교육학박사학위논문

A Real-Time Analysis of Korean EFL Students' Speech Fluency: Focusing on Korean Raters' Construction of Fluency Profile

한국인 EFL 학생들의 말하기 유창성에 대한 실시간 인식 변화에 대한 연구: 한국인 채점자의 총체적 유창성 판단 과정을 중심으로

2022년 8월

서울대학교 대학원

외국어교육과 영어전공

김재희

A Real-Time Analysis of Korean EFL Students' Speech Fluency: Focusing on Korean Raters' Construction of Fluency Profile

한국인 EFL 학생들의 말하기 유창성에 대한 실시간 인식 변화에 대한 연구: 한국인 채점자의 총체적 유창성 판단 과정을 중심으로

by

Jaehee Kim

A Dissertation Submitted to the Department of Foreign Language Education in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in English Language Education

At the

Graduate School of Seoul National University

August 2022

A Real-Time Analysis of Korean EFL Students' Speech Fluency: Focusing on Korean Raters' Construction of Fluency Profile

한국인 EFL 학생들의 말하기 유창성에 대한 실시간 인식 변화에 대한 연구: 한국인 채점자의 총체적 유창성 판단 과정을 중심으로

지도 교수 안현기

이 논문을 교육학박사 학위논문으로 제출함

2022년 6월

서울대학교 대학원
외국어교육과 영어전공
김재희

김재희의 교육학박사 학위논문을 인준함

2022년 7월

위 원 장 _____ (인)

부위원장 _____ (인)

위 원 _____ (인)

위 원 _____ (인)

위 원 _____ (인)

A Real-Time Analysis of Korean EFL Students' Speech Fluency: Focusing on Korean Raters' Construction of Fluency Profile

by
Jaehee Kim

A Dissertation Submitted to the Department of Foreign Language Education in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in English Language Education at the Graduate School of Seoul National University

JULY 2022

APPROVED BY DISSERTATION COMMITTEE:

_____
KITAEK KIM, COMMITTEE CHAIR

_____
SUN-YOUNG OH

_____
IN YOUNG YANG

_____
SUYEON IM

_____
HYUNKEE AHN

ABSTRACT

A Real-Time Analysis of Korean EFL Students' Speech Fluency: Focusing on Korean Raters' Construction of Fluency Profile

Jaehee Kim

Department of Foreign Language Education (English Major)

Graduate School of Seoul National University

The oral fluency of a second language (L2) speaker is often used as a criterion in the second language assessment. A number of studies have investigated the characteristics and features of L2 fluency within the quantitative paradigm. The present study extends the field by analyzing raters' perceptions of L2 fluency during evaluation from a qualitative perspective. While previous quantitative studies have attempted to determine the factors or variables which influence L2 fluency, the present study focuses on L2 raters' subjective perception of speech features used as the criteria for rating fluency. Using Idio-dynamic software and stimulated interview skills, the raters' real-time assessment process is investigated and the raters' constructions of fluency profiles are examined.

This study investigated the features of L2 learners' oral production

that influence perceptions of L2 fluency in the speaking test. Listening to audio clips of six Korean college students with varied English proficiencies, the goal of the study was to examine the factors that affect raters' evaluations of fluency. Seven raters graded fluency dynamically, using the Idio-dynamic Software to upgrade or downgrade fluency over the course of the listening task. The rating process was followed with stimulated recall; raters were interviewed to determine which aspects of L2 fluency were associated with enhanced or diminished fluency. Qualitative analysis revealed that raters' fluency judgment continuously changed indicating moment-to-moment shifts over time, which can be referred to as dynamic. The main speech features influencing raters' perceptions of L2 fluency were speech rate, pauses, self-repair, grammatical/phonological accuracy, and automatized production. These speech features were inherently intertwined and combined during the fluency rating process, rather than being clearly distinguished from each other and independently applied to fluency assessments.

The effects of rating modes (i.e., audio-only ratings, video-mediated ratings) on perceiving features of fluency were analyzed. The effects of the rating modes were not statistically different in the test-takers' scores. However, the number of comments on certain features changed between the rating modes. Raters tended to focus on automaticity

and pause phenomena during the video-mediated ratings while they were more sensitive to accuracy in audio-only ratings. In addition, raters' comments reflected their overall impression of the speech samples by watching the video clips while they recognized more specific errors in each sentence by listening to the audio clips. However, the speech features were categorized in the same theme and emphasized to be balanced for fluent speech regardless of rating modes.

Although inter-rater reliability was being observed and their correlation was high, each rater's assessment process varied referring to different linguistic features for their judgment. The results provide important insights accounting for the complexity of perceiving L2 fluency by non-native raters in the language testing context. In addition, raters' definitions of fluency varied and raters tended to judge students' fluency levels in relation to various linguistic dimensions. Therefore, it is necessary to discuss the ways to develop rubrics that include major variables influencing fluency assessment in English speaking tests, and train raters to have a better understanding of fluency assessment.

Keywords : speaking fluency, fluency rating, perceived fluency,
qualitative approach, dynamic rating, rating modes

Student Number : 2007-30391

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

The present study uses the Idio-dynamic method to investigate the real-time assessment process of fluency perception among those learning English as a second language. The first section introduces the motivation and purpose of the current study, the second section presents the research questions, and the last section outlines the organization of the thesis.

1.1 Background and Purpose of the Study.

Fluency is one of the criteria for second language (L2) assessment, and it is specified in the scoring rubrics of many English speaking tests such as TOEIC Speaking (Test of English for International Communication Speaking), TOEFL iBT (Test of English as a Foreign Language internet Based Test), and ACTFL OPIc (Oral Proficiency Interview – computer of the American Council on the Teaching of Foreign Language). For example, TOEIC Speaking test describes fluency in relation to long pauses and frequent hesitations, and the ACTFL OPIc mentions fluency, referring to hesitations, pauses, and reformulations. In the TOEFL iBT, “degree of automaticity” and “pace” are mentioned in reference to fluency. Though fluency has always recurred in scoring rubrics, the statements concerning fluency among tests are discrepant in respect of terms and

considered factors, which leaves room for subjective interpretation for scoring fluency in English speaking tests.

In the field of L2 assessment, fluency is not a synonym for overall oral proficiency (Chambers, 1997; Luoma, 2004), but it is one component of oral proficiency, which includes accuracy and complexity of linguistic forms (Housen & Kuiken, 2009). Many previous studies have examined factors influencing raters' judgment (Iwashita et al., 2008). A number of studies have used a systematic research approach to evaluate L2 fluency, using temporal variables and native-speaker judgements within the quantitative paradigm (Derwing et al., 2004; Freed et al., 2004; Ginther et al., 2010; Kormos and Dénes, 2004; Lennon 1990; Riggenbach 1991; Towell et al., 1996). The results of these various studies raise a question: what features are actually recognized and applied by raters in real time to evaluate L2 fluency? While the aforementioned purely quantitative studies have attempted to determine the factors or variables influencing L2 fluency, the present study focuses on raters' subjective perceptions of L2 fluency during evaluation. An approach to investigate perceptions of fluency using qualitative data can yield crucial insights to account for the complexity of raters perceiving L2 fluency, especially in L2 language assessments.

In the present study, raters' perceptions of the L2 fluency rating are compiled and analyzed qualitatively. The intent of this study is not to

reproduce the extensive quantitative L2 fluency research using temporal variables. Rather, this study tries to extend the field by analyzing listeners' perceptions of L2 fluency in English and the features influencing "rating fluency" during assessment. In addition, the research aims to detect the linguistic processing experience of raters when evaluating L2 speech fluency in real time. In previous research, fluency was measured at a single time using a Likert-type scale and rated by listeners who lacked fluency linguistic knowledge. The present study focuses on variables that raters consider important and compares dynamic ratings with holistic ratings. As a complete understanding of listeners' perceptions of fluency requires an examination of the assessment processes of raters, it is important to examine listeners' sensitivity to (temporal) factors as well. It is worth understanding the assessment processes of raters because speaking proficiency is still commonly measured by means of a human listener's judgment in high-stakes language proficiency tests as well as performance tests in a classroom. Furthermore, understanding the criteria that the raters apply to evaluate L2 fluency will be the first step toward developing valid and reliable rubrics and rater training for fluency assessment in L2 speaking tests.

1.2 Research Questions

This study aimed to investigate raters' real-time perception of fluency evaluation in L2 speech and determine the linguistic features of L2 learners' oral production that influence raters' decision during assessment. In addition, the effect of non-linguistic factors such as rating mode is also examined because two rating modes such as face-to-face interview and audio-recorded assessment are commonly used for speaking tests. For example, TOEFL iBT and TOEIC adopted computer delivery mode on speaking tests while IELTS and Cambridge English Assessment Test still adhere to interview format. Most previous studies focused on objective measures of oral fluency and related linguistic features, with few studies attempting to investigate raters' subjective perception of L2 fluency during the assessment process and non-linguistic variables. In light of these research needs, the specific research questions addressed in the present study are as follows:

1. How listeners evaluate fluency for L2 English speakers and what speech features influence their moment-to-moment fluency judgments?

2. What are the effects of different rating methods (i.e. listening to audio clips and watching video clips) on recognizing features which influence the perceived fluency rating?

1.3. Organization of the Thesis

This thesis is composed of five chapters. Chapter 1 introduces the motivation and research questions of the present study. Chapter 2 reviews how fluency has been defined and operationalized along with some major findings of experimental researches on the linguistic features affecting fluency. Chapter 3 describes the research method of the main experiment including speaking test, rating procedure, and data analysis. Chapter 4 discuss findings and related issues. Chapter 5 summarizes the major findings and concludes the study with pedagogical implications, limitations, and suggestions for further studies.

CHAPTER 2. LITERATURE REVIEW

This chapter provides an overview of the body of literature pertinent to this study. First, the definition and measures of fluency in L2 research are presented. Then, the Idio-dynamic method is discussed. Finally, the literature on test mode variability and rater variability in the speaking assessment context is discussed.

2.1 Defining Fluency in L2 Research

Even though a number of studies have evaluated L2 fluency, there is no singular definition of fluency. However, Nevertheless, some definitions are as follows. Fillmore (1979) described four elements of fluency: 1) the ability to talk at length with few pauses; 2) the ability to talk cohesively and logically using "semantically dense" sentences; 3) the ability to talk in a wide range of contexts or situations; and 4) the ability to be creative and imaginative with language use. Crystal (1987) defined fluency as "smooth, rapid, use of language." Lennon (1990) explained two senses of fluency: the broad and narrow sense of fluency (p. 389). The broad sense of fluency roughly corresponds to overall oral proficiency, and the narrow sense of fluency refers to the speed and smoothness of oral proficiency. Lennon (2000, p. 26) defined the narrow sense of fluency as the "rapid, smooth, accurate, lucid, and efficient translation of thought or

communicative intention under the temporal constraints of on-line processing."

The discrepancies among definitions of fluency arose from different viewpoints: that of the listener and that of the speaker (De Jong, 2018). The definition by Lennon (1990) describes the listener's impression of the ease of speech articulated by the speaker. Conversely, other definitions assume the viewpoint of the speaker, focusing on the speaker's ease or trouble in the speech production process. Similarly, Luoma (2004, p. 88) recognized the different viewpoints stating that temporal characteristics were not simply descriptions of a person's speech but markers of the listener's perception.

Segalowitz (2010, p. 165) distinguished three notions of fluency: cognitive, utterance, and perceived fluency. Cognitive fluency is "the efficiency of operation of underlying processes responsible for the production of utterances," which is the speaker's capacity to utilize the underlying cognitive processes. Utterance fluency concerns "the features of utterances that reflect the speaker's cognitive fluency" (p. 165) referring to the temporal, pausing, and repair characteristics of utterance, which can be acoustically measured. Perceived fluency relates to "the inferences that listeners make about the speaker's cognitive fluency based on the utterance fluency" (p. 165).

With regard to temporal behavior of a speaking performance, the notion of temporal fluency is divided into three subconstructs: speed, breakdown, and repair fluency (Skehan, 2003, 2009; Tavakoli & Skehan, 2005). Speed fluency reflects the speed of delivery. Breakdown fluency refers to pausing behavior, including the frequency, location, and duration of pauses. Repair fluency is concerned with dysfluency phenomena, such as repetitions and false starts. This study focuses on the raters' perception of temporal features influencing fluency rather than measuring breakdown, repair, and speed fluency qualitatively.

2.2 Measuring Fluency in L2 Research

To identify speech features affecting fluency, a number of studies identified acoustic characteristics of L2 speech (Bosker et al., 2013; Chambers, 1997; Cucchiarini et al., 2000; Cucchiarini et al., 2002; Ginther et al., 2010; Kormos & Dénes, 2004; Lenno, 1990; Möhle, 1984; Riggenbach, 1991; Towell et al., 1996; Wood, 2004). Kormos and Dénes (2004) examined ten temporal features to measure fluency based on a monologic narrative task with fixed content. The features were 1) speech rate, 2) articulation rate, 3) phonation to time ratio, 4) mean length of run, 5) the number of silent pauses per minute, 6) the mean length of pauses, 7) the number of filled pauses per minute, 8) the number of disfluencies

per minute (repetition, restarts, and repairs), 9) pace (the number of stressed words per minute), and 10) space (the proportion of stressed words to the total number of words).

According to previous studies, speech rate and pausing phenomena are the best predictors of fluency (Derwing et al., 2004; Lennon, 1990; Riggenbach, 1991; Rossiter, 2009). Speech rate (i.e., the number of syllables per minute, including pause time) and mean length of run (i.e., the mean number of syllables between two silent pauses) were consistently and strongly correlated with L2 fluency (Kormos & Dénes, 2004; Towell et al., 1996). Lennon (1990) investigated the longitudinal development of fluency. Ten native English teachers judged four German advanced EFL learners' fluency at the start and end of a six-month residence in Britain. According to the findings, the speaking rate and mean length of run increased, and the frequency of filled pauses decreased as fluency developed. Cucchiarini et al. (2000) concluded that speech rate appeared to be the best predictor of listeners' fluency among beginner learners. They further concluded that the number (as opposed to length) of unfilled pauses affected perceived fluency. In their research, ten teachers of Dutch rated the fluency of spontaneous speech from Dutch learners at the intermediate and beginner level. They compared their subjective ratings of fluency to objective fluency indicators and discovered that speech rate and articulation rate were the best indicators

among beginning level learners. Mean length of run on the other hand was more predictive of fluency among intermediate learners. Kormos and Dénes (2004) compared temporal features of speech produced by intermediate and advanced learners of English and found that there were statistically significant differences between fluent and non-fluent participants in speech rate, phonation to time ratio, and the mean length of run, but not in articulation rate. In contrast, Ginther, Dimova, and Yang (2010) investigated relationships between oral English proficiency and the temporal measure of fluency and found that articulation rate strongly correlated with speaking scores, though less strongly than speech rate. Though there have been discrepancies in the relationship between articulation rate and fluency rating, it is obvious that speech rate is the best indicator of L2 fluency.

Pauses, another predictor of fluency, are normally defined as a break in speech or a moment of silence. O'Connel and Kowal (1983) define pauses as "the absence of speaking." Kowal and O'connel (2008) redefined pauses as periods in which vocalization was absent, and these moments were referred to by names like "silence, pause, gap, lapse, and offtime." However, pauses are not limited to silent gaps. There are two kinds of pauses examined in disfluency research: silent pauses and filled pauses (Crystal, 1987). Silent pauses are silent periods of non-articulation. In contrast, filled pauses are vocalized pauses, such as "um"

and “uh” (Clark & Tree, 2002). Many researchers have explored the role of pauses in L2 speaking and investigated the relationship between fluency and pause.

Studies on pause phenomena measure pauses by including the number, the length, and the location of the pauses. The majority of studies on the relationship between pauses and L2 fluency focused on the number and length of pauses. Interestingly, the research findings on pausing effects provide inconsistent results. Ginther et al. (2010) reported that both pause frequency and pause length were negatively correlated with proficiency scores and fluency ratings, respectively. Similarly, in Bosker et al. (2013), pause frequency and pause length were negatively correlated with fluency judgments. De Jong and Perfetti (2011) found that pause length along with phonation/time ratio and mean length of fluent run correlated with fluency ratings. However, in Kang’s (2008) study, the number of silent pauses predicted the judgment of oral proficiency, but the length of pauses was not a strong predictor of fluency ratings. A similar result was found by Kormos and Dénes (2004). The number of silent and filled pauses was not significantly associated with listeners’ fluency ratings. Instead, the raters in that study focused more on temporal characteristics including speech rate, mean length of utterance, phonation time ratio, and the number of stressed words/minute. By contrast, Cucchiarini et al. (2002) found the opposite pattern. Fluency correlated

with pause frequency but not with pause length.

Some studies on pause location classified pauses into two categories: pauses between clauses and pauses within clauses. Such studies suggested that fluent speech tended to contain pauses at grammatical boundaries, whereas non-fluent L2 speech often had pauses within clauses or utterances (Davies, 2003; de Jong, 2016; Kahng, 2014, 2018; Riazantseva, 2001, Riggenbach, 1991; Tavakoli, 2011). Language seems to be encoded one clause at a time in fluent speech (Pawley & Syder, 2000), and pausing within clauses seems to reflect difficulties in planning or encoding speech (Cenoz, 1998; Lennon, 1984; Wood, 2010). Kahng (2014) pointed out that one of the biggest differences between L1 and L2 utterance fluency is the number of pauses within a clause. Another study by Kahng (2018) examined the influence of pause location on perceived fluency of L1 and L2 speech. The findings suggested that pauses within clauses lowered fluency ratings compared to pauses between clauses. In Kang's (2010) study on International Teaching Assistants (ITAs) accentedness, the proportion of atypical topic boundary pauses within clauses revealed a strong effect on fluency judgments.

Self-repair was investigated in most research examining fluency. Self-repair, a form of reformulation, refers to self-initiated corrections of a problem arising in one's speech-production processes (Kormos, 2000). Levelt (1983) distinguished self-repair into two subtypes:

appropriacy repairs and error repairs. Appropriacy repairs are false starts, the abandonment of an utterance followed by its immediate revision with the intention of improving coherence. Error repairs are the attempted replacement of perceived non-standard output (e.g., of syntax, lexis, or pronunciation) with a form that a fluent speaker would recognize as standard. Repair fluency relates to the number of corrections and repetitions present in speech. According to Lennon (1990), self-repetition may reflect planning processes, and a decrease in self-repetitions may be interpreted as an increase in oral fluency, although self-corrections did not appear to be a reliable indicator of fluency. The findings of prior research on repair fluency are inconsistent (e.g., Bosker et al., 2013; Cucchiariniti et al., 2002; Kormos & Dénes, 2004; Tavakoli et al., 2020). There has been little agreement on the extent to which repair measures accurately capture the fluency of L2 speakers. For example, Kormos and Dénes (2004) indicated that repair fluency was not a predictor of fluency, though speed and breakdown fluency were highly correlated with perceived fluency. Conversely, Bosker et al. (2013) observed that repair did contribute a small but significant amount to perceived fluency. In Kahng's (2014) study, L2 speakers used more self-repetitions than L1 speakers, presenting a weak negative correlation between self-repetitions and overall speaking scores. She argued that self-correction was affected by personality and L2 learning experience. Similarly, Suzuki et al. (2021) indicated that repair fluency was more

strongly associated with an individual's speaking style than L2 proficiency, demonstrating that fluency was stable throughout L1 and L2 production and across L2 proficiency levels.

In L2 acquisition and language testing literature, definitions of fluency are generally linked to quantitative temporal aspects, such as speed, pause phenomena, and the ability to produce fluent runs of speech (Brumfit, 2000; Ejzenberg, 2000; Fillmore, 2000; Kormos, 2006; Pawley & Syder, 1983; Sajavaara, 1987; Schmidt, 1992; Segalowitz, 2010). In addition to the temporal factors influencing perceptions of speaking fluency, general follow-up discussions with raters in several L2 studies have suggested that fluency judgments may also be affected by non-temporal variables, such as accent, grammar, vocabulary, intonation, and confidence (Freed, 1995; Lennon, 1990, 2000; Rossiter, 2009; Wennerstrom, 2000). Riggenbach (1991) asserts that "in order for there to be fluency … it appears that many different conditions have to be met – some proficiency in grammar, pronunciation, and vocabulary to mention a few" (p. 439). This study investigates the relationship between listeners' sensitivity to L2 speech characteristics and L2 fluency perception. We examine raters' subjective perception the features they use to evaluate fluency rating. This study focuses on the subjective rating processes rather than objective measures of oral fluency.

2.3 Idio-dynamic Method

Language can be seen as a dynamic system, which means all factors or variables involved in language development are interconnected, interact with each other over time, and affect each other differently over time on different time scales. MacIntyre et al. (1998) examined the variability and interaction of willingness to communicate and L2 language use by means of their "Idio-dynamic method". The method includes videotaped interviews, question-and-answer exercises, and conversations among L2 learners. Using a form of stimulated recall, learners review the video recording in order to rate their fluctuating affective reactions. Learner ratings are graphed as a continuous curve, printed immediately, and reviewed by the learner and a research assistant. The transcripts of each learner's L2 speech can thus be linked to peaks and valleys observed on the graph, for example, to study verbal or non-verbal markers of changing affective states.

Nagle et al. (2019) took a dynamic approach to L2 comprehensibility and examined how listeners construct comprehensibility profiles for L2 Spanish speakers during the listening task and what features enhance or diminish comprehensibility. Listeners rated comprehensibility dynamically using Idio-dynamic software to upgrade or downgrade comprehensibility over the course of the listening

task. Dynamic ratings for an audio clip were video-captured for stimulated recall, and listeners were interviewed to understand which aspects of L2 speech was associated with enhanced versus diminished comprehensibility.

The present study employed the Idio-dynamic method to investigate how listeners evaluate fluency for L2 English speakers and what speech features influence their moment-to-moment fluency judgments. Speaking and listening are dynamic acts whose properties fluctuate over time. In the language testing context, as L2 speakers produce varying levels of speech features such as speed and pause or repair over time, listeners must continuously process these variabilities to evaluate the speakers' fluency levels. Even in the case of L1 speech, for example, speakers generally appear to alternate between periods of fluent and disfluent speech, and these temporal cycles occur on a time scale of 10–30 seconds (Pakhomov et al., 2011). The present research examined raters' subjective experience of perceiving features which affected fluency rating rather than the actual temporal features.

2.4 The Effects of Test Mode on Scoring

Tests of speaking ability are considered subjective in nature since they involve human judgement (Carr, 2011). In addition, whether live or

recorded, speaking assessment involves raters having to rely on their listening skills and oftentimes their short-term working memory. This is in contrast to raters of written samples, who always have a document to rely on (Ginther, 2013). In test validation, language testers have long held an interest in specifying and minimizing the factors that confound score interpretation.

One factor possible unintended effect on scoring is delivery mode of the test, that is, computer-delivered speaking tests or face-to-face tests. The delivery mode of a speaking test and its effects on the assessment process has primarily been studied in relation to their impact on test-taker performance in computer-mediated tests. For instance, previous studies have addressed the issue of face validity (Kenyon & Malabonga, 2001) and the effectiveness of technical aspects (Malabonga et al., 2005). There are also studies that have examined test takers' strategic behaviors on the speaking section of the TOEFL iBT (Swain et al., 2009) and have compared test takers' performance on the test with their actual academic performance (Brooks & Swain, 2014). Zhou (2015) compared the computer-delivered mode to face-to-face interviews, focusing on the test scores assigned to analytical scales. The results showed no mode effect on the test score. Raters assigned similar ratings to speakers' performance during each of the modes because the participants performed similarly between the two modes. The studies on

delivery mode focus more on the test takers' performance than the raters' performance.

Some research has been conducted to compare the scoring of audio-recording samples in comparison to the scoring of live performance tests. When only audio is provided, it has been found that the more proficient examinees are affected since their actual level of proficiency is underestimated by the raters. In contrast, examinees with adequate use of nonverbal behavior received higher scores when their performance was video recorded and shown to raters in this format (Nambiar & Goon, 1993). These findings align not only with the fact that higher-ability language learners synchronize speech with nonverbal behavior (Neu, 1990), but also with the interactional competence approach to defining speaking ability which posits that nonverbal behavior is, in fact, a part of speaking ability (Ducasse & Brown, 2009). The study of Joo and Kim (2011) revealed that examinees performed significantly more fluently in a face-to-face interview than in the computer-mediated speaking test. They hesitated and reformulated sentences more frequently in the computer-mediated speaking test.

In terms of scoring of recorded speech samples of speaking performance, little has been said about whether the type of recorded speech sample may have an effect on the consistency or severity of rating. Studies of speaking assessment and rater biases choose one of the

speech sample types, either audio or video. Nakatsuhara (2007) and O'Sullivan (2002), for instance, made use of videotaped interviews to study interviewee-interviewer effects in the assessment of a speaking test while Ekes (2005) and Winke et al. (2011) made use of audio recordings only.

Few studies compared the two types and their impact on the rating. Lavolette (2013) conducted one of the studies that compared ratings of audio and audio-visual speech samples. Lavolette examined the ratings of audio-only samples, video samples, and samples with audio from the video samples in the context of formative assessment. In their rating of 39 ESL examinees performing the TOEFL iBT direct speaking task, raters were found to significantly favor both types of audio-only samples, contrary to Nambiar and Goon's (1993) findings. Thus, it was determined that the choice of speech sample type could be a factor of unexpected rater variance. Beltrán (2016) examined the effects of audio-only format and audio-visual format on the scoring of speaking test performance. The findings of his study suggested that the inclusion of visual stimuli did not have significant effects on assigned scores or internal consistency. Yet, raters prefer audio-visual speech samples to audio-only speech samples because video provided a more authentic experience. The intended message of the speaker and the delivery of the speech were better understood. The present study also focuses on the effect of rating modes

and explored whether audio-only input or video input in the scoring process may have an impact on raters' behavior. In addition, this study attempts to compare the effects of rating modes on perceiving linguistic features influencing fluency scores.

2.5 Non-native Rater Variability in Speaking Assessment

It is general practice to use rater judgments in speaking proficiency testing. However, it has been shown that raters' knowledge and experience may influence their ratings, both in terms of leniency and varied focus on different aspects of speech. In the process of language proficiency rating in general, the knowledge and experience of the raters play a central role (Lumley, 2005). Regarding the rating of speaking proficiency, the raters' assignment of scores appears to be related to their severity/leniency (Brown, Iwashita, & McNamara, 2005; Carey, Mannell, & Dunn, 2011; Hsieh, 2011; Kang, 2008; Rossiter, 2009). Furthermore, several studies have established that the raters' backgrounds determine the language performance features on which they are apt to focus (Eckes, 2008; Hsieh, 2011; Zhang & Elder, 2011).

Numerous studies have been conducted to identify and analyze rater effects on the scoring of speaking tests. Some studies have focused on the effects of rater characteristics (McNamara & Adams, 1994), in

relation to inherent qualities of the raters such as language background, gender, or educational training. For example, Ceban (2003) conducted a study to determine whether the differences found between four groups of raters assessing four interviews could be attributed to their language background or academic training. The raters were either L1 English speakers or L1 Japanese speakers and were from one of four educational background groups (graduate students with EFL or ESL background, ESL teachers, or ESL students). After conducting a Facets analysis to identify possible biases, it was determined that the variation between the four rater categories in this study could not be attributed to language or educational background, even though tendencies of leniency or severity could be observed in the data.

Raters' language background is one widely-investigated factors in oral assessment. The native speaker/non-native speaker (NS/NNS) status was one of the predictors of oral performance rating and non-native speakers were found to be harsher with proficiency ratings than were native speakers (e.g., Brown, 1995; Fayer & Krasinski, 1987; Kang, 2008; Santos, 1988). For example, Brown's (1995) results pertaining to the Japanese Test for Tour Guides showed that Japanese raters were substantially harsher than English NS raters on linguistic items such as pronunciation. Santos (1988) reported that when NNSs rated other NNSs' language ability, the raters' effort in attaining a high level of proficiency

led them to attribute errors to a lack of commitment on the learner's part. On the other hand, NS raters may have a more global view and not worry about non-native features as long as they do not seriously impede communication. Kang (2008) also found that NNSs were significantly severe with the comprehensibility and proficiency ratings in her research. She added that NNS assessors, who have gone through this complex learning procedure themselves, tend to be less tolerant of others' mistakes. On the contrary, Kang (2013) found that native English raters tended to assess more strictly than Korean raters. He investigated validity and reliability of Native English raters and Korean raters for Korean English speaking assessments. Even though native raters were stricter in rating oral proficiency, Korean raters evaluated more strictly on the grammar section. Lee (2010) found that some non-native English teachers are strict in the grammar part while native English teachers are generous in assessing grammar. It seemed that some Korean teachers were deeply grammar oriented and therefore strict on grammar in the oral assessment and in overall language learning.

There has been research identifying the effects of rater variability, especially on fluency rating. A variety of types of listeners have been found to reliably rate speech fluency. These range from relatively expert raters (i.e., linguists, teachers, speech therapists) in Cucchiarini et al. (2002) to untrained native speakers (Derwing, Rossiter, Munro, &

Thomson, 2004; Freed, 1995) and L2 learners (Riggenbach, 1991; Rossiter, 2009). Raters from a variety of backgrounds may rely on similar cues in the speech stream (e.g., Kormos & Dénes, 2004). For example, Préfontaine (2013) found that French L2 learners' self-perceptions of fluency were moderately correlated with native French listeners' fluency ratings. Most commonly, these include speech rate and pausing phenomena (e.g., Derwing et al., 2004; Lennon, 1990; Riggenbach, 1991; Rossiter, 2009). According to Kang and Ahn (2012), in particular, Korean raters placed more focus on pause duration in fluency evaluation, while native English raters focused on speech rate. In terms of the difference between non-native and native raters in fluency rating, non-native raters tended to be more severe in general (Fayer & Krasinski, 1987). Additionally, L2 learners may rate the speech of fellow L2 learners as less fluent than native speakers (Rossiter, 2009). The present study also targets Korean raters who evaluate L2 English learners. Using the Idio-dynamic method, it is expected to provide more data on how non-native raters perceive and evaluate L2 fluency in time-sensitive constructs.

In summary, this study investigates L2 raters' subjective perception of which speech features they are using as criteria for rating fluency. Using Idio-dynamic software and stimulated interview skills, raters' real-time assessment process is examined. The present study also focuses on the effect of audio-only delivery mode or video-mediated

mode on raters' scoring and perception of fluency. Subsequently, the subjective rating process of individual raters is analyzed in terms of nonnative rater characteristics.

CHAPTER 3. METHODOLOGY

This chapter first describes the methodology used to collect speech samples in Section 3.1. Detailed descriptions of the rating process for the current study are provided in Section 3.2.

3.1. Speaking Tests

3.1.1 Participants

The speakers included in this study were college students who learned English as a foreign language in Seoul, Korea after age eight. All speakers were female and from the same college. Their native language was Korean, and none spoke any other languages at home during their childhood. Their English language proficiency levels varied.

Eleven students volunteered to participate in this study, but two students dropped out for personal reasons. The remaining nine students participated in the speaking test which was conducted one-on-one through the video communication system Zoom (www.zoom.us), a video communication application. However, only six students' speech samples could be used for analysis. One speech sample was abandoned due to frequent use of Korean, because the participant's English language proficiency was too poor to perform the given tasks. She often spoke

Korean to explain herself and sometimes asked the researcher for the correct expressions when she could not express herself in English. Two students took advantage of the video conference setup and cheated on their tests. The students prepared and typed their answers in English and then read their answers. Since this study did not focus on reading, the speech samples of those two students were excluded. Therefore, the speech samples from six students were analyzed. For the analysis, each speaker was assigned an ID number (from Speaker 1 to Speaker 6)

3.1.2. Procedures

This research was originally designed to see if the fluency assessment changed during grading in face-to-face tests and in audio-recorded tests. However, the face-to-face test method had to be changed to a contactless format due to onset of the COVID-19 pandemic. To create an assessment environment similar to face-to-face test evaluation, a video recording method through Zoom was used. Through this method, speaking tests were conducted in the form of interviews, and the raters could obtain visual information as in a face-to-face situation while the raters listened to the test-takers answering.

Participants were asked to access Zoom and joined in a Zoom Meeting at the agreed time. Each student participant had a Zoom account provided by their college and was familiar with the tool, having used this

application to take classes for several months. Test questions on the researcher's computer monitor were presented on the student's computer monitor using Zoom's screen-sharing function. The test process was recorded using Zoom's screen-recording function and saved as a video file (mp4.) at the end of the test. The video files were converted to audio files (e.g., mp3., wav.) using the program Wondershare Uniconverter (https://videoconverter.wondershare.com). This conversion was done to observe the difference in rating modes, especially the difference of fluency features recognized by the raters in audio-recorded speech samples versus video-recorded speech samples.

The speech samples were transcribed by a researcher using Praat (http://www.fon.hum.uva.nl/praat), a program widely used for voice analysis. The length and location of all pauses were recorded(see Appendix A for an example of transcription).

3.1.3. Tasks

Two types of monologic tasks were used: a picture description task and an opinion task (see Appendix B, C). The picture description task required students to describe pictures on the screen in as much detail as possible. The opinion task required students to express their opinions on a specific topic. These two tasks are traditionally used in pausological research and

commonly used in L2 assessments. Although dialogue, given its interactive nature, represents a more natural and authentic environment (Riggenbach, 1991; Guillot, 1999; Van Lier, 2004), research in second language acquisition (SLA) frequently uses monologic performance, including reading-aloud, sentence repetition, information transfer, and oral presentation (O'Sullivan, 2008). Tavakoli (2016) pointed out that L2 fluency research has predominantly focused on measuring monologues due to the difficulty of measuring fluency in dialogue. She suggested that monologic task performances are more controllable and predictable due to the simpler pragmatic demands for speech planning. In addition, the procedure for measuring monologues is easier than dialogues, which reflect interactive aspects, such as overlap, unclaimed between-turn pauses, and the interdependence of the interlocutors' performances. The complex pragmatics involved in dialogue leading to less controllable and predictable performance is an additional factor making analysis difficult.

Since this study focused on computer-based tests widely adopted in Korea, monologic speaking tasks were chosen. Monologues have the advantage of control and the procedure for measuring pauses is clearer and simpler. Further, integrated tasks were excluded to eliminate factors affecting test performance, such as listening or reading comprehension. The picture description task and the opinion task do not require preceding listening or reading comprehension.

Two questions were presented for each task in the event a picture or a topic was unfamiliar to or difficult for the participants. Therefore, speakers answered four questions (two picture description questions and two opinion questions). Questions were presented one at a time. The first picture appeared on the screen and students had one minute to prepare their responses followed by 45 seconds to speak about the picture. The process was repeated with the second picture. After the picture description tasks, the opinion tasks were proctored to students. A question about the first topic, which was related to vacation, appeared on the screen, and students had two minutes to prepare followed by one minute to speak. The process was repeated with the second topic, which concerned social media. Students were not interrupted while speaking, even after the allotted answer time expired. However, each recorded video clip was adjusted to cap the length of speech at 45 seconds for the picture description task and one minute for the opinion task. Individual speaking tests lasted between 10–20 minutes. However, only one response for each task, with the larger speech sample, was used for analysis.

3.2. Rating

3.2.1 Raters

The seven raters were experienced English teachers who worked as

college English instructors in Korea. Each was an English education expert who had majored in English education; four of them held a Ph.D. in English education and three of them held an M.A. in English education. All raters worked as English L2 teachers in Korean universities and taught English L2 for 6 years or more (mean 12.4 years, maximum 20 years). Ages ranged from 38–47 years and all were female. For the analysis of their scores and interview responses, each rater was assigned an ID number, from Rater 1 to Rater 7.

All raters had experience judging the L2 oral fluency of Korean college students. Each had experience evaluating the speech of L2 English students by rating other tests such as classroom assessments. Raters had proctored oral examinations in the form of an interview or a presentation to their students. Four raters (Rater 1, Rater 3, Rater 5, Rater 6) had experience teaching TOEIC speaking preparation classes. Given this experience, the raters are more likely to provide a consistent and accurate assessment of fluency (Préfontaine et al., 2016).

It was explicitly stated during the recruitment process that raters should be English education majors and teachers who have experienced language teaching and testing. This was important given the need to understand the concepts of fluency used in the English speaking test. It was necessary that raters be able to distinguish fluency in the narrow sense from fluency in the broad sense. Fluency in the broad sense is

equivalent to overall proficiency (Chambers, 1997), which considers grammar, vocabulary, and pronunciation. For example, being "fluent" in English may refer to error-free grammar, a large vocabulary, and/or native-like English pronunciation. However, this study is concerned with fluency as a component of speaking proficiency, especially in oral examination; thus, the flow and smoothness of the speech must be assessed separately from grammar and vocabulary. Experienced raters scored each scoring area independent of other areas (e.g., task performance, language use, grammar, pronunciation, and so on), while the inexperienced raters decided the scores of given scoring areas interdependently (Cumming et al., 2002; Song & Lee, 2015; Wolfe et al., 1998). Therefore experienced English L2 teachers who majored in English were recruited with the expectation that they evaluated fluency more independently. Another expectation was that raters give more specific reasons for their decisions in the context of L2 acquisition or L2 language testing.

3.2.2 Procedures

Individual raters' judging fluency sessions took place in a quiet location and lasted between 90-120 minutes. Raters were informed that the goal of this study was to distinguish variables which influenced the evaluation

of L2 fluency in speaking tests. To elicit information about the raters' definitions of fluency, the researcher posed the following question: "For evaluating fluency in English speaking tests, what do you think is the factor determining the fluency score?" Fluency was defined as the impression of how easily and smoothly speech was delivered. Prior to initiating fluency judgments, the researcher emphasized that raters needed to assess the flow and smoothness of the speech separate from grammar and vocabulary. Raters received examples which were "fluent but grammatically inaccurate" or "grammatically correct but not very fluent." A written scoring rubric was not provided to the raters, and the temporal features of fluency rating were not emphasized in order to avoid imposing a self-fulfilling construct of L2 fluency upon the raters.

To record fluency ratings, Idio-dynamic software (MacIntyre, 2012) was used. This software is freely available (http://faculty.cbu.ca/pmacintyre) and allows users to record time-locked ratings (in one second increments) by clicking to raise or lower the rated level to values ±5 relative to the baseline (marked by a straight line crossing 0). The raters were instructed to click the button labeled "increase fluency" when they felt that the speaker was fluent and click the button labeled "decrease fluency" when they felt that the speaker became less fluent. The raters were told that each successive click of the mouse corresponded to an additional increase or decrease in their

rating. Each click appeared as an upward or downward block on a color bar graph. In the absence of rating activity from the user, the software engaged a built-in auto-zero function, returning the rating to the baseline at the rate of one click-point per second (see Appendix D & E).

Raters listened to each speech sample several times: to rate fluency, during the stimulated interview. Raters first heard and rated the audio files which were converted from the original video files. The audio-only files were rated before the video files, because visual memory has a longer duration and more accurate recall than auditory memory (Butcher, 2006; Cohen et. al., 2009; Kargopoulos et. al., 2003; Lindner et. al, 2009). Further, rating the audio-only files first, eliminated or reduced the effects of information recall about the speakers.

Prior to rating the audio clips, the raters practiced employing the software with a speech sample clip not selected for this study. Following confirmation that each rater understood the task and the computer program, they listened to the audio clips selected for the study and rated them using the onscreen interface. First, speech samples from the picture description task were assessed, followed by the opinion task samples. The raters' reliability improved if the order of the samples was presented by task rather than by speaker. The speech samples of the same task were loaded in random order.

Raters were permitted to listen to the audio clip as many times as necessary. However, they were not allowed to rewind or fast forward the files because the Idio-dynamic software failed to show the clicks over time with these operations. The researcher marked positive clicks and negative clicks on the speech transcription during the rating to quickly pinpoint spots on the audio clip. At the end of each clip, raters provided a fluency rating using a nine-point scale (1=extremely disfluent, 9=extremely fluent). Immediately after the rating, an interview was conducted with each rater following Gass and Mackey's (2007) recommendations for stimulated recall research. The audio clip was played immediately after the rating in order to exploit recent memory and reduce recall interference. Raters were told that they could stop the audio clip at any time to share their comments, mitigating researcher's interference. The researcher waited for the raters to stop the video and provide comments, though the researcher asked specific follow-up questions. Raters were instructed to focus on their thoughts at the time they clicked upward or downward to indicate their rating. The researcher intervened with questions to avoid obvious spikes and dips in the ratings without comment (e.g., Can you tell me what made you click at this point?). The interview process was recorded using the computer application VoiceNote and transcribed (see Appendix F for an example of interview transcription).

The raters listened to the same audio clip again to track the raters' recognition of pauses because some raters perceived pauses which were not identified in the sound analysis program (Praat). They reported that excessive pauses led to fluency decrease. In order to pinpoint the pauses which decreased fluency, raters were instructed to click the button labeled "increase pause" when the raters heard a pause using the same Idio-dynamic software. Raters clicked and continued clicking the button if they perceived the pause becoming longer. Each successive click of the mouse corresponded to an additional increase, which indicated a longer pause.

At least one week after rating the audio clips, the raters watched the video clips and repeated the same rating process using Idio-dynamic software and a 9-point scale (1=extremely disfluent, 9=extremely fluent). The video clips were the original files which had been converted to audio files for audio-only assessment. Before rating the video clips, the raters practiced using the software to review the functions of Idio-dynamic Software. After the raters confirmed comprehension, the video clips were presented. As in audio testing mode, the video-recorded speech samples from the picture description task were presented prior to the video-recorded speech samples from the opinion task. The order of the speech samples in each category was randomized. Shortly after rating each speech sample, the video clip was played during a stimulated interview

with each rater. The raters were told to comment on anything that influenced their decision while watching the video clip. The researcher posed follow-up questions referring to graphs showing the timing, magnitude, and direction of mouse clicks extracted from the software. The interview process was recorded. Next, the raters watched the video clips again and marked pauses and perceived long pauses using the same software.

3.3 Data Analysis

In terms of raters' understanding of fluency as the target dimension, two questions were asked to individual raters: "What do you think is fluent speech?" and "What factors are considered when you judge fluency in a language test?" Most of the raters agreed that fluent speech means a natural and smooth speech without pauses while one rater (rater 2) mentioned immediate understanding. Rater 2 answered that fluent speech should be understood easily and immediately while it is being delivered but she added that speech is easily understood when the speech rate, intonation, and flow are natural. Regarding the question about the factors affecting their fluency assessment, all raters considered speech rate as an important factor and six of them (except rater 2) reported pauses as well. Three raters (rater 1, 2, and 3) also mentioned the importance of

natural accent and intonation for judging fluency.

For quantitative analyses, the timing, magnitude, and direction (upgrade, downgrade) of the rating activity from the Idio-dynamic software's data output were extracted and global fluency rating (1-9) was tabulated. The key feature of the Idio-dynamic software is that it allows users to provide ratings visualized as deviations from the baseline. The magnitude and direction (±5) over time were plotted on the graph and exported to an Excel sheet after each rating. However, the graphs sometimes failed to indicate whether the value was approaching 0 because of raters' clicking activity or a built-in auto-zero function of this software. For example, some raters clicked two different buttons ("increase fluency" and "decrease fluency") consecutively within one second. To decide whether the graph approaching zero was due to raters' clicking or not, the excel sheet displaying the number of clicks was used.

The global ratings were collected for comparison with the real-time assessments, which were the focus of this research. For the global ratings, interrater reliability was estimated using a two-way mixed, consistency intraclass correlation coefficient (ICC) and Cronbach's alpha. The consistency ICC of the raters reached 0.96 ($p < .001$) and Cronbach's alpha was 0.928. The summary of global ratings is provided in Table 1.

Table 1. Descriptive Statistics for Fluency Ratings (1-9 scale)

Task Type	Speaker1		Speaker2		Speaker3		Speaker4		Speaker5		Speaker6	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Picture	5.29	.95	4.71	.76	6.43	.79	8.29	.49	6.43	.98	4.71	.76
Opinion	6.43	.98	6.00	.58	5.43	1.13	7.86	.90	7.57	.53	3.14	.90

The audio recordings of the stimulated interviews were transcribed by the researcher. The responses were categorized based on common themes. Themes were coded from the transcribed comments.

CHAPTER 4. RESULTS AND DISCUSSION

This chapter discusses the major findings concerning the research questions based on the descriptive statistics and qualitative analyses of the collected data. First, real-time assessment profiles from the dynamic approach are presented in Section 4.1. Second, speech features influencing assessment of L2 fluency are described in Section 4.2 and the effects of different rating modes in Section 4.3. Finally, Section 4.4 gives a general brief discussion of the raters' characteristics in judging fluency.

4.1. Real-time Assessment Profiles

To investigate fluency assessment as a dynamic construct, which means continuous and time-sensitive processing of different linguistic features, individual rater data were inspected to determine the extent to which the raters' assessments were dynamic and whether their approach changed from audio-rating mode to video-rating mode.

Table 2 reports the number of clicks for each rater and Figure 1 illustrates the number and magnitude of raters' clicking activity.

Table 2. The Number of Clicks For Real-time Assessments of Fluency.

Raters	Audio			Video		
	Positive clicks	Negative clicks	total	Positive clicks	Negative clicks	total
Rater 1	269	197	466	300	167	467
Rater 2	119	156	275	87	67	154
Rater 3	79	189	268	56	116	172
Rater 4	37	103	140	58	124	182
Rater 5	146	158	304	92	61	153
Rater 6	76	87	163	126	32	158
Rater 7	193	58	251	139	49	188
Total	919	948	1867	858	616	1474

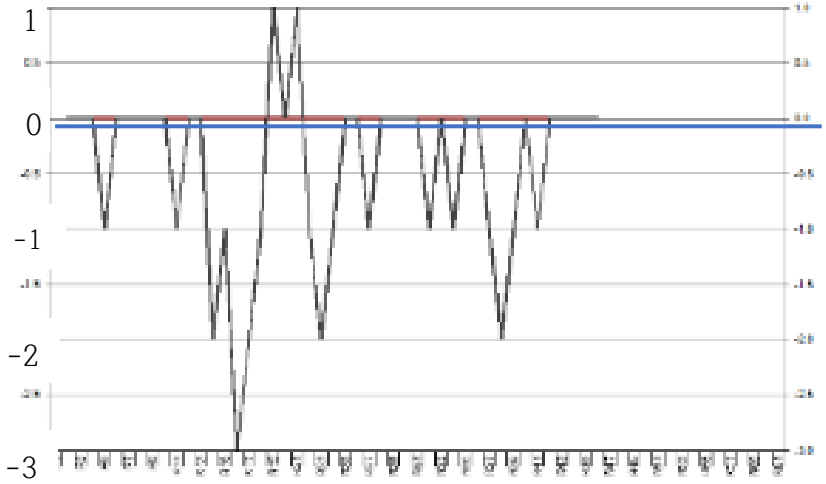
Following Nigel et al.'s (2019) categorization, the raters were classified as dynamic, semi-dynamic, or non-dynamic based on the frequency and magnitude of click activity. Nigel et al. (2019) explained that dynamic raters showed high click frequency and magnitude, semi-dynamic raters showed high frequency but lower magnitude, and non-dynamic raters low frequency and magnitude. Although click frequencies for semi- and non-dynamic raters partially overlapped, the semi-dynamic raters utilized a larger portion of the scale. For non-dynamic raters, ratings of ±1 were common.

Figure 1: The Example of Fluency Rating Graphs from Idio-dynamic software

<Speaker 2, Task 1>

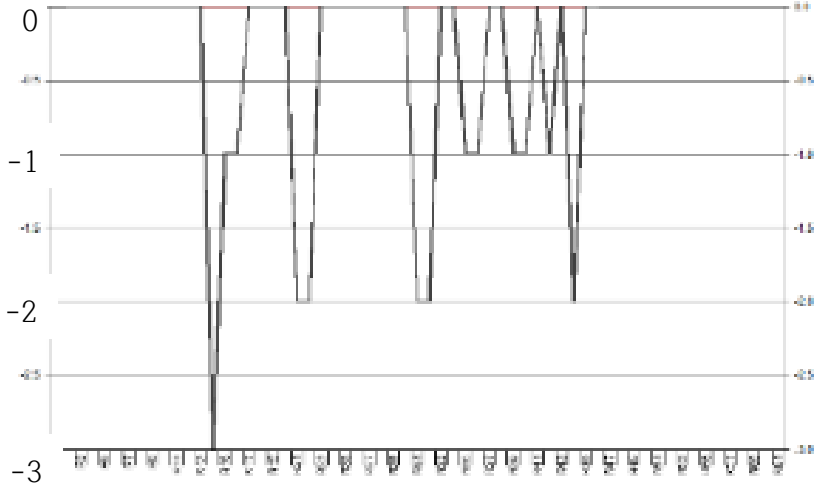
1. Audio

- Dynamic



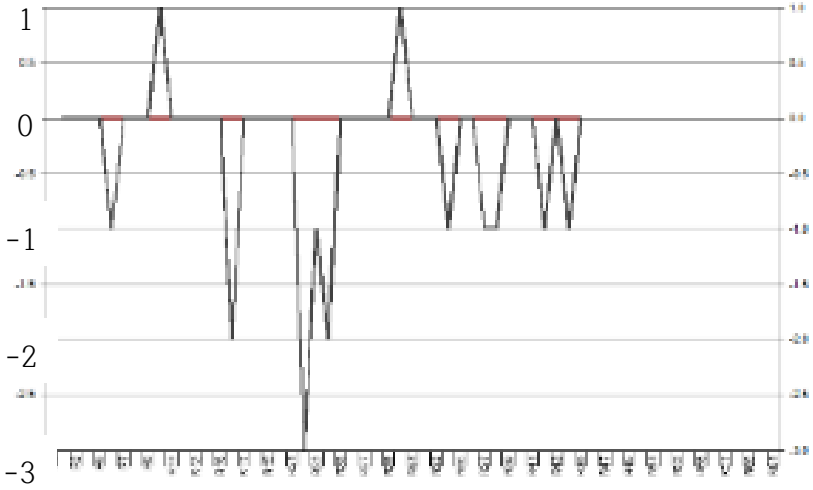
(Rater 1)

- Semi-dynamic

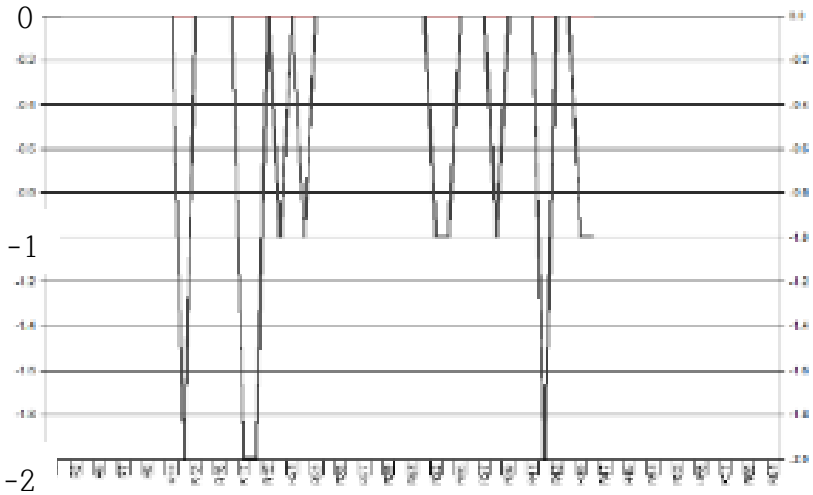


(Rater 2)

- Semi-dynamic



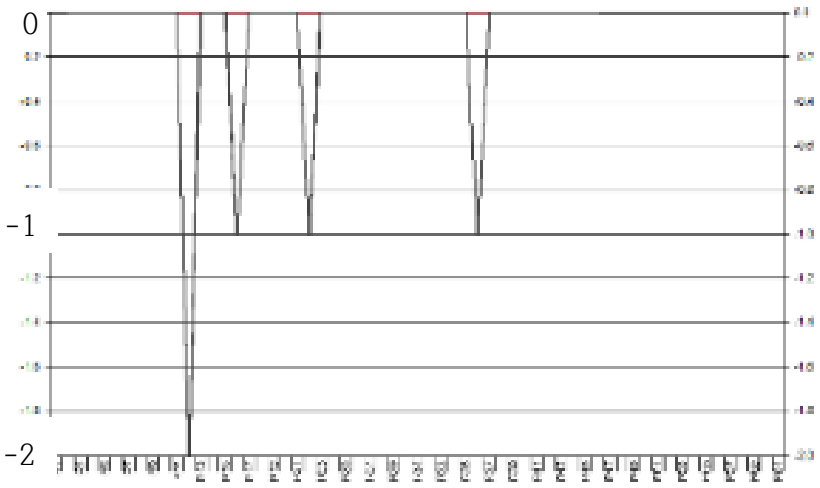
(Rater 3)



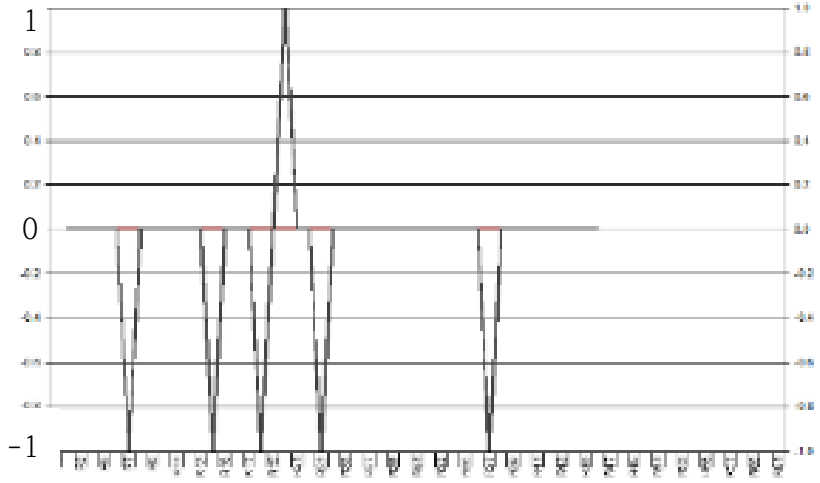
(Rater 5)

(Rater 2) (Rater 3) (Rater 5)

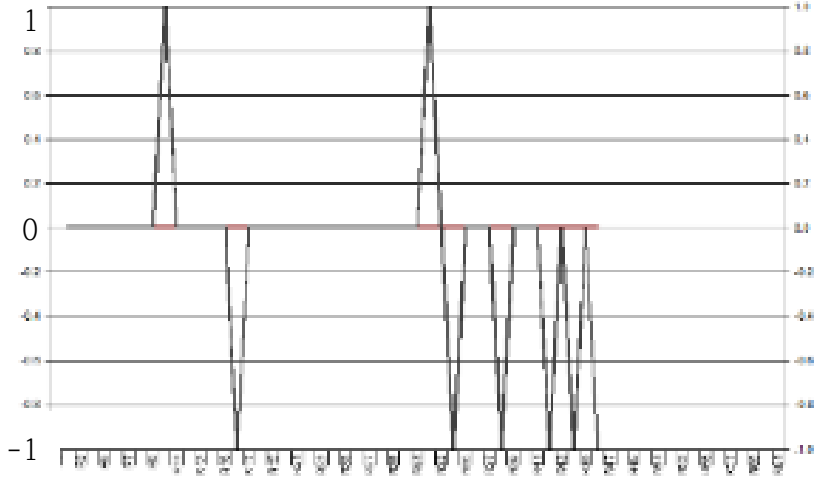
- Non-dynamic



(Rater 4)



(Rater 6)

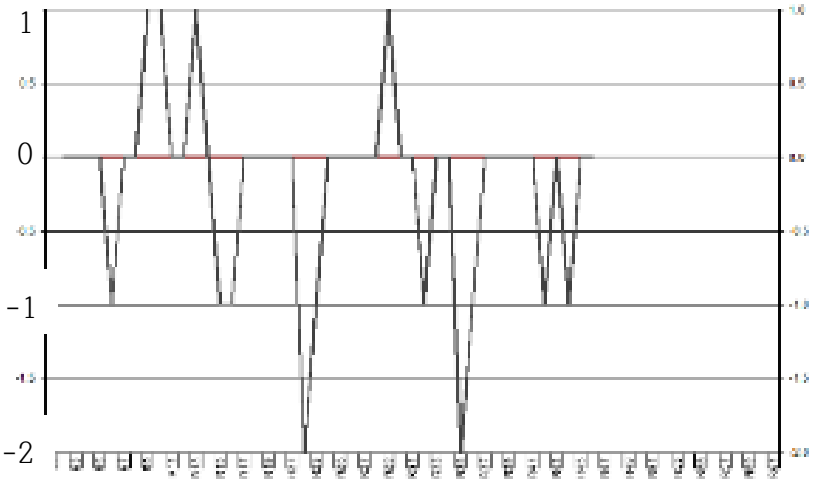


(Rater 7)

<Speaker 2, Task 1>

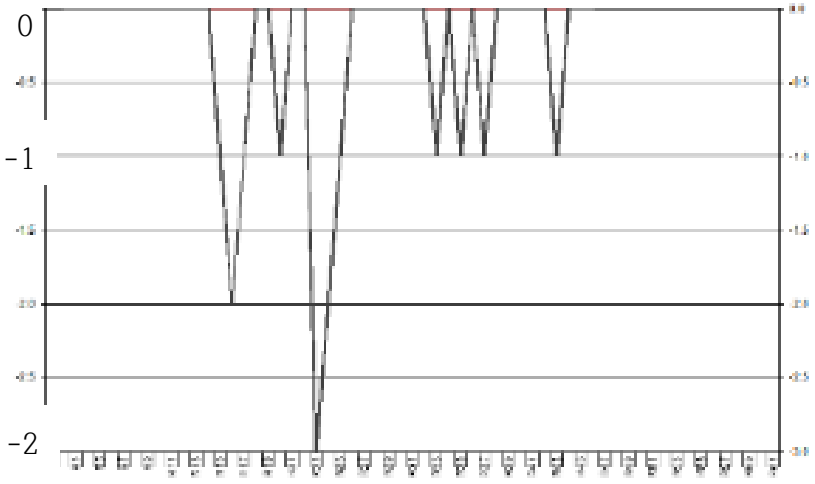
2. Video

- Dynamic



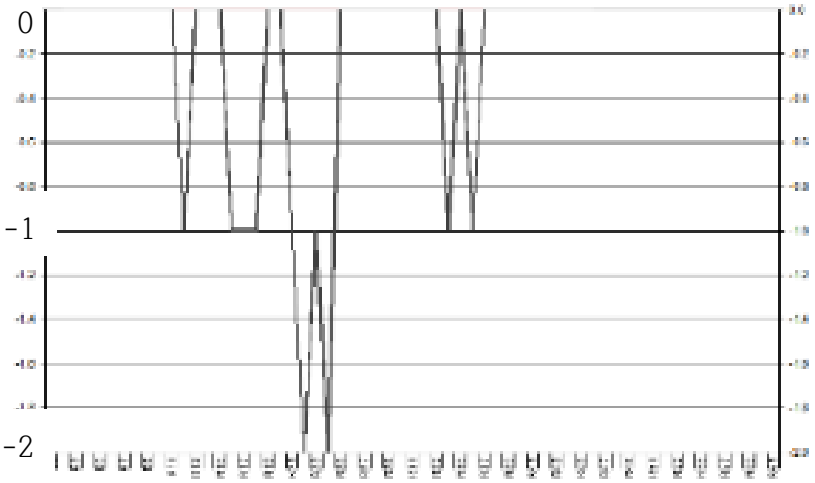
(Rater 1)

- Semi-dynamic

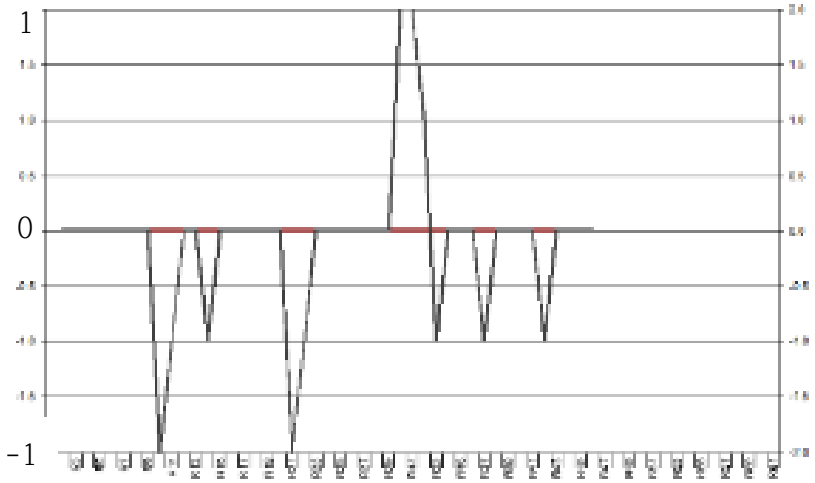


(Rater 3)

- Semi-dynamic

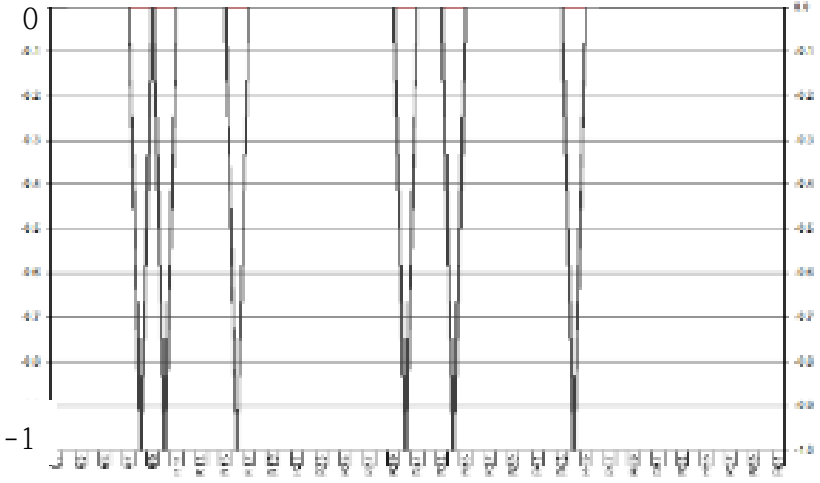


(Rater 4)

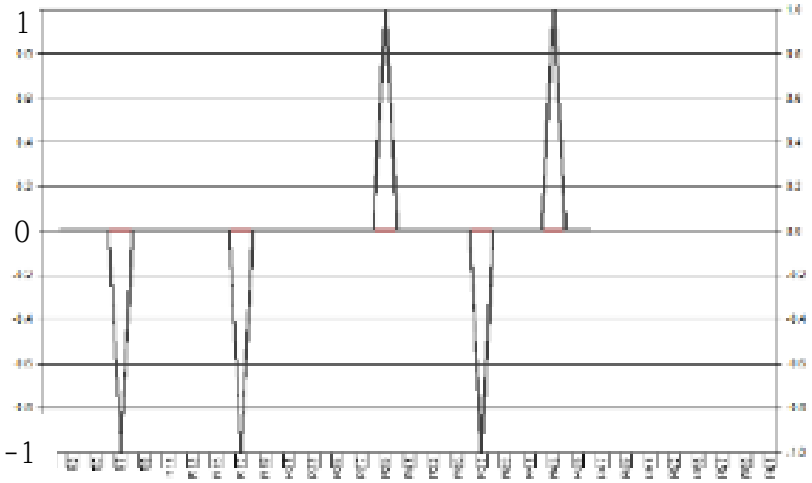


(Rater 5)

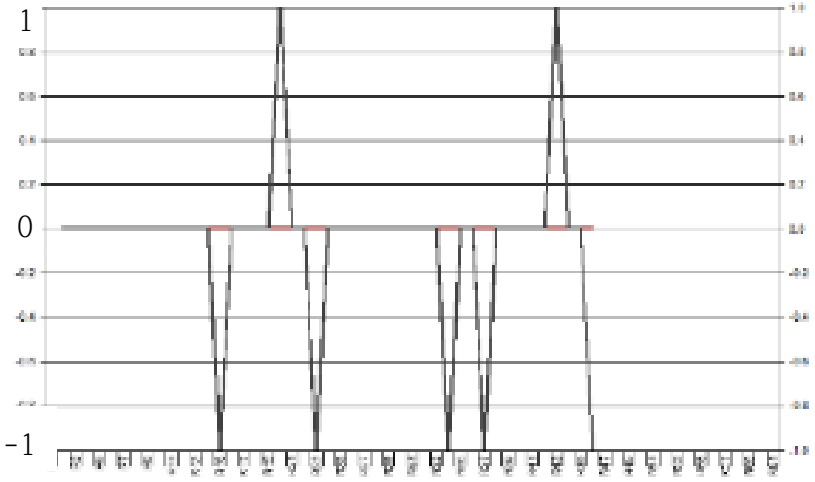
- Non-dynamic



(Rater 2)



(Rater 6)



(Rater 7)

Rater 1 is a dynamic rater who displayed higher click frequency and magnitude than other raters. She continuously evaluated fluency over the clips, upgrading or downgrading speakers. Consequently, graphs were characterized by high peaks and deep valleys for Rater 1. Rater 2, 3, 4, and 5 tend to be semi-dynamic raters, who displayed the same pattern of continuous ratings as Rater 1 but the magnitude of the click activity was less pronounced than Rater 1. Their ratings were typically centered on zero. Rater 6 and 7 are non-dynamic raters. The frequency of clicking is not significantly different from semi-dynamic raters but the magnitude of their rating is small (±1). In particular, Rater 7 did not upgrade or downgrade the speech at the precise moment. She rather waited for a sentence to be finished and clicked once at the end of each sentence. Although there is a difference in degree, most raters registered shifts in fluency evaluation as they listened to or watched the clips.

As for the two rating modes, the number of clicks in video rating was less than that of audio rating. According to the table, the difference in these figures appeared mostly in the number of negative clicks. Moreover, some raters' approach to the ratings was not consistent according to the testing mode. As shown in Figure 1, Rater 2 showed characteristics of a semi-dynamic rater when listening to speaker 2 but became non-dynamic while watching and evaluating the video clip of the speaker. Meanwhile, Rater 4 showed a completely opposite tendency

(from non-dynamic to semi-dynamic). Unlike previous studies that investigated the effects of test modes on test scores and reported no significant effects on scoring (e.g., Beltrán, 2016; Brown, 2005; Chalhoub-Deville, 1995; Fulcher, 2003; McNamara & Lumley, 1997; Zhou, 2008, 2015), a real-time fluency assessment is affected by test modes. Raters downgraded speech more frequently when they listened to the speakers than when they watched the video clip. Compared to one rater (Rater 4) who thought that her rating had turned stricter in video-mediated rating, four raters (Rater 2, 5, 6, 7) commented that they had become more lenient when they watched and assessed video clips, as illustrated by the following comment:

> *" I can see the students cheating. I didn't noticed it when I just listened to them. Looking at the video, I can see the students are reading the prompt on the screen. That's why some students sounded fluent when they started answering even though their overall English speaking skills were not good,"* (Rater 4)

> *"I think I'm less picky now. While I'm looking at speakers, I focus on what they're saying. I mean the content, not the errors. Also, I can see them smiling or rolling their eyes, so it's like I pay less attention to what they're saying, and I sometimes forgot the errors I*

caught." (Rater 2)

When Rater 4 watched the video clips, she noticed that the students were reading out the expressions presented as questions on the computer screen during answering. When some students answered quickly without hesitation while staring at the screen, Rater 4 suspected the students were reading rather than speaking the answer and left negative comments on them. On the contrary, four other raters (Rater 2, 5, 6, 7) admitted that they felt like having a conversation with the students when evaluating on video, so they seemed to focus more on the content and less on the speech errors.

4.2 Speech Feature Influencing Perceptions of L2 Fluency

4.2.1 General Patterns

To answer the research question concerning the linguistic features associated with the real-time assessment, stimulated interview comments were analyzed. Table 3 shows the major themes that emerged from the qualitative analysis and the frequency of comments provided per coded theme. The comments categorized as negative were associated with raters downgrading a speaker's fluency, while positive comments

were linked to raters upgrading a speaker's fluency. Although some issues pertaining to L2 speech assessment were identified such as comprehensibility (e.g., "I don't understand what she is talking about"), or task completion (e.g., "She didn't finish her answer."), the issues unrelated to fluency were excluded from the analysis. The detailed analysis focused on the speech features of speech rate, pauses, repair, and fillers (a sub-category of hesitant), which were frequently considered as factors influencing fluency.

Table 3. Frequency of Coded Comments in the Stimulated Interviews.

	Audio			Video		
Coded theme	Upgrade	Downgrade	Total	Upgrade	Downgrade	Total
Speech rate	57	14	71	78	26	104
Pause	0	203	203	1	194	195
Repair						
Correction	6	38	44	3	36	39
Repetition	0	29	29	0	25	25
Fillers	0	57	57	0	0	0
Automaticity	74	35	109	153	52	205
Accuracy						
Grammar	16	25	41	4	15	19
Pronunciation	18	101	119	13	72	85
Total	308	502	673	252	420	672

Two themes were additionally investigated in this study, which are automaticity and accuracy. The automaticity theme was adopted from the study by Préfontaine and Kormos (2016), in which they referred to automaticity as "Efficiency and Effortlessness." This theme, in the

context of L2 communication, refers to speaking ease or difficulty and underlying speech planning and processing efficiency. Segalowitz (2010) also assumed that the efficiency of speech, whether the process was easy and effortless, seemed to have a direct impact on listeners' perceptions of L2 fluency. Therefore, in the present study, raters' comments highlighting the fine balance among several features (i.e., speech rate, pause, intonation, lexical retrieval, and content organization) were counted in the automaticity category. Lastly, grammatical and pronunciation errors, usually considered a component of accuracy, were exceptionally investigated in this study. The reason for this focus was that error-related comments were not negligible, one of the characteristics of non-native raters as reviewed in chapter 2. Non-native teachers were more sensitive to L2 student-errors and stricter in assessing L2 speaking proficiency. In fact, some raters of this study insisted that understandable speech was a prerequisite for fluency, so fluency could not be dissociated from comprehensibility. Since the raters' real-time perception was important and could provide implications for further research, the researcher encouraged raters to share their comments on their ratings freely.

In total, raters made 673 and 672 comments in audio-only mode and video-mediated mode, respectively, during the stimulated interview. Raters most often referred to pause phenomena for their fluency ratings

in both rating modes, especially for downgrading speech samples. Only one comment indicated that pauses improved fluency in this study (i.e., "…, the locations of pauses are proper and make the speech go fast and smooth"). Comments pertaining to the automaticity and speech rate showed the opposite tendency. There were more positive comments for upgrading fluency in automaticity and speech rate categories. Between the two categories, automaticity was the second most popular theme and the sum of comments from both rating modes was 314. Interestingly, the number of positive comments on automaticity in video-mediated rating was almost twice as much as in the audio-only rating.

Self-repair and fillers are a form of hesitation often regarded as markers of dysfluency (Kormos, 2000; Tavakoli & Skehan, 2005). Comments for repair were counted according to the subcategories: correction (alterations of the original material before an interruption) and repetition (verbatim iterations of syllables or words). All comments about repetition were negative, while some comments about correction were positive. Comments for fillers also showed similar patterns to repetition in that they negatively affect fluency rating. Although the number of comments referring to speech delays is not as large as other themes, the interesting thing is that the raters' perceptions differ greatly between the two rating methods. There were 57 comments mentioning fillers in audio-only ratings but no raters mentioned fillers when they watched video clips.

These features will be discussed in the later section with raters' comments.

Lastly, accuracy was included among the themes despite not usually being considered a component of fluency. That is because accuracy had 264 comments, the third-largest number among all themes. The comments were counted under two subcategories: grammar and pronunciation. Grammar accuracy refers to lexical errors, morphological errors, and syntactic errors, while pronunciation accuracy refers to segmental errors, syllable structure errors, word stress errors, and rhythm. There were more comments on pronunciation (204) than grammar (60). The number of comments on grammatical errors or phonological errors was significantly reduced in the video clip ratings. Accuracy comments totaled 160 (34 upgrading and 126 downgrading) in the audio clip ratings and 104 (17 upgrading and 87 downgrading) in the video clip ratings.

An examination of the themes indicated that raters' perceptions of fluency are determined by a number of dimensions, not all of which are merely temporal. The raters' comments for each theme were investigated to uncover the relevance of specific dimensions to fluency and the linguistic processing experience of raters when they evaluate L2 fluency.

4.2.2 Speech Rate

Previous research has shown the importance of speech rate and the mean length of run as predictors of L2 fluency (Bosker et al., 2013; Cucchiarini et al., 2002; Derwing et al., 2004; Freed 2000; Freed et al., 2004; Ginther et al., 2010; Iwashita et al., 2008; Kormos & Dénes, 2004; Lennon, 1990; Préfontaine, 2013; Préfontaine et al., 2015; Riggenbach, 1991; Towell et al., 1996). In these studies, a faster speech rate and longer run-length were consistently related to higher fluency scores and levels of proficiency. Compatible with these findings, the qualitative data revealed that speed is a salient quality of speech perception.

From the perspective of speech perception, the raters expressed that slow speech is problematic because it does not catch the listener's attention.

1. *"Speech rate is too slow so I can't understand what she is saying. I just hear words one by one not the content. I even cannot remember what she said in the previous sentences."*

(Rater 6 on Speaker 2)

2. *"She speaks really slowly. It's boring or somewhat frustrating to wait for her to finish each sentence."*

(Rater 1 on Speaker 4)

In addition to comments on the speakers' speech rate, raters paid attention to speech changes and explained shifts in the L2 speakers' speed and flow.

3. *"The speech rate is good in this sentence. Her speech rate becomes faster and faster toward the end of the recording. The speaker seems to hesitate and drawl at the beginning but speaks more quickly now."*

(Rater 1 on Speaker 4)

4. *While upgrading: "Speech sound is good. It's natural and easy to understand."*
While downgrading: "She speaks too slowly. It doesn't sound fluent. It is strange because the speech rate was quite good before. Ah, she might read some part of the question prompted on the screen."

(Rater 4 on Speaker 6)

There were particular points at which raters perceived the speech to become faster. For example, when the speakers identified the location of people or items, especially in the picture description task, they used idiomatic expressions such as "on the right/left side of the picture" or "in the foreground/background of the picture." Raters upgraded fluency at points where they perceived the speech rate becoming faster. One

example is presented below. The length of the pause is provided in parentheses.

The excerpt from Speaker 2 is as follows:

"On the right side of the picture (1.16) there are some colorful flowers."

5. *"Speech rate is good and natural. I think it's because she speaks in chunks."*

(Rater 5 on Speaker 2)

6. *"She spoke pretty slowly, but the speech rate is good and intonation contour is natural here. It seems like she speaks this part in chunks."*

(Rater 1 on Speaker 2)

The term "chunk" was used by raters. Speaking in chunks refers to speaking in phrases, combining words to seamlessly connect them, rather than speaking words individually. When speaking in chunks, there tended to be fewer pauses and more words spoken between pauses. It seemed that the raters' perceptions of fluency were formed in relation to the number of words between pauses. Previous research has confirmed the importance of the mean length of runs measured quantitatively (Freed et al., 2004; Raupach, 1980, 1987; Towell, 2002; Towell et al., 1996).

The *chunking process* is further discussed in the general discussion section of this paper.

The speakers' speech rate was not mentioned independently. Raters described a speaker's speed or flow in relation to other fluency criteria, such as silent pauses, pronunciation, and grammar. Some raters tried to explain a slow speech rate with frequent or long pauses. Interestingly, most raters upgraded or downgraded fluency considering speech rate in relation to complex grammar structures.

The examples of upgrading are as follows:

7. *"She didn't speak slowly even though the sentence structure is complex."*

(Rater 2 on Speaker 4)

8. *"The sentence is long with a complex structure, but the speech rate does not become slow at all."*

(Rater 5 on Speaker 1)

9. *"I upgraded this part because the speech rate is good, and this is a long and grammatical sentence."*

(Rater 5 on Speaker 3)

10. *"The speech rate was good even though it is a passive form. Korean EFL learners feel difficulties using passive forms correctly."*

(Rater 6 on Speaker 5)

11. "*It is a quite difficult expression, but she speaks it in natural speech rate"*

(Rater 7 on Speaker 4)

The examples of downgrading are as follows:

12. *"This is a really simple sentence, but she speaks really slowly and hesitatingly."*

(Rater 1 on Speaker 3)

13. *"The sentence structure is simple, but the speech rate is slow.*"

(Rater 6 on Speaker 6)

14. *"She speaks slowly so she might have enough time to correct her answer, but she makes an ungrammatical sentence here."*

(Rater 7 on Speaker 1)

Raters were sensitive to the speech rate in relation to the sentence

structure. This does not seem to be related to their perception of fluency and speech rate but rather to their grading strategies. The raters should decide how to grade "ungrammatical simple sentences spoken at high speed" and "grammatically compound-complex sentences spoken at low speed." According to the raters' comments, they did not deduct points when students spoke slowly while producing sentences with complex grammatical construction. In contrast, they would give students lower marks when the students produced ungrammatical sentences despite their speech rate not being significantly slow. It seemed that raters attempted to balance fluency and accuracy as they graded the speech samples.

4.2.3 Pause

In the field of L2 fluency, a number of studies have used systematic research to examine the relationship between pauses and fluency using temporal variables within the quantitative paradigm (Derwing et al., 2004; Freed et al., 2004; Ginther et al., 2010; Kormos & Dénes, 2004; Lennon, 1990; Riggenbach, 1991; Towell et al., 1996). The majority of studies on L2 utterance fluency focused on the frequency, duration, and distribution of pauses. Disfluent speech, as indicated by excessive pausing, has previously been reported as one of the major impediments to L2

intelligibility and a source of negative perceptions of speech performance. Previous L2 English fluency perception research has reported that pauses within clauses, rather than between clauses, sounded disfluent to native speakers (Ejzenberg, 2000; Kahng, 2018; Pawley & Syder, 2000; Riggenbach, 1991; Wennerstrom, 2001).

In the present study, pause phenomena are categorized by their duration, location, and frequency. Comments could be as simple as, "there are too many pauses in this sentence," or "it's a long pause." Additionally, raters combined two or more pause problems, such as in the example below.

The excerpt from Speaker 1 is as follows:

"And (0.61) maybe your (1.21) dream (0.89) could be changed."

15. *"A pause between 'your' and 'dream' is not proper at all. This pause is long as well."*

(Rater 4 on Speaker 1)

The frequencies of each category are presented in Table 4. The most common comments were about the length of pauses followed by the location of pauses and finally the frequency of pauses. There was not

much difference in comments according to the rating modes but the type of tasks influenced raters' perception of pauses. There were more comments concerning pause location in the picture description task and more comments concerning pause frequency in the opinion task. It seems that the difference is due to the characteristics of each task. In the picture description task, all information was provided in the picture; thus, raters were more sensitive to the pauses following simple or easy words.

Table 4. Frequency of Comments on Pause Duration, Location, and Frequency.

	Audio				Video			
Task Type	Duration	Location	Frequency	Total	Duration	Location	Frequency	Total
Picture Description	79	25	9	113	72	19	12	103
Opinion	84	12	29	125	99	20	22	141

The excerpt from Speaker 1 is as follows:

"Two women (1.16) is wearing (0.35) glasses. …There are many trees (0.33) and (0.39) people."

16. *"Pause after 'wearing' is not necessary at all. 'Glasses' is not a difficult word. 'Wearing glasses' is more like a phrase, isn't it? … There are no difficult words or structures in this sentence. I don't understand*

why she puts pauses before saying 'people.''

(Rater 1 on Speaker 1)

The raters' comments revealed that they are sensitive not only to the location of pauses but also to the purpose of pauses. Raters downgraded fluency when they heard unexpected pauses while speakers searched for the correct vocabulary. When the vocabulary was easy and consisted of commonly-used words, the raters concluded that the pause was unnecessary and improper. They downgraded the speech fluency. Below are two examples relating to a pause and its perceived purpose. Interestingly, all seven raters downgraded fluency when they heard these pauses in the picture description task.

The excerpt from Speaker 5 is as follows:

"I can see various (0.36) and colorful (1.97) flowers."

17. *"The pause between 'colorful' and 'flowers' is awkward. The location of the pause is not appropriate, and it is long at the same time. She is describing a picture and flower is not a difficult word. She must know this word. Since there is a pause in the location searching for such an easy word, it sounds disfluent."*

(Rater 1 on Speaker 5)

18. *"There is a pause between an adjective and a noun. They are really close because this adjective modifies this noun. 'Colorful flowers' is not a difficult expression, but a long pause is inserted. It seems like the speaker's proficiency problem."*

(Rater 2 on Speaker 5)

19. *"I don't understand why the speaker pauses here. She must be looking at this picture and describing it. So, there's no need to think of the next word, an easy word like this."*

(Rater 7 on Speaker 5)

The excerpt from Speaker 1 is as follows:

"One woman (1.74) wore (0.58) sky blue short sleeved shirt."

20. *"There are pauses and both are long. The flow is choppy in between words even though this sentence is not complex, and she doesn't need to think of difficult words."*

(Rater 1 on Speaker 1)

21. *"The pauses are too long for a simple sentence like this. Words and expressions are not difficult. It seems like her proficiency level is*

low."

(Rater 2 on Speaker 1)

22. *"This is a simple sentence with no difficult expressions but there are pauses. When the simple sentence seems to require a lot of elaboration, I regard the speaker as low-level L2 learner, and I cannot give her a high score."*

(Rater 5 on Speaker 1)

23. *"Even though the speaker pauses here, long pauses, she is using the wrong tense here. If there were no pauses, I could ignore the tense error. Usually, after pauses, I expect correct sentences. She might be not a fluent speaker."*

(Rater 6 on Speaker 1)

As is evident in these examples, raters determined the speakers' L2 proficiency levels according to the purpose of the pauses. Once the speaker's L2 proficiency level was ascertained to be low, the raters were reluctant to attribute a high score in the fluency assessment. In fact, pause phenomenon is closely related to the L2 development stage (Cenoz, 1998; Izumi, 2003; Kormos, 1999a, 2006; Swain, 1985). Cenoz (1998) observed an increased frequency of unpredictable pauses (non-juncture pauses) in the case of lower proficiency speakers. Kahng (2014) reported

that the lower proficiency speakers paused or hesitated because of grammar and vocabulary more often than the higher proficiency learners. Therefore, raters had the general understanding that pauses occur when speakers experience processing difficulties, resulting in disfluency.

An absence of comments indicated that pauses improve fluency in this study. Comments pertaining to pauses implied that fluency improved as the number of pauses decreased. Some comments did pertain to pauses perceived but ignored; these were not counted, because they did not affect the fluency score. However, listeners' perception and understanding of pauses were well illustrated in these comments.

24. *"It is ok to have a pause before new information or unfamiliar words. I can ignore pause when students give new information like words or expressions that we don't usually use."*

(Rater 1 on Speaker 4)

25. *"I can wait when the speaker says, 'first of all' or 'for example' and then pause a little bit longer. That's because I expect something is coming up soon. When the students stop speaking for a quite long time, I expect their speech is over. Then suddenly they start speaking again and I put effort to remember what they said before. This kind of situation is*

more bothering."

(Rater 1 on Speaker 2)

Per Rater 1's comments above, when listeners heard a pause, they anticipated an upcoming short or long delay (FoxTree, 2001), new information rather than given information (Arnold et al., 2003, 2004), and an unknown object rather than a known object (Arnold et al., 2007). Listeners seemed to understand the role of pauses in the speech. Planned and grammatical pauses which generally occur at the boundary of a clause due to the need to parse and plan the sentence, do not downgrade fluency. However, ungrammatical and unplanned pauses which occur at inappropriate locations indicate a breakdown in composing the speech stream as planning, production, and lexical access are disrupted (O'Shaughnessy, 1992; Rochester, 1973). Therefore unplanned or unexpected pauses may cause raters to downgrade fluency.

There was one interesting speech sample which belonged to Speaker 4. Examined with waveforms and spectrogram using Praat, her speech did not show many pauses. However, raters perceived many small pauses between words in her speech sample, which resulted in raters' downgrading her fluency rating. This example is provided below. Parentheses without the pause length represent the pauses perceived by the raters but not observed by the sound analysis program.

The excerpt from Speaker 4 is as follows:

"I think (0.29) for ( ) a university student (0.66) it's ( ) better ( ) to ( ) spend (0.5) a long vacation (0.6) doing ( ) an internship."

26. *"There were too many small pauses in her first sentence, and as a result, her speech rate was very slow. She seems to read the question on the monitor so it may not need a lot of preparation. But pauses occur very often in this sentence."*

(Rater 1 on Speaker 4)

27. "*I don't know why many small pauses were inserted in her speech. Because of the frequent pauses, it is like staccato. She speaks too slowly. Her speech sounded prolonged at the same time."*

(Rater 3 on Speaker 4)

Both raters pointed out that frequent pauses resulted in a slow speech rate despite some pauses not being observed. However, it seemed that a slow speech rate caused the raters to perceive pauses which were not detected in waveforms or spectrograms. This misperception is also related to a failure of "connected speech." In comments on other speech samples with excessive pausing, failure of connected speech due to slow speech rate was mentioned.

The excerpt from Speaker 1 is as follows:

"Many books (0.17) are piled up (0.39) on the table. (1.05) Four people (0.78) are standing around the (0.63) book."

28. *"This speaker speaks slowly. She pronounces words in isolation. Every single word is stressed so the intonation is also unnatural."*

(Rater 1 on Speaker 1)

29. *"Her intonation is awkward. Each word finishes with falling intonation. It seems that she pays attention to articulation so there are no linking sounds."*

(Rater 2 on Speaker 1)

There is often a significant difference between a word's pronunciation in isolation versus its pronunciation in connected speech. Connected speech is spoken language in a continuous sequence, as in normal conversation. In connected speech, words or syllables are clipped, phrases are run together, and words are stressed differently than they would be in isolation. In English, the stress pattern of a word is generally influenced by its context.

In Speaker 1's speech sample, extra pauses indicated clear-cut borders between each word and interrupted connected speech. Likewise,

careful articulation of individual sounds could lead to the same result. When speakers speak slowly, they tend to pronounce each word without linking. For example, when one word ends with a consonant and the following word begins with a vowel, the sounds are often linked. In Speaker 1's speech sample, pauses disturbed linking (e.g., between "books" and "are," "up" and "on," "people" and "are"). In Speaker 4's speech sample, raters perceived pauses between the words "for" and "a university," as well as "doing" and "an." Another reason for the absence of connected speech is that each syllable in individual words seems to be stressed. For instance, a reduced vowel such as in "to" was stressed in speaker 4's speech, which made raters judge that pauses were in between words. (e.g., "… it's ( ) better ( ) to ( ) spend (0.5) a long vacation…"). As described above, it seems that a slow speech rate influenced some raters' pause perception.

4.2.4 Self-repair

Earlier research examined the association between fluency and repair, comparing the fluency rating and the number of repairs (e.g., the number of corrections per minute and the number of repetitions per minute). However, the findings are conflicting. Several studies (e.g., Cucchiarini et al., 2002; Kormos & Dénes, 2004) indicated that repair

fluency was not a good predictor of fluency. Conversely, Bosker et al. (2013) reported that repairs were found to add a small but substantial amount of explanatory power to perceived fluency. In Kahng's (2014) study, only repetitions showed a weak negative correlation with speaking scores. In the present study, the qualitative analysis showed a similar tendency to the previous quantitative research. All repetitions correlated to downgrading fluency but not all corrections correlated.

Repetitions had a negative effect on fluency ratings. Some raters' comments were simple: "She repeated this part, so it did not sound fluent." Some raters provided specific justifications for deducting the fluency score.

30. *"The speaker repeated the same word, so it sounded like she stuttered."*

(Rater 3 on Speaker 2)

31. *"There's no reason for repetition. It's like a rehearsal for what she's going to say. Definitely not necessary."*

(Rater 1 on Speaker 1)

32. *"She repeated the subject of the sentence several times. I think she needs some time to think of what to say and how to make the next*

sentence."

(Rater 7 on Speaker 2)

Rater 7's comment reflects those of Tavakoli et al. (2020), who claim that verbatim repetitions do not necessarily reflect repair behavior but rather indicate a breakdown, as a speaker may employ repetition to buy time. The role of repetition in this sense is similar to a pause rather than reformulation. In fact, speakers inserted pauses before and within repetitions. Repetition seemed like a component of silent or filled pauses allowing the speaker some more time for speech planning.

Self-corrections, however, were appreciated, and the raters reacted favorably to speakers making an effort to self-correct.

33. *"She makes self-correction here. But it is OK, because she produced a grammatically correct sentence anyway."*

(Rater 7 on Speaker 3)

34. *"If the speaker has a grammatically correct sentence after self-correction, it's totally fine with me. I don't want to deduct the fluency score because this is a natural process. It means she is conscious of her grammatical errors."*

(Rater 6 on Speaker 6)

The raters indicated that fluency decreased when the speaker repeated self-corrections and when the speaker produced sentences with errors after self-correction. One of the reasons is that self-corrections in this study were closely related to linguistic errors. As the present study only dealt with monologic oral performances, self-corrections in this study were closely related to phonological and grammatical errors. Therefore, raters often associated the frequency and the result of self-correction with the speaker's overall L2 proficiency.

35. *"She makes a lot of self-corrections in her speech. I didn't downgrade when I heard first one or two because she could correct the sentences anyway. But I think there are too many self-corrections here. Clicking here is actually a result of accumulation of previous self-corrections."*

(Rater 2 on Speaker 3)

36. *"The speaker puts so much effort into correcting grammatical errors, but she still has ungrammatical sentences. She also seems to doubt her own corrections. She must be a beginner level L2 learner."*

(Rater 6 on Speaker 6)

37. *"There's a trial for self-repair but she cannot choose right vocabulary and expressions for this sentence. She may not have enough*

proficiency to make English sentences for this kind of task. So, her fluency score cannot be higher than this."

(Rater 5 on Speaker 6)

The raters mentioned pauses in their comments about self-repair (e.g., "I heard pauses and repetition here" or "There were long pauses with self-correction"). Several studies reported a connection between these two disfluency types with L2 proficiency. In a section headed "Reformulation pauses," Tavakoli (2011) referred to the importance of the compound occurrence of pauses with other forms of disfluency, such as repairs. She noted that such symptoms of disfluency are mutually interactive, and planning takes place during the pauses prior to the start of a reformulation. There have been calls for a better understanding of the relationship between self-repair (number, location, and structure) and general speech performance through the study of pause behavior (Kormos, 1999a). Ejzenberg (2000) associated lower proficiency L2 speakers with more pauses, and she reported that the corrections and false starts of those speakers resulted in intra-clausal repetition that sounded like "debilitating hesitation" (p. 302). Riggenbach (1991) noted that silent pauses are mostly found within clusters of disfluencies that comprise repetitions and false starts as repair; non-fluent speakers produced more of such clusters than fluent speakers.

Several studies have examined the duration of reformulations (Plug & Carter, 2014; Van Hest, 1996). For example, Van Hest (1996) reported that false starts take significantly longer to produce than corrections, a finding confirmed by Kormos (2000b). This relationship remained true when comparing L1 and L2 repairs (at all proficiency levels). When participants speak in their L2, the duration of their false starts is longer than when speaking in their L1. Previous studies measured duration of repairs and noted that pauses occurred at the start of reformulation. However, little research has analyzed the direct relationship of pauses and reformulations and their dependence on fluency. Speakers' production of pauses inside reformulations may be an indicator of the relationship or reformulations to proficiency level. Williams and Korko (2018) quantitatively examined pauses produced inside two reformulation types with reference to proficiency level.

Since the sample size of speakers is small in this study, the relationship between pauses (frequency, location) and repairs could not be analyzed. Only 10 repetitions and 10 corrections occurred, and the raters' comments merely noted their occurrence. However, as previously noted, pauses occurring before and within repetitions seemed to behave like filled pauses, which are typically employed to buy time and are negatively associated with fluency ratings.

4.2.5 Fillers

Fillers are known as “discourse markers” (e.g., “you know” or “so”) or “filled pauses” (e.g., “uh” or “um”). Filled pauses are defined as lexical (e.g., “well,” “like,” “you know,” or “actually”) or non-lexical voiced utterances (e.g., “uh” or “um;” Riggenbach, 1991) and prosodic markers, like laughter and sighs, that interrupt the stream of speech (Schüller et al., 2013). Fillers are considered relevant to cognitive function in speech planning. When speakers have difficulty with time constraints underpinning their speech planning and execution, they are likely to use filled pauses (Clark, 2002).

Raters easily noticed the fillers because they are often preceded, and may be followed, by a silent pause (Beattie, 1977). When the raters heard “uh” or “um” with a silent pause, they reported a “long pause.”

38. *“She couldn’t start her answer immediately. Pause here is too long with ‘uh’ and ‘um’.”*

(Rater 2 on Speaker 3)

Raters were sensitive to the functions of fillers, and they easily found lexical fillers. When the listeners heard fillers, they assumed that the speaker was trying to process information that might be difficult or

complex for them (Clark, 2002; Shriberg, 2005; Stenstroem, 1994). When speakers were unfamiliar with what they were talking about, they used fillers. These fillers indicated that the speaker planned to continue speaking even if he/she paused for a moment. In this experiment, when the speakers prolonged certain words or phrases several times, raters immediately analyzed the reason for the prolongation and repetition. They recognized fillers as vehicles for planning speech or searching for the appropriate linguistic forms.

The excerpt from Speaker 1 is as follows:

"And the internship (0.24) is (1.49) maybe (0.25) uh (0.66) the internship maybe(0.44) is not essential to your job. … And (0.61) maybe your (1.21) dream (0.89) could be changed."

39. *"I downgrade here because of frequent pauses and long pauses. Besides, she may use 'maybe' as a filler. When she doesn't know what to say, she used maybe and buys time. … I downgraded because I heard 'maybe' again, a repeating filler here and long pauses."*

(Rater 3 on Speaker 1)

40. *" 'Maybe' here sounds like a filler. She frequently uses maybe, and it doesn't sound fluent*."

(Rater 4 on Speaker 1)

One rater exhibited interesting behavior in relation to fillers. Rater 5 distinguished non-native fillers from native-like fillers and rated their use as incompetent or highly proficient L2 learners' characteristics. When the rater heard non-native fillers which Korean L2 learners prefer to use, she named them "Korean's English discourse markers," and selected "decrease fluency." She explained that discourse markers could be categorized according to L2 proficiency development and there were several discourse markers reflecting poor language skills. She added that the speakers' proficiency level as well as the fillers themselves influenced fluency assessment.

41. *"There are some discourse markers showing the learner's English proficiency. For lower level L2 learners, typical ones are 'and', 'but', 'so', 'I think', 'because', and 'maybe'. …She said, 'in my opinion' and 'I believe' instead of saying "I think". She must be a high level L2 learner."*

(Rater 5 on Speaker 1)

There has been significant research examining the effects of fillers on listeners' comprehension and the results were discrepant. Discourse markers and filled pauses may help non-native listeners' comprehension of speech (Blau, 1991) and compensate for disruptions and delays in speech (Brennan and Schober, 2001). Conversely, some

researchers have argued that fillers are primary obstacles to a listener's perception and comprehension of speech (Voss, 1979). Per the present study's results, fillers are viewed as negatively impacting fluency ratings like silent pauses.

4.2.6 Automaticity

The raters' comments encompassed a wide range of features, such as speech rate, pausing, pronunciation, and grammar. It seemed that fluent speech skillfully balanced these features. The raters' explanations for upgrading fluency simultaneously cited multiple categories as reasons, including speech rate, pauses, and grammatical and phonological accuracy.

42. *"The sentences are long, but all are grammatical. Speech rate, pause, intonation are all natural."*

(Rater 7 on Speaker 4)

43. *"As it goes on, she produced longer sentences and they are grammatically correct. Speech rate and pauses are all good."*

(Rater 6 on Speaker 1)

44. *"I like the expression, especially word choices. Her pronunciation is good and intonation as well."*

(Rater 5 on Speaker 1)

Raters judged the speakers' proficiency level and fluency at the start of the speech samples. Proficient speakers spoke long grammatically-correct sentences with few pauses and no repairs at a good speed. Moreover, little attention and effort is needed for proficient speakers to produce fluent speech; the production process seemed automatized. Automaticity refers to the absence of attentional control in executing a cognitive activity (Kahneman, 1973) and includes several characteristics, such as rapidity, effortlessness, and unconscious and ballistic nature (Segalowitz & Hulstijin, 2005). Kormos (2006) pointed out that while L1 speech production only requires focus on speech planning and monitoring, L2 speech has not fully automatized syntactic and phonological encoding, slowing down speech. Per the raters' observations, the psycholinguistic process of speech planning and encoding is particularly salient. This is because their effortlessness seems to directly impact raters' perceptions of L2 automaticity, as demonstrated by the qualitative perceptions.

With regard to automatized production, three patterns emerged from the qualitative data. First, automaticity is represented as speakers' L2 proficiency level in the raters' comments, as shown in the below

examples.

45. *"It seems like she can make only simple sentences. I clicked 'decrease' because she started sentences using the same phrase, 'some people are' several times. She must be a low-level learner so I cannot give her a high score."*

(Rater 2 on Speaker 2)

46. *"Most sentences started with 'there are.' It may be hard for her to make other types of sentence structures. She uses simple sentences repeatedly and seems to need more effort to make complex sentences. 'There is' or 'there are' is a typical expression usually used by low-level learners."*

(Rater 5 on Speaker 2)

In the above examples, the rater describes the speaker's sentence structure in delivering L2 speech. As we have seen, raters tend to underestimate speakers' L2 proficiency when the sentence structure is simple and the same structures are repeated in speech. The speakers' linguistic abilities immediately affected the raters' judgment of fluency. Second, the overall organization of speech was a feature that the raters considered when reflecting on automaticity. Regardless of temporal features or grammatical errors, raters upgraded speech when the

answers were well-organized and the topic was developed logically and coherently, as shown in the below examples.

47. *"The description of the picture is well-organized. She described the picture by giving details. Overall, there's no ambiguous part. Her description was clear and easily understood."*

(Rater 5 on Speaker 5)

48. *"There are no salient errors, actually. Her speech sounds fluent. But there is no story. Especially this part lacks relevance. I cannot associate 'new and touched,' with what she said before."*

(Rater 7 on Speaker 1)

49. *"She abruptly states her opinion and experience here. It's awkward and unexpected. This part is not necessary and not relevant. She might not have an idea, so she is just filling time. There should be a bridge sentence."*

(Rater 2 on Speaker 3)

From the raters' comments, it appears that discourse structure is the speech feature underlying perceptions of fluency. These findings are reminiscent of Song (2017), where non-native English raters determined

fluency scores considering other factors such as task completion, utterance volume, or the level of the test concerned. Especially, raters judged fluency level in relation to task completion levels which evaluated the appropriateness of the content or the degree of performance of the task.

Third, the raters considered naturalness a sign of automaticity. When upgrading speech, raters often explained their reason citing the speech as "generally natural." In these instances, the researcher posed followed-up questions asking what made the speech seem natural. Every feature was mentioned, such as good speech rate, few pauses, no hesitation, grammatical sentences showing the complex structure, and pronunciation with very little Korean accent. Comments pertaining to this set of features indicating naturalness were counted and categorized under automaticity as shown in the comment below.

50. "*Speech becomes natural because of chunk processing. Because of chunks, the locations of pauses are proper and make the speech go fast and smooth.*"

(Rater 1 on Speaker 4)

Chunks are groups of words found together in language. Chunks can be words always paired together, such as fixed collocations, or words commonly paired, such as certain grammatical structures following

language rules. According to usage-based accounts of L2 acquisition, language is formulaic in nature, with language exemplars stored as "chunks" in the mental lexicon (Bybee & Hopper, 2001). In the speech samples from the picture description task, fixed phrases to describe location were commonly used (e.g., "on the right side of the picture," "in the background of the picture"). Some raters even upgraded a speech sample of the low proficiency speaker (Speaker 2) when they noticed these expressions mentioning naturalness and speech rate.

For the "natural" and "effortless" part, all raters mentioned "chunk speech" or "chunk processing" and attributed the appropriate chunk speech to high scores. The researcher asked raters why they placed emphasis on "speaking in chunks." Examples of the raters' comments are shown below.

51. *"Chunk processing is important. The speech rate goes up when you speak fluently in chunks. Plus, the speech rate doesn't slow down as there are no unnecessary pauses. When one speaks in chunks, the accent becomes natural and connected speech is possible, so the speaker doesn't slow down."*

(Rater 1 on Speaker 4)

52. *"When the speakers speak in chunks and their speech rate is fast, I judge them as fluent speakers. Speakers quickly go over the parts which*

aren't important, and they don't stress every word. They are well-informed on English structures and expressions. And they can automatically make English sentences."

(Rater 2 on Speaker 4)

"Chunks" are frequently referred to by the raters when upgrading the speech sample. Chunk speech reduces unnecessary pauses in the middle of clauses. A small pause usually follows each chunk. Chunks provide a pleasing sentence rhythm with small pauses. In addition, silent pauses at grammatical boundaries help listeners comprehend the intended meaning (Arons, 1993; Bower & Springston, 1970; Griffith, 1991; Lass & Leeper, 1977; Reich, 1980; Sugito, 1990).

Moreover, pauses between chunks are different from "hesitant pauses." Hesitant pauses are related to delays in speech planning and production processes and can occur when a speaker needs to plan an upcoming speech or encounters difficulty. However, pauses between chunks are prosodic pauses (Ferriera, 1993, 2007), which separate utterances into intonational phrases (i.e., a speech segment which occurs with a single prosodic contour), and thus are part of the rhythmic structure of speech. Indeed, in L1 speech, most pauses tend to occur at clause boundaries (junctures) (Boomer, 1965; Hawkins, 1971; Holmes, 1988; MacGregor, 2008). Expressly, listeners use their knowledge of chunks to help them predict meaning and therefore are able to process

language in real-time. Accordingly, breaking speech into short chunks is an essential component of speaking fluently.

Using chunks seems to be a matter of automaticity and is closely related to L1 production. Lìeven et.al. (1992) estimated that 20% of the speech of young English-speaking children is "frozen phrases," the phrases of words that tend to come together in a single chunk. They suggested that children remember and use several words as a single unit. Erman and Warren (2022) reported that multiword combinations constitute approximately half of written and spoken English. These "formulaic sequences" (Wray, 2005) such as idioms and multiword expressions are particularly characteristic of oral discourse among native speakers (Biber et al., 1999) In this respect, chunk speech may be deemed similar to L1 speech, considering the degree of automaticity.

In addition to being associated with automatized speech, naturalness seemed to be related to the speech sounding genuine, which is suggestive of authenticity. It is consistent with giving one's full and undivided attention to the person or matter at hand (Préfontaine & Kormos, 2016), as shown in the examples below.

53. *"For the last part of the clip, I felt like she was telling a story. It appeared that the information was delivered comfortably and clearly."*

(Rater 1 on Speaker 3)

54. *"The speaker guesses and explains the situation behind the picture as well as describes it in detail. I think, she has a language skill such as inference that only advanced learners have."*

(Rater 2 on Speaker 1)

55. *"To support her opinion, she is talking about her experience of internship. It is natural and her idea is easily understood."*

(Rater 5 on Speaker 2)

According to the data, demonstrating diverse expressions, organizations of speech or coherence, and naturalness seemed to be important in perceiving L2 automaticity. However, these three qualifiers often appeared as general descriptors (e.g., language use or topic development in the TOEFL iBT speaking rubric) on speaking tests rather than indicators of fluency. As illustrated above, raters seemed to judge fluency according to external factors which were not generally included in the fluency feature.

4.2.7 Accuracy

Perceptions of fluency are formed not only in relation to temporal features but also in relation to grammar. Accuracy refers to the ability to

produce error-free language (Foster & Wigglesworth, 2016; Polio & Shea, 2014). Accuracy is largely associated with learners' linguistic knowledge representations, whereas fluency is a performance phenomenon, typically defined as the ability to produce smooth and eloquent speech.

Among six raters, three were sensitive to grammatical or phonological errors while two were not. One rater was sensitive to pronunciation but not to grammatical errors. Rater 2, Rater 6, and Rater 7's fluency judgments were associated with non-temporal features of L2 speech. During the stimulated interviews, the researcher asked for further explanation in cases where rater's comments were focused on accuracy more than fluency. The raters' answers to the researcher's question were as follows

56. *"It is difficult to grade speakers' fluency separately from their proficiency or comprehensibility. If the meaning isn't delivered well and if there is no comprehensibility, then I believe it is impossible to evaluate fluency. … Fluency cannot be highly evaluated if they speak quickly without delivering the content accurately."*

(Rater 2 on Speaker 1)

57. *"When I say that I understand what students are saying, …that*

means they must not have any grammatical errors. It is difficult to comprehend an ungrammatical sentence. If I wonder what they are saying, then I think the process is slow, and that also makes me think they aren't fluent. Then, the pronunciation error can cut points for fluency."

(Rater 6 on Speaker 6)

58. *"I also can't say they are fluent when they speak naturally but can't deliver the content properly due to grammatical mistakes, wrong pronunciation or making wrong choices for words. Comprehensibility is a prerequisite for fluency."*

(Rater 7 on Speaker 1)

Their ratings essentially relied on their own standards and rules for evaluating fluency. Each emphasized "understandable" and "comprehensible" sentences, which included grammatical sentences with no pronunciation errors.

The effects of grammatical accuracy on fluency judgments have been reported in previous studies (Kormos & Dénes, 2004; Rossiter, 2009). Suzuki and Kormos (2020) observed that morphological accuracy and pronunciation are highly correlated with fluency ratings. It seems that even expert raters were confused in rating fluency if they were provided explicit guidelines for grading.

Conversely, the other four raters seldom mentioned grammatical errors during the experiment. In speech samples containing a grammatical error, these raters downgraded the fluency rating at pauses before and after the error instead of mentioning the error as a problem.

59. *"I usually evaluate students on accuracy if their pronunciation is wrong and they make grammatical mistakes. For comprehensibility and expressions, I sometimes make another category for clarity and evaluate them."*

(Rater 1 on Speaker 6)

60. *"I will give high points for fluency if their speech rate isn't too slow, and the speech does not have a lot of pauses or lengthy pauses. …errors will be separately categorized in accuracy. I don't think too much about fluency …as long as I can comprehend what they are saying."*

(Rater 4 on Speaker 6)

Only 13 of 264 comments on accuracy came from Rater 1 (10 comments) and Rater 4 (three comments). Rater 1 mentioned grammatical problems related to self-repair situations twice, whereas Rater 4 did not mention it. Their comments on pronunciation were mostly related to vowel prolongation and speech rate (five from Rater 1 and three from

Rater 4). The remaining three comments from Rater 1 noted that the speakers' natural intonation and pronunciation upgraded their fluency rating. Raters' fluency perceptions are easily influenced by their own standards and understanding of fluency despite being experienced teachers and raters.

4.3 The Effects of Rating Modes

Non-linguistic feature such as rating mode is investigated in this study in order to find the effects of different rating methods on fluency assessment. This study attempted to examine whether the raters equally sensitive to the linguistic factors when they listen to the speech samples or when they watch people speaking.

Descriptive statistics for the fluency rating is provided in Table 5. As shown, the scores each student received did not differ between the rating modes (audio-only and video-mediated). The consistency ICC(Intraclass Correlation Coefficitn) of the raters reached .936 (p<.001) and Cronbach's alpha was .928. Therefore, raters seemed to assign similar scores to the speech samples in each mode.

Table 5. Descriptive Statistics for Fluency Ratings of Each Mode (1-9 scale).

	Picture Description				Opinion			
	Audio		Video		Audio		Video	
Task Type	M	SD	M	SD	M	SD	M	SD
Speaker 1	5.29	.95	5.43	.98	6.43	.98	6.29	1.11
Speaker 2	4.71	.76	5.86	1.07	6.00	.58	6.14	1.07
Speaker 3	6.43	.79	6.86	1.07	5.43	1.13	5.71	.76
Speaker 4	8.29	.49	8.29	.76	7.86	.90	7.71	1.11
Speaker 5	6.43	.98	6.86	1.07	7.57	.53	6.86	.38
Speaker 6	4.71	.76	4.57	1.13	3.14	.90	3.43	.53

Table 6 is the same as Table 3 and is presented again to compare the major themes and frequencies of the coded comments in each rating mode. The theme of comments from the audio-rating mode and the video-mediated mode are the same with the difference in the frequency that each feature was mentioned in. When the raters listened to the audio files, pause (203 comments) was mentioned most frequently, followed by accuracy (160), automaticity (109), and speech rate (71). In the video-mediated rating, automaticity (205) received the most comments, followed by pause (195), speech rate (104), and accuracy (104). Compared to audio-only ratings, the comments on automaticity and speech rate increased from 109 to 205 and from 71 to 104, respectively, whereas the total number of accuracy comments decreased from 160 to 104.

Table 6. Frequency of Coded Comments in the Stimulated Interviews of Each Rating Mode.

	Audio			Video		
Coded theme	Upgrade	Downgrade	Total	Upgrade	Downgrade	Total
Speech rate	57	14	71	78	26	104
Pause	0	203	203	1	194	195
Repair						
Correction	6	38	44	3	36	39
Repetition	0	29	29	0	25	25
Fillers	0	57	57	0	0	0
Automaticity	74	35	109	153	52	205
Accuracy						
Grammar	16	25	41	4	15	19
Pronunciation	18	101	119	13	72	85
Total	308	502	673	252	420	672

First of all, comments pertaining to automaticity increased significantly when upgrading fluency in video-mediated rating. However, raters' descriptions and explanations tended to be simple and not as specific as in the audio-only mode. The raters' answers shared more of an overall impression rather than specific reasons. An example of this difference in comments is shown below. These comments come from the same rater (Rater 5) for the same speech sample (picture description task by Speaker 1).

The comment in audio-rating mode is as follows:

61. *"The last sentence was spoken without hesitation or pauses. It*

seems like she did not make an effort to make this sentence. It means that she can easily make this kind of sentence, so she is not a low-level learner."

(Rater 5 on Speaker 1)

The comment in video-rating mode is as follows:

62. *"It sounded natural. There were no awkward parts. Just everything was fine and went smoothly."*

(Rater 4 on Speaker 1)

Many raters simply commented "there's no serious error," "it's natural," or "it's a generally good sentence." In order to determine whether these comments belonged to the accuracy theme or automaticity theme, the researcher asked the raters for precise descriptions. For example, if the rater described only the grammatical structure or sentence length, the comment was categorized as accuracy. However, comments referring to grammatical structure with a rapid speech rate, good word choice, or appropriate expressions were categorized as automaticity.

Though the descriptions pertaining to video-clip became short and

ambiguous, the underlying characteristics were the same as audio-clip analysis: complex grammatical structure, organization of speech, and naturalness. For instance, raters referred to automaticity if speakers were able to form complex, compound sentences without grammatical errors using good word choice. Concurrently, raters expected speakers to produce sentences with good pronunciation and natural intonation.

In terms of speech rate, the number of comments increased slightly in video clip rating compared to the audio clip interviews. However, no significant difference was observed in comments between audio-rating mode and video-mediated rating mode except for one example. After taking a closer look at the results, Rater 1's judgment on Speaker 4 was found to be completely different between the two testing modes. When Rater 1 listened to the audio clip, she commented "speech rate is natural" and "the speaker divides the speech in chunks." Contrastingly, in the video-rating mode, she judged that, "she speaks too slow," "she pronounces every single word so the speech rate is too slow," and "it's a slow prolonged speech." At the end of the video-rating interview, the researcher showed the result of the audio-only rating and enquired of the reason for her contrasting judgments. However, she could not clarify why the speech sounded slower in the video clip than in the audio clip.

63. *"She speaks really slowly. It's boring or somewhat frustrating to wait her to finish each sentence. I don't know why. Perhaps it's because of the facial expression. She doesn't have facial expressions? Or because she is not making eye contact? She is looking at the camera, though."*

(Rater 1 on Speaker 4)

There could be many reasons explaining Rater 1's contrasting judgments on Speaker 4's speech rate. The repetitive rating process may have affected her judgment, or her condition at that moment may have affected her perception of speech rate. As other raters' judgments on Speaker 4 did not differ between the two modes, the contrast could be attributed to the rater's personal characteristics.

The number of comments on grammatical errors or phonological errors was significantly reduced in the video clip ratings. The total number of comments on accuracy in the audio clip ratings was 160 (41 upgrading and 119 downgrading) and in the video clip ratings was 104 (85 upgrading and 19 downgrading). When raters watched the video clips, they tended to be less sensitive to errors. During the audio clip ratings, raters often used various grammatical terms like "passive voice," "present perfect," or "objective complement." Conversely, not many grammatical terms were used in the video-clip ratings. In the case of pronunciation, raters pointed out overall contours or intonations in the

video-clip ratings rather than mentioning the pronunciation of a single word or phrase like in the audio-only ratings.

The examples of the comments in audio-only rating are as follows:

64. *"She is using difficult structure here. The form of objective complement is correct."* (Rater 7 on Speaker 3)

65. *"There is no subject in her sentence. 'going' should be a correct form, not just 'go."* (Rater 6 on Speaker 2)

66. *"Vowel sound in the world 'abroad' is strange."*

(Rater 2 on Speaker 1)

67. *"The pronunciation of 'cameras' is not natural. She pronounced it with very strong Korean accent"* (Rater 6 on speaker 3)

The examples of the comments in video-mediated rating are as follows:

68. "*She made a grammatically wrong sentence and repeated it several times."* (Rater 2 on Speaker 2)

69. *"Her intonation is awkward. Most sentences were finished with rising intonation."* (Rater 3 on Speaker 1)

A few studies have investigated the effects of computer-based tests versus face-to-face tests on speaking assessments (Elder & Iwashita, 2005; Iwashita et al., 2001; Wigglesworth, 1997). The previous studies focused on validity issues (Kenyon & Malabonga, 2001), test takers' strategic behavior (Swain et al., 2009), and test takers' performance (Brooks & Swain, 2014; Jeong et al., 2011; Zhou, 2008; Zhou, 2015). However, it was difficult to find research investigating raters' perceptions on audio clips versus video clips. Ultimately, more research is needed to understand how raters grade speaking tests using the different testing modes.

4.4 General Discussion

4.4.1 Dynamic Ratings of Fluency

This study examined L2 fluency to clarify the extent to which different linguistic features are associated with fluency perception across time, as raters listen to L2 speech. Raters' perception of fluency continued to change as they listened to the speakers. Most raters continued evaluating

the speakers' L2 fluency in real-time by upgrading or downgrading the speech samples. However, the frequency and magnitude were different according to the raters. A dynamic rater showed higher click frequency which led to greater fluctuation in her real-time assessment plots. Most raters fell into semi-dynamic group. Semi-dynamic raters also frequently evaluated the speakers' fluency while they were listening to or watching the speakers, but the range of their ratings was narrower, limited to ±1/ ±2 in most cases. Non-dynamic raters tended to reserve their judgment until the end of each sentence so they evaluated fluency far less frequently compared to other groups. Perception of fluency changes over time and the timing and the location of the error might produce a variable response in different listeners. This point was particularly salient in relation to specific speech features, which are provided in Section, 4.3.2.

4.4.2 Factors Influencing L2 Fluency Assessment

First, various speech features influenced raters' evaluations of L2 fluency and these features were intertwined. With respect to raters' explanations for their click activity, multiple categories were cited as reasons for upgrading or downgrading fluency, including speech rate, pause, self-repair, grammatical and phonological accuracy, and automatized production. As the excerpts from the qualitative comments indicate, these

speech features and concepts are inherently intertwined and cannot be easily distinguished from each other. According to the raters' reactions, an L2 speaker is considered fluent when they can combine all the features while speaking easily and relatively quickly, with pauses at appropriate junctures, and without grammatical errors. While these factors collectively influence L2 fluency, the speech features most frequently commented on by the raters in this dataset were pauses. In fact, pause phenomena were closely connected with other temporal features. Ginther et al. (2010) suggested that filled pauses should not be examined separately and should be incorporated with silent pauses or vocalization when examining fluency. In their study of L2 learners of Dutch, de Jong et al. (2013) found that the number of pauses along with repetitions and repairs were related to perceived fluency ratings.

Second, in examining the possible effects of the rating modes for fluency assessment, no significant differences in the scores were observed. Scores on audio-only delivered speech could be interpreted similar to those scores on video-mediated speech samples. This means that in spite of the observed variability between each testing mode, and given the high internal-consistency reliability attained in both rating modes, these remain within a comparable range. Although there was no significant difference in global rating scores, the stimulated interview revealed a difference in the raters' perceptions. In audio mode, raters

evaluated the errors of the speech more strictly while they evaluated the overall impression and naturalness of the speech in video mode. For this reason, there were more references to the accuracy section in audio-only rating, while references to automaticity increased significantly in the video-mediated rating. The inclusion of visual information such as facial expressions and gestures may have effects on the comprehension of intended messages or distract raters from focusing on the actual performance of the test takers. Additional study and review is required.

Thirdly, the fluency rating appeared to depend on the particular response strategy adopted by the raters. Raters referenced their own status as non-native English teachers and seemed to approach ratings from that perspective. For example, raters in this study seemed to excuse a slow speech rate even though speech rate is considered the best indicator of L2 fluency according to previous quantitative studies (Cucchiarini, Strik & Boves, 2002; Derwing, Rossiter, Munro & Thomson, 2004; Kormos & Dénes, 2004; Riggenbach, 1991; Rossiter, 2009). Further, raters did not expect the student participants to speak as quickly as native English speakers. The raters mentioned a slow speech rate as a reason for disfluency, but if the other features were satisfactory, they upgraded the speech or alternated between upgrading and downgrading.

Despite raters' efforts to understand the speakers as fellow L2 learners, there were times when grammatical constructions or

pronunciation negatively affected comprehension, and these comprehensibility issues downgraded fluency. Accuracy was the third most stated theme. This may be due to the fact that problems related to grammar or structure are easily perceived. Conversely, the effects of accuracy problems on fluency ratings varied based on each rater's perspective of fluency. Some raters considered comprehensible content delivery important, and thus, accuracy is necessary for fluency. However, some raters distinguished between fluency and accuracy matters. It seemed that a specific and explicit test scoring rubric is needed, and effective training is necessary for consistent fluency judgments. Another possible reason for using accuracy measures in fluency judgment was the characteristics of non-native raters. All raters in this study were non-native L2 speakers of English. As reviewed in Section 2.5, non-native listeners tend to be more sensitive to errors and more severe with comprehensibility and proficiency.

Moreover, the concept of fluency is confusing. Fluency as a component of oral proficiency does not seem to be fully understood by raters. The features associated with dysfluent speech may also include those related to the more proficiency-oriented view of fluency. Suzuki and Kormos (2020) reported that a strong association was found between raters' judgments of fluency and comprehensibility. Fluency and comprehensibility are not only conceptually overlapping but are also

difficult to distinguish while evaluating L2 learners' speech. In addition to temporal features, some studies suggested that perceptions of speaking fluency were affected by non-temporal features such as grammar, pronunciation, vocabulary, accent, or intonation (Freed, 1995; Lennon, 1990, Riggenbach, 1991; Rossiter 2009). Rossiter (2009) found that the fluency ratings were affected by non-temporal measures including pronunciation, grammar, and vocabulary as well as temporal features. It may be that listeners in that study made use of the broad definition of fluency that equates fluency with proficiency (Chamber, 1997).

Finally, automaticity was considered as an important factor for L2 fluency judgment. In the same vein, chunks seemed to be an important aspect of improving fluency. Chunk speech was a commonly mentioned reason for raters to upgrade their speech in this study. The importance of naturalness and the importance of using chunks to achieve naturalness have been recognized by a number of researchers (e.g. Erman & Waren, 2000; Ellis, 2001; Nattinger & Decarrico, 1992; Pawley & Syder, 1983; Wray, 2002). For example, Pawley and Syder (1983) concluded that the language learner's task is not only to master the generative rules of a language and produce grammatically correct sentences, but also to acquire knowledge regarding "which of the well-formed sentences are native-like". Ellis (2001) claimed that speaking natively is speaking idiomatically, using frequent and familiar collocations. As learners

increase their L2 proficiency, they need to acquire a vast number of chunks. Therefore, chunk acquisition is important for L2 learners to develop natural and fluent speech.

Chunks are relevant for usage-based accounts of language structure and language learning. Usage-based accounts hold that language learning is exemplar-driven. Through frequent encounters (usage-based events), learners save exemplars of utterances (tokens), some of which may later serve the basis for analysis into abstract constructions. This can be used productively later. As frequency of forms in input is a major driver of language acquisition, L2 learners with sufficient exposure to authentic input will acquire chunks just as they acquire other aspects of language. However, for L2 learners, the acquisition and use of native-like chunks may not be an easy task because L2 learners tend to focus on individual words as meaning units rather than on multi-word chunks. This is perhaps a result of classroom instruction, which tends to encourage separate attention to grammar and vocabulary (Wray, 2002). In addition, some chunks may not be sufficiently frequent or salient for the L2 learners (Granger & Paquot, 2008). Therefore, it is necessary to provide learners exposure to authentic and native-like chunks, and develop classroom materials that help them identify chunks and practice them.

CHAPTER 5 CONCLUSION

This chapter draws a conclusion based on the results and discussion proposed in the previous chapter. Section 5.1 presents a summary of the key findings of the present research, followed by some pedagogical implications. Section 5.2 discusses the limitations of the present study and provides some suggestions for future research.

5.1 Major Findings and Pedagogical Implications

Understanding fluency within the context of L2 speech production and perception is a critical challenge facing language assessment. This study sought to examine L2 fluency and clarify the extent to which different linguistic dimensions of speech are associated with moment-to-moment shifts in fluency while also evaluating speaking tests. The data suggested that raters consistently emphasized a fine balance between speech rate, pause, self-repair, grammar, pronunciation, and automized speech. In addition, the speech features and themes are intrinsically intertwined and are not easily distinguished from each other. L2 speakers were considered fluent when they combined all the features, enabling them to speak easily and quickly, with appropriately located pauses and without grammatical errors or strange pronunciation. Pertaining to automaticity, which depends on procedural knowledge, “chunk speech”

was often mentioned by the raters, and comments pertaining to "speaking in chunks" correlated with upgrading the speakers' fluency. Therefore, it would be a good strategy for test-takers to practice chunking in order to earn high scores on English speaking tests; chunking can help learners speak more naturally and understand natural sentence breaks. From a usage-based perspective, where input is one of the main drivers of language learning, one of the ways to recognize and develop chunking is by helping L2 learners gain sufficient exposure to authentic input. L2 learners in high input conditions were more successful in their development of native-like chunks than low-input learners (Erman & Warren, 2000; Hana, 2013; Verspore et al., 2010; Verspore, Schmid, & Xu, 2012). Therefore, it is necessary to develop teaching materials and methods with authentic chunks. L2 learners' difficulty with chunks can be caused by the limited capacity of working memory (Ellis, 2001). Chunks in L1 are learned, stored, and processed as whole units, while post-childhood L2 learners tend to analyze input for individual words. As a result, chunks containing more phonological units can be difficult for learners and the memory of them will fade unless chunks continue to be encountered and regularly used. In this regard, teaching students to speak in chunks could concurrently facilitate speech production and its fluency perception.

One of the speech features most frequently commented on by the

raters in this dataset was grammatical/phonological errors, usually considered as part of the accuracy segment. This may be due to the fact that grammatical and phonological errors are easily perceived features and, given all raters all non-native English teachers, they may be more conscious of these speech features in the case of L2 learners. The effects of grammatical accuracy on fluency judgments have been reported in previous studies (Kormos & Denes, 2004; Rossiter, 2009; Santos, 1988; Suzuki & Kormos, 2020), and non-native raters stricter in the case of L2 learners' errors (Brown, 1995; Kang, 2008; Kang, 2013; Lee, 2010). The definition of fluency was not understood identically among the raters. Fluency is sometimes confused with comprehensibility or proficiency (Chamber, 1997; Suzuki & Kormos, 2020) .

It seems that even expert raters were confused in rating fluency if they were provided explicit guidelines for grading.

Therefore, a key question is whether, given all other variables in L2 speech, it is possible to focus on fluency while ignoring grammatical errors, unexpected lexical choices, and the like, to the extent that individuals can detect existing fluency differences in the same speaker. In Song (2017) and Song and Lee (2015), although raters were relatively reliable and experienced, most of them did not apply appropriate scoring criteria in evaluating Korean students' English fluency and pronunciation levels. This is why education and rater training for speaking assessment is important. A complex and detailed scoring rubric is also key. Jeong

(2015) suggested that rater training should not only focus on rating practices but prior to ratings, sufficient time should also be given to learning the rubric. For classroom assessment, teachers usually develop their own rubric. However, it is recommended that teachers get together and develop or adapt a rubric by comparing and contrasting criteria, scales, and rubric styles. If an appropriate rubric is developed and raters correctly use the rating, it will be valid and reliable.

5.2 Limitations and Suggestions for Further Research

One possible limitation of this study is that each theme and object was identified by one researcher. It would be more reliable if the themes and objects were identified and coded by multiple researchers. Thus, in future studies, a process to achieve consensus on the themes and coding decisions should be implemented.

In this study, each rater's assessment process varied referring to different linguistic features for their judgment even though inter-rater reliability was being observed and their correlation was high. With larger samples, future studies should focus on the relationship between raters' subjective perceptions and objective measures such as temporal features from previous studies. Comparing subjectively perceived linguistic features and objectively measured temporal features can advance the

understanding of the raters' assessment process in oral performance.

REFERENCES

Arnold. M. E, Fagnano, M., & Tanenhaus, M. K. (2003). Disfluencies signal the, um, new information. *Journal of Psycholinguistic Research, 32*, 25-36.

Arnold, J. E., Hudson Kam, C. L., & Tanenhaus, M. K. (2007). If you say -thee uh- you're describing something hard: The on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 914-930.

Arnold, J. E., Tanenhaus, M. K., Altmann, R. J., & Fagnano, M. (2004). The old and thee, uh, new: Disfluency and reference resolution. *Psychological Science, 15,* 578-582.

Arons, B. (1993). Speech skimmer: Interactively skimming recorded speech. *Proceedings of 6th Annual ACM symposium on User Interface Software and Technology, USA, 6*, 187-196.

Beattie, G. (1977). The dynamics of interruption and the filled pause. *British Journal of Social and Clinical Psychology, 16*, 283-284

Beltrán, J. (2016). The effects of visual input on scoring a speaking achievement

test. *TESOL & Applied Linguistics, 16*(2), 1-23.

Biber, D., Johansson, S., Leech G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. London: Longman.

Blau, E. K. (1991). *More on comprehensible input: The effect of pauses and hesitation markers on listening comprehension*. From ERIC database. Paper presented at the Annual Meeting of the Puerto Rico Teachers of English to Speakers of Other Languages (San Juan, PR, November 15, 1991)

Bosker. H. R., pinget, A., Quene, H., Sanders, T., & de Jong, N.H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing, 30*, 159-175.

Boomer, D. S. (1965). Hesitation and grammatical encoding. *Language and Speech, 8*, 148-158.

Bower, G. H., & Springston, F. (1970). Pauses as recoding points in letter series, *Journal of Experimental Psychology, 83*, 421-430.

Brennan, S. E., & Schober, M. F. (2001). How listeners compensate for disfluencies in spontaneous speech, *Journal of Memory and Language, 44*(2), 274-296.

Brooks, L., & Swain, M. (2014). Contextualizing performances: comparing

performances during TOEFL iBT and real-life academic speaking activities. *Language Assessment Quarterly, 11(4)*, 353-373.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing, 12*(1), 1-15.

Brown, A. (2005). *Interviewer variability in oral proficiency interviews*. Frankfurtam Main: Peter Lang.

Brown, A., Iwashita, N., & McNamara, T. F. (2005). *An examination of rater orientations and test taker performance on English for academic purposes speaking tasks* (Monograph Series 29th ed.) Educational testing Service.

Brumfit, C. (2000). Accuracy and fluency: The basic polarity. In H. Riggenbach, (Ed.), *Perspectives on fluency* (pp. 61-73). Ann Arbor: University of Michigan Press.

Butcher, K. (2006). Learning from text with diagrams: Promoting mental model development and inference generation. *Journal of Educational Psychology, 98*(1), 182-197.

Bybee, J. L, & Hopper, P. J. (Eds.) (2001). *Frequency and the emergence of*

linguistic structure: Volume 45. Amsterdam: John Benjamins.

Carey, M. D., Mannell, R ., & Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing, 28*(2), 201-219.

Carr, N. (2011). *Designing and Analyzing Language Tests*. Oxford, UK: Oxford University Press.

Ceban, H. L. (2003). Rater group bias in the speaking assessment of four L1 Japanese ESL students. *Second Language Studies, 21*(2), 1-44.

Clark, H. H. (2002). Speaking in time. *Speech Communication. 36*, 5-13.

Clark, H. H. & Fox Tree J.E. (2002). Using uh and um in spontaneous speaking. *Cognition 84(1)* 73-111.

Cenoz, J. (1998). *Pauses and communication strategies in second language speech*. Rockville: Educational Resources Information Center.

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

Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing, 12*(1), 16-33.

Cohen, M. A, Horowitz, T. S., & Wolfe, J. M. (2009). Auditory recognition memory

is inferior to visual recognition memory. *Psychological and Cognitive Sciences, 106*(14). 6008-6010.

Cucchiarini, C., Strik, H., & Boves, L. (2000). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *Journal of the Acoustical Society of America, 107*, 989-999.

Cucchiarini, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *Journal of the acoustical Society of America, 111*, 2862-2873.

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal, 86*(1), 67-96

Crystal, D. (1987). *The Cambridge Encyclopedia of language*. Cambridge: Cambridge University Press, Cambridge.

Davies, A. (2003). The native speaker: Myth and reality (2nd ed.). Tonawanda, NY: Multilingual Matters.

Dehaene, S., Dupoux, E., mehler, J., Cohen, L., Paulesu, E., Perani, D., van de Moortele, P. F., Lehericy, S. & Le Bihan, D. (1997). Anatomical variability in the cortical representation of first and second language. *Neuroreport,*

8, 3809-3815.

De Jong, N. H., & Perfetti, C. A. (2011). Fluency training in the ESL classroom: An experimental study of fluency development and proceduralization. *Language Learning, 61*(2), 533-568.

De Jong, N. H., Groenhout, R., Schoonen. R., & Hulstijin. J.H. (2013) Second language fluency; Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics, 36*(2), 223-243

De Jong, N. H., Steinel, M. P., Florijin. Al, Schoonen, R., & Hulstijin, J. H. (2013). Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics*, 34, 893-916.

De Jong, N. H. (2016). Predicting pauses in L1 and L2 speech: The effects of utterance boundaries and word frequency. *International Review of Applied Linguistics in Language Teaching, 54*, 113-132.

De Jong, N. H. (2018). Fluency in second language testing: insights from different disciplines. *Language Assessment Quarterly, 15(3*), 237-254.

Derwing, T., Rossiter, M., Munro, M., & Thomson, R. (2004). Second language fluency: Judgments on different tasks. *Language Learning, 54*, 655-679.

Ducasse, A., & Brown, A. (2009). Assessing paired orals: Rater's orientation to interaction. *Language Test, 26*(3), 423-443.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly, 2*(3), 197-221.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*(2), 155-185.

Ejzenberg, R. (2000). The juggling act of oral fluency: A psycho-sociolinguistic metaphor. In H. Riggenbach (Ed.), *Perspectives on fluency*, 287-313. Ann Arbor: University of Michigan Press.

Elder, C. & Iwashita, N. (2005). Planning in language testing. In R. Ellis (Ed.), *Planning and task performance in a second language (pp*. 217-238). Amsterdam: John Benjamins.

Ellis, N. C. (2001). Memory for language. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 33-68). Cambridge: Cambridge University Press.

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text Interdisciplinary Journal for the Study of Discourse, 20*, 29-62.

Fayer, J. M., & Krasinski, E. (1987). Native and nonnative judgments of intelligibility and irritation. *Language Learning, 37*(3), 313-326.

Ferreira, F. (1993). Creation of prosody during sentence production. *Psychological Review*, *100*, 233-253.

Ferreira, F. (2007). Prosody and performance in language production. *Language and Cognitive Processes, 22*, 1151-1177.

Fillmore, C. J. (1979). On Fluency. in C. J. Fillmore, D. Kempler, & W. S. Wang (Eds.) *Individual differences in language ability and language behavior* (pp. 85-101). New York, NY: Academic Press.

Fillmore, C. J. (2000). On fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 43-60). Ann Arbor, MI: University of Michigan Press.

Fox Tree, J. E. (2001). Listeners' uses of um and uh I speech comprehension. *Memory & Cognition, 29*, 320-326.

Freed, B. F. (1995). Do students who study abroad become fluent? In B. F. Freed (Ed.), *Second language acquisition in a study abroad context* (pp. 123-148). Amsterdam: John Benjamins.

Freed, B. F. (2000). Is fluency, like beauty, in the eyes and ears for the beholder?

In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 243-265). Ann Arbor: University of Michigan Press.

Freed, B. F., Segalowitz, N., & Dewey, D. P. (2004). Context of learning and second language fluency in French: Comparing regular classroom, study abroad, and intensive domestic immersion programs. *Studies in Second Language Acquisition, 26,* 275-301.

Fulcher, G. (2003). *Testing second language speaking*. Harlow, North Dakota: Pearson Education.

Gass, S. M., & Mackey, A. (2007). *Data elicitation for second and foreign language research*. NY: Rutledge.

Ginther, A. (2013). Assessment of Speaking. In C. A. Chappelle (Ed.), *The Encyclopedia applied linguistics*. Oxford, UK: Wiley-Blackwell.

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

Granger, S., & Paquot, M. (2008). Disentangling the phraseological web. In S. Granger & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 28-49). Amsterdam: John Benjamins.

Griffiths, R. (1991). Pausological research in an L2 context: A rationale, and review of selected studies. *Applied Linguistics, 12,* 345-364.

Guillot, M. (1999). *Fluency and its teaching*. Clevedon: Multilingual Matters.

Hana, S.-G. (2013). *Chunks in L2 development: A usage-based perspective*. Doctoral dissertation, University of Groningen, Groningen, Netherlands.

Hawkins, R. R. (1971). The syntactic location of hesitation pauses. *Language and Speech, 14*, 277-288.

Holmes, V. M. (1988). Hesitations and sentence planning. *Language and Cognitive Processes, 3*, 323-361.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics, 30(4*), 461-473.

Hsieh, C. (2011). Rater effects in ITA testing: ESL teachers' versus American undergraduates' judgments of accentedness, comprehensibility, and oral proficiency. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 9*, 47-74.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: how distinct? *Applied Linguistics,*

29(1), 24-49

Iwashita, N., McNamara, T., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to task design. *Language Learning, 51*(3), 401-436.

Izumi, S. (2003). Comprehension and production processes in second language learning: In search of the psycholinguistic rationale for the output hypothesis. *Applied Linguistics, 24*, 168-96.

Jeong, H. (2015). Rubrics in the classroom: do teachers really follow them? *Language Testing in Asia, 5*(1), 1-14

Jeong, H., Hashzume, H., Sugiura, M., Sassa, Y., Yokoyama, S., Shiozaki, S., et al. (2011). Testing second language oral proficiency in direct and semidirect settings: a social-cognitive neuroscience perspective. *Language Learning, 61(3*). 675-699.

Joo, M., & Kim, Y. (2011). Investigation of the effects of a computer-mediated English test on speaking performance in terms of accuracy, fluency, and complexity, *Korean Journal of English Language and Linguistics, 8*(3), 677-699.

Kahneman, D. (1973). Attention and effort. Englewood Cliffs, NJ: Prentice Hall.

Kahng, J. (2014). Exploring utterance and cognitive fluency of L1 and L2 English speakers: Temporal measures and stimulated recall. *Language Learning, 64*, 809-854.

Kahng, J. (2018). The effect of pause location on perceived fluency. *Applied psycholinguistics 39*, 569-591.

Kang, O. (2008). Ratings of L2 oral performance in English: Relative impact of rater characteristics and acoustic measure of accentedness. *Spaan Fellow. 6*, 181-205.

Kang, O. (2010). Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness. *System, 38*, 301-315.

Kang, S. (2013). The study on Korean raters' characteristics for Korean English oral performance. *Studies in Linguistics, 26*, 1-21.

Kang S., & Ahn, H. (2012). A comparative study on criteria and tasks in Korean English speaking assessment by native and non-native raters. *Language Research, 48*(2), 1-25.

Kargopoulos, P., Bablekou, Z., Gonida, E, & Kiosseoglou, G. (2003). Effects of face and name presentation on memory for associated verbal descriptors. *The American Journal of Psychology, 116(3*), 415-430.

Kenyon, D. M. & Malabonga, V. (2001). Comparing examinee attitudes toward computer-assisted and other oral proficiency assessments. *Language Learning Technology, 5(2)*, 60-83.

Kim, M. S. (2018). *The Effects of Pause and Speech Rate in Evaluating English Speech*, Doctoral Disseration, Hankuk University of Foreign Studies.

Kormos, J. (1999a). Monitoring and self-repair in L2. *Language Learning, 49*, 303-342.

Kormos, J. (2000). The role of attention in monitoring second language speech production. *Language Learning, 50*(2). 343-384.

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

Kormos, J. (2006). *Speech production and second language acquisition*, London: Lawrence Erlbaum.

Kowal, S. H., & O'Connell, D. C. (2008). *Communicating with One Another: Toward a psychology of spontaneous spoken discourse*. Springer Science and Business Media.

Lass, N. J., & Leeper, H. A. (1977). Listening rate preference: Comparison of two-

time alternation techniques. *Perceptual and Motor Skills, 44*, 1163-1168.

Lavolette, E. (2013). Effects of technology modes on ratings of learner recordings. *IALLT Journal of Language Learning Technologies, 43*(2), 1-27

Lee, C.-H. (2010). Improving inter-rater reliability in oral proficiency test at college level. *Modern Studies in English language & Literature. 54*(1), 367-387.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning, 40*, 387-417.

Lennon, P. (2000). The lexical element in spoken second language fluency. In H. Riggenbach (Ed.), *Perspective on fluency* (pp. 25-42). Ann Arbor, M: University of Michigan Press.

Levelt, W. (1983). Monitoring and self-repair in speech. *Cognition, 14*, 41-104.

Lieven, E. V. M., Pine, J. M., & Barnes, H. D. (1992). Individual differences in early vocabulary development: Redefining the referential-expressive distinction. *Journal of Child Language, 19*(2), 287-310,

Lindner, K., Blosser, G., Cunigan, K. (2009). Visual versus auditory learning and memory recall performance on short-term versus long-term tests.

Modern Psychological Studies, 15(1) 39-46.

Lumley, T. (2005). *Assessing second language writing: the rater's perspective*. Frankfurt am Main: Peter Lang.

Luoma, S. (2004). *Assessing speaking*. Cambridge, UK: Cambridge University Press.

MacGregor, L. J. (2008). *Disfluencies affect language comprehension: Evidence from even related potentials and recognition memory* (Doctoral dissertation). Retrieved from Edinburgh Research Archive(http://hdl.handle. net/1842/3311).

MacIntyre, P. D. (2012). The idiodynamic method: A closer look at the dynamics of communication traits. *Communication Research Reports, 29*, 361-367.

Malabonga, V., Kenoyn, C. M., & Carpenter, H. (2005). Self-assessment, preparation and response time on a computerized oral proficiency test. *Language Test, 23*(2), 131-166.

McNamara, T. F. (1997). 'Interaction' in second language performance assessment: Whose performance? *Applied Linguistics, 18*, 446-466.

McNamara, T. F., & Adams, R. J. (1994). Exploring rater characteristics with

Rasch techniques. In Selected papers of the 13th Language Testing Research Colloquium (LTRC). Princeton, NJ: Educational Testing Services, international Testing and Training program Office.

McNamara, T. F., & Lumley, T. (1997). The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational settings. *Language Testing, 14*(2), 140-156.

Möhle, D. (1984). A comparison of the second language speech production of different native speakers. N. H. W. Dechert, D. Möhle, & M. Raupach, (Eds.), *Second language production*. (pp. 26-49). Tübingen: Günter Narr.

Nagle, C., Trofimovich, P. & Bergeron, A. (2019). Toward a dynamic view of second language comprehensibility. *Studies in Second Language Acquisition 41*, 647-672.

Nakatsuhara, F. (2006). Impact of inter-interviewer variation on analytical rating scores and discourse in oral interview tests. *Newcastle Working Paper in Linguistics, 12*, 55-68.

Nambiar, M. K., & Goon, C. (1993). Assessment of oral skills: A comparison of scores obtained through audio recordings to those obtained through f ace-to-face evaluation, *RELC Journal, 24*(1), 15-31.

Nattinger, J. R. & DeCarrico, J. S. (1992). *Lexical phrases and language teaching*. Oxford: Oxford University Press.

Neu, J. (1990). Assessing the role of nonverbal communication in the acquisition of communicative competence in L2. In R. Scarcella, E. Andersen, & S. D. Krashen (Eds.), *Developing communicative competence in a second language* (pp. 121-138). New York, NY : Newbury House.

O'Brien, I., Segalowitz, N., Freed, B., & Collentine, J. (2007). Phonological memory predicts second language oral fluency gains in adults. *Studies in Second Language Acquisition, 29*, 557-582.

O'Connell, D. C., & Kowal, S. H. (1983). Pausology. *Computers in Language Research, 2*(19), 221-301.

O'Shaughnessy, D. (1992, October). *Analysis of false starts in spontaneous speech*. Paper presented at the International Conference on Spoken Language Processing, Banff, Alberta, Canada. (ERIC Document Reproduction Service No. ED 356 506)

O'sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing, 19*(3), 277-295.

O'sullivan, B. (2008) *Modelling performance in tests of spoken language*.

Frankfurt, Germany: Peter Lang.

Pakhomov, S. V., Kaiser, E. A., Boley, D. L., Marino, S. E., Knopman, D. S., & Birnbaum, A. K. (2011). Effects of age and dementia on temporal cycles in spontaneous speech fluency. *Journal of Neurolinguistics, 24*, 619-635.

Pawley, A., & Syder, F. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. Richards & R. Schmidt (Eds.), *Language and communication* (pp. 191-226). New York: Longman.

Pawley, A., & Syder, F. (2000). The one clause at a time hypothesis. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 163-191). Ann Arbor, MI: University of Michigan Press.

Ploug, L., & Carter, P. (2014). Timing and tempo in spontaneous phonological error repair. *Journal of Phonetics, 45*, 52-63.

Préfontaine, Y. (2013). Perceptions of French fluency in second language speech production. *Canadian Modern Language Review, 69*(3), 324-348.

Préfontaine, Y., Kormos, J., & Johnson, D. E. (2015). How do utterance measures predict raters' perceptions of fluency in French as a second language? *Language Testing*, 1-21.

Préfontaine, Y., & Kormos, J. (2016). A qualitative analysis of perceptions of fluency in second language French. *IRAL, 54*(2), 151-169.

Raupach, M. (1980). Temporal variables in first and second language speech production. In H.D. Dechert & M. Raupach (Eds.), *Temporal variables in speech: Studies in honour of Frieda Goldman-Eisler (pp. 263-270)*. Hague: Mounton

Raupach, M. (1987). Procedural learning in advanced learners of a foreign language. In J. A. Colman & R. Towell (Eds.), *The advanced language learner (pp. 123-155)*. London: CILT

Reich, S. S. (1980). Significance of pauses for speech perception. *Journal of Psycholinguistic Research, 9*, 379-389.

Riazantseva, A. (2001). Second language proficiency and pausing. *Studies in Second Language Acquisition, 23*, 297-526.

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

Rochester, S. (1973). The significance of pauses in spontaneous speech. *Journal of Psycholinguistic Research 2*(1), 51-81.

Rossiter, M. J. (2009). Perceptions of L2 fluency by native and non-native speakers of English. *The Canadian Modern Language Review, 65*(3), 395-412.

Sajavaara, K. (1987). Second language speech production: Factors affecting fluency. In H. Dechert & M. Raupach (Eds.), *Psycholinguistic models of production (pp. 45-65)*. Norwood: Ablex.

Santos, T. (1988). Professors' reactions to the academic writing of nonnative-speaking students. *TESOL Quarterly, 22*(1), 69-90

Schmidt, R. (1992). Psychological mechanisms underlying second language fluency. *Studies in Second Language Acquisition, 14(3*), 357-385.

Schüller, B., Steidl, S., Batliner, A., Bürkhardt, F., Devillers, L., Muller, C., & Narayanan, S. (2013). Paralinguistics in speech and language – State-of-the-art and the challenge. *Computer Speech & Language*, *27*, 4-39.

Segalowitz, N., & Hulstijin, J. (2005). Automaticity in bilingualism and second language learning. In F. F. Kroll & A.M.B De Groot (Eds.), *Handbook of bilingualism: Psycholinguistics approaches (pp. 371-388).* Oxford, UL: Oxford University Press.

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. New York, NY: Routledge.

Shriberg, E. E. (2005). Spontaneous speech: How people really talk, and why engineers should care. In: Proc. 9th European Conf, on Speech Communication and Technology, Lisbon, Portugal, pp 1781-1784.

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

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

Song, M. Y. (2017). Nonnative raters' perceptions and judgments of Korean English learners' fluency and pronunciation level. *Korean Journal of English Language and Linguistics, 17*(4), 787-815

Song, M. Y., & Lee, Y. S., (2015). Scoring behavior of English speaking raters: Suggestions for rater training. *Journal of Research in Curriculum & Instruction, 19*(4). 1081-1101.

Stenstroem, A. (1994). An Introduction to Spoken Interaction. London and New York: Longman.

Sugito, M. (1990). On the role of pauses in production and perception of discourse. *Proceedings of the 1st international Conference on Spoken Language Processing, Japan, 1,* 513-516.

Suzuki, S., & Kormos, J. (2020). Linguistic dimensions of comprehensibility and perceived fluency: An investigation of complexity, accuracy, and fluency in second language argumentative speech. *Studies in Second Language Acquisition*, *42*, 143-167.

Suzuki, S., Kormos, J., & Uchihara, T. (2021). The relationship between utterance and perceived fluency: A meta-analysis of Correlational studies. *The Modern Language Journal, 105*(2), 435-463.

Swain, M. (1985). *Communicative competence: some roles of comprehensible input and comprehensible output in its development*. In S. M. Gass & C. G. Madden (Eds.), *Input in second language acquisition*. Rowley: Newbury House.

Swain, M., Huang, L.S., Barkaoui, K., Brooks, L., & Lapkin, S. (2009). *The Speaking Section of the TOEFL iBT: Test-takers' reported strategic behaviors* (TOEFL iBT Research Rep. No. 10). Princeton, NJ: Educational Testing Service. http://www.ets.org/Media/Research/pdf/RR-09-30.pdf.

Tavakoli, P. (2011). Pausing patterns: Differences between L2 learners and native speakers. *ELT Journal, 65*, 71-79.

Tavakoli, P. (2016). Fluency in monologic and dialogic task performance:

Challenges in defining and measuring L2 fluency. *International Review of Applied linguistics in Language Teaching 54*(2), 133-150.

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing, In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239-273). John Benjamins.

Tavakoli, P., Nakatsuhara, F., & Hunter, A-M. (2020). Aspects of fluency across assessed levels of speaking proficiency, *The Modern Language Journal, 104* (1), 169-191.

Towell, R. (2002) Relative degrees of fluency: A comparative case study of advanced learners of French. *international Review of Applied Lingistics in Language Teaching 40*(1), 117-150.

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

Trofimovich, P., & Baker, W. (2006) Learning second language suprasegmentals: Effects of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition, 28*, 1-30.

Van Hest, E. (1996). *Self-repair in L1 and L2 production*. Tilburg, the Netherlands: Tilburg University Press.

Van Lier, L. (2004). *The ecology and semiotics of language learning: A sociocultural perspective*. Boston: Kluwer Academic.

Verspoor, M. H., Lowie, W. M., & van Dijk, M. (2008). Variability I second language development from a dynamic systems perspective, *Modern Language Journal, 92*, 214-231.

Verspoor, M. H., Schmid, M. S., & Xu, X. (2019). A dynamic usage-based perspective on L2 writing. *Journal of Second Language Writing, 21*(3), 239-263.

Voss, B. (1979). Hesitation phenomena as sources of perceptual errors for non-native speakers. *Language Speech 22(2)*, 129-144.

Wennerstrom, A. (2001). *The music of everyday speech: Prosody and discourse analysis*. Oxford: Oxford University Press.

Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Testing, 14*(1), 85-106.

Williams, S. A., & Korko, M. (2019). Pause behavior within reformulations and the proficiency level of second language learners of English. *Applied Psycholinguistics, 40*(3), 723-742.

Winke, P., Gass, S., & Myford, C. (2001). *The relationship between raters' prior language study and the evaluation of foreign language speech samples*. Princeton, NJ: Educational Testing Service.

Wolfe, W. E., Kao, C.-W., Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication, 15*(4), 465-492.

Wood, D. (2004). An empirical investigation into the facilitating role of automatized lexical phrases in second language fluency development. *Journal of language and learning, 2*(1), 27-50.

Wood, D. (2010). *Formulaic language and second language speech fluency: Background, evidence and classroom applications.* London: Continum.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Zhang, Y., & Elder, C. (2011). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing, 28*(1), 31-50.

Zhou, Y. J. (2008). A comparison of speech samples of monologic tasks in speaking tests between computer-delivered and face-to-face modes. *Japan Language Test Association, 11*, 189-208.

Zhou, Y. J. (2015). Computer-delivered or face-to-face: effects of delivery modern the testing of second language speaking, *Language Testing in Asia, 12(5*), 1-16.

APPENDICES

1. Appendix A: An Example of Transcription of a Speech Sample

2. Appendix B: Picture Description Tasks

3. Appendix C: Opinion Tasks

4. Appendix D: Idio-dynamic Software (Anionvariable tester V2)

5. Appendix E: Image and Excel from Idio-dynamic Software
   (Anionvariable tester V2)

6. Appendix F: An Example of Transcription of Stimulated Interview

Appendix A: An Example of Transcription of a Speech Sample

Speaker 1

Task 1 (Picture description)

(0.36) Many books *(0.17)* are piled up *(0.39)* on the table. *(1.05)* Four people *(0.78)* are standing around the *(0.63)* book. *(1.9)* I think one woman is talking about *(0.2)* the *(0.14)* book. *(1.37)* Two women *(1.16)* is wearing *(0.35)* glasses. *(1.39)* One woman *(1.74)* wore *(0.58)* sky blue short sleeved shirt. *(2.45)* It's bright *(0.61)* outside *(1.14)* because it's daytime. *(3.36)* Outside of the building *(0.77)* there are many trees *(0.33)* and *(0.39)* people.*(2.99)*

Appendix B: Picture Description Tasks

1. Please describe the picture in as much detail as you can. You have 45 seconds to speak about the picture.

2. Please describe the picture in as much detail as you can. You have 45 seconds to speak about the picture.

Appendix C: Opinion Tasks

3. Give your opinion about the topic below. Be sure to say as much as you can in the time allowed. You have 60 seconds to speak.

- Which is a better way for a university student to spend a long vacation: traveling abroad or doing an internship?

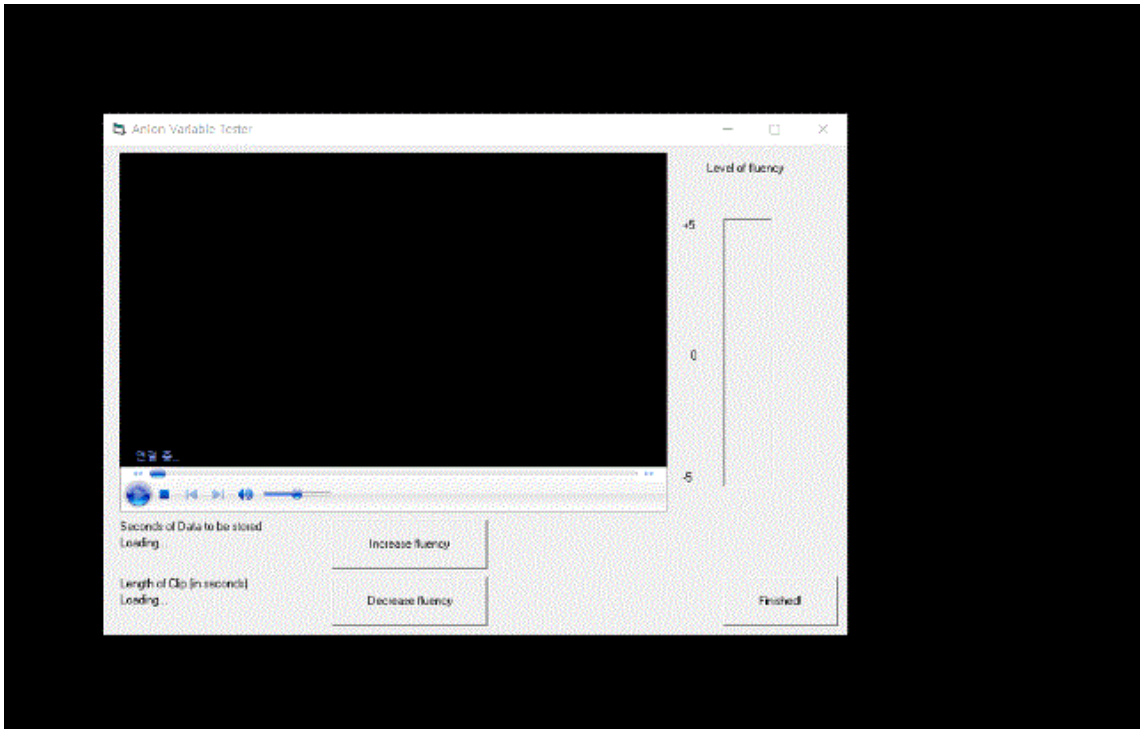
Use specific reasons and examples to support your answer.

4. Give your opinion about the topic below. Be sure to say as much as you can in the time allowed. You have 60 seconds to speak.

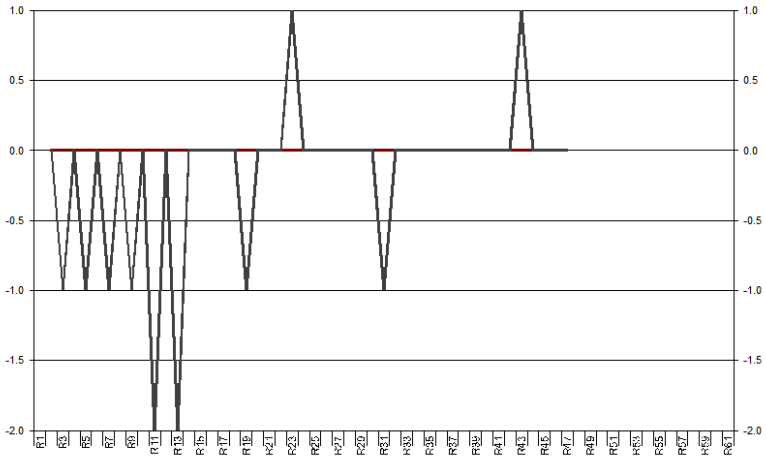
- What are some advantages of using social media as a marketing tool?

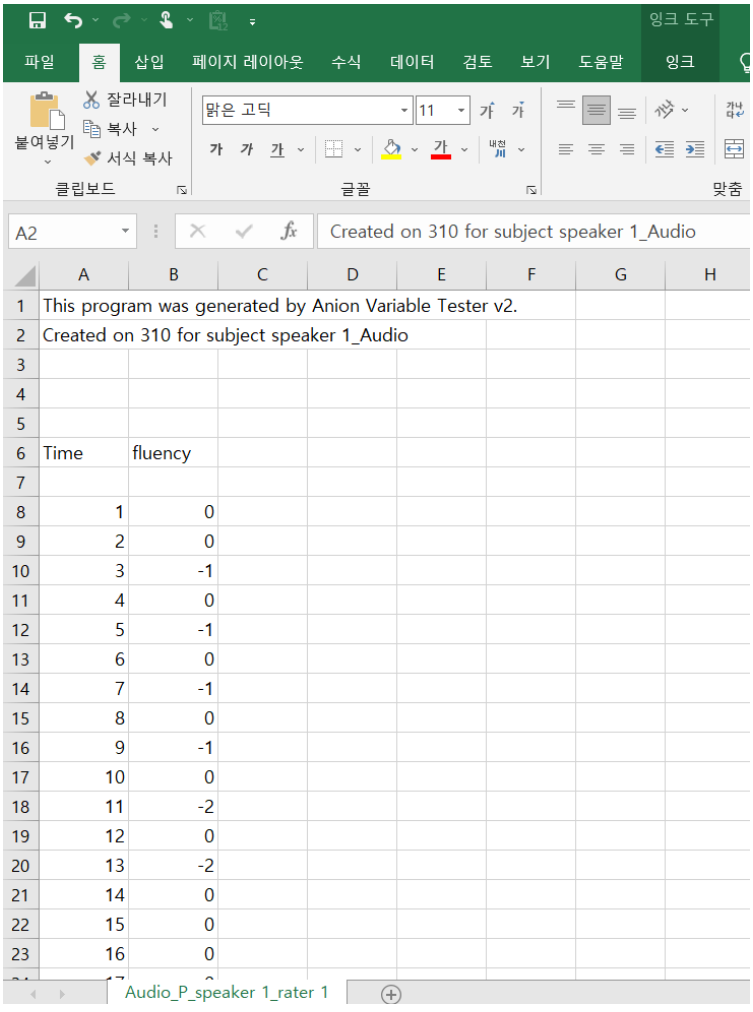
Use specific reasons and examples to support your opinion.

Appendix D: Idio-dynamic Software(Anionvariable tester V2)



Appendix E: Image and Excel from Idio-dynamic Software (Anionvariable tester V2)





잉크 도구

파일 홈 삽입 페이지 레이아웃 수식 데이터 검토 보기 도움말 잉크

잘라내기 복사 서식 복사 붙여넣기 클립보드

맑은 고딕 11 가 가 글꼴

맞춤

A2 Created on 310 for subject speaker 1_Audio

	A	B	C	D	E	F	G	H
1	This program was generated by Anion Variable Tester v2.							
2	Created on 310 for subject speaker 1_Audio							
3								
4								
5								
6	Time	fluency						
7								
8	1	0						
9	2	0						
10	3	-1						
11	4	0						
12	5	-1						
13	6	0						
14	7	-1						
15	8	0						
16	9	-1						
17	10	0						
18	11	-2						
19	12	0						
20	13	-2						
21	14	0						
22	15	0						
23	16	0						

Audio_P_speaker 1_rater 1

Appendix F: An Example of Transcription of Stimulated Interview

Rater 1 on Speaker 5, Task 1(picture description)

Pause	Length	Speech	±	Comment
1	0.84			
		This picture was		
2	0.24		-	This pause is natural.
		taken		
			--	After 'taken', there is a small pause. Pauses for word searching are natural but once finishing word searching, this part should be spoken continuously without a break. But there is another small pause after 'taken'. I think it sounds really unnatural.
		(pause)		
		at a book store		
3	1.71			
		There are		
4	0.19			
		four		
		(pause)	-	This part didn't sound fluent. This part should be spoken without a break too but there are small pauses here again. These expressions are not difficult and more like idiomatic expressions. But it took a long time for her to come up with this expression. When she speaks these kinds of easy expressions slowly, it seems that these structures are not automatized and she put effort to process these structures. So it doesn't sound fluent at all. Easy expressions should be spoken fast. In case of novel expression, pauses are natural.
		people		
		(pause)	-	
		in this picture		

				You know what I mean. When students say idiomatic expressions and easy expressions slowly, I consider that they are less fluent in English. For example, there was a pause in the first sentence but that pause was not awkward because she was thinking about what to speak. But she does not say anything special or important here but it's slow and there are pauses in the middle so she sounds less fluent.
5	1.31			
		And		
6	0.87			
		many books are stacked on the table.	+	Intonation is very natural here.
7	2.22			
		Three of them are women		
8	0.86			
		but		
9	0.16		-	There is a long pause after 'but'. This part is not actually minus minus.
		I'm not sure about the one	++	Intonation is very natural.
10	0.83			R: There are more pluses in this sentence.
		who's behind		This part, plus here(a blue cardigan) and
11	0.44			here(on the right side of the picture),
		the woman		intonation is very natural and speech rate is
12	0.37			also natural here just like she is expressing her
		wearing		thoughts without delay.
13	0.07			
		a blue cardigan	+	
14	0.92			

		because		
15	0.14			
		I can't see the face of the person	+	
16	1.84			
		And it seems that	+	Contour is also very natural.
17	0.64			
		the woman		
18	0.13			
		on the right side of the picture	+	
19	0.54			
		is		
20	0.43			There are frequent pauses between words, 'is' 'explaining' 'the book' 'to the woman' She carefully articulates each word to say it clearly. Then it sounds less flowy and less communicative. It doesn't sound like a chunk. It feels like she was saying every word so I also cannot understand what she was saying right away because I also needed some time to combine the words to understand the meaning.
		explaining (pause) the book	-	
21	0.38			It sounds fluent to me when she speaks idiomatic expressions in chunks such as 'stacked on the table' or 'three of them are women'. Then, I also focus on the meaning. But I cannot understand immediately when she speaks each word with word stress, even not word stress but very when she speak each word clearly without contour.

R: You mean connected speech?
Yes, yes. Something like connected speech. In my case, connected speech or chunk processing is considered fluent and processing word by word is considered less fluent.
And the location of pauses is important. For example, the pause after 'This picture was' was not strange at all but pauses between 'four,' 'people' or 'explaining', 'the book' which are inserted within a clause and chunks are awkward even though the pauses are really short. These kinds of small pauses downgrade fluency.
Another example is 'taken'. Considering argument structure, there needs the argument after 'taken'. 'this picture was taken' is grammatical but semantically incomplete. This is an adverbial phrase but I consider it as an argument so it should be a chunk. So a small pause between 'taken' and 'at a book store' is really strange. You know what I mean. This sentence is passive and it(at a book store) is not an object. But semantically this part(at a book store) is necessary to tell where this picture was taken.
When a pause, even a really short pause, is inserted in a chunk, it sounds less fluent. Native speakers usually speak in chunks and of course with many pauses but the location of the pauses is important. Pauses within a chunk made me evaluate that the student is not fluent or proficient in English enough to automatize those given expressions and

process them in chunks. For example, the pause between 'but' and 'I'm not sure' is unnatural. After 'but' if the student added something like 'she is not…' the pause could be natural. But even in Korean, 'but I'm not sure' can be semantically considered as a chunk, right? It doesn't contain new information after 'but'. I am not sure. The pauses (here, here, and here) which I think are unnatural and less fluent are all within chunks. For example, 'explaining' and 'the book to' is closely connected to form a meaning so should be spoken in a row but there is a small pause between 'explaining' and ' the book'. Then I wonder why she showed a pause in a chunk. And I feel like she was not fluent. But there's a part that she sounds really fluent such as 'stacked on the table' and 'I'm not sure about the one'.

But it is ok to have a pause before new information or unfamiliar words. I can ignore pauses when students give new information like words or expressions that we don't usually use. For example, the pause between 'wearing' and 'a blue cardigan' is OK. If she tried to say 'a jacket,' not 'a cardigan', there should be no pause after wearing. 'A jacket' is not a special word so I would wonder why she paused here and judge that the speech is less fluent. But 'cardigan' is not a common word she uses every day and is more like a meaning-embedded word, so a pause before 'a cardigan' is not awkward. Rather this kind of

				pause is natural because she paused here to think about how to express the item.
		to the woman next to her.		
22	1.03			
		so		
		(pause)		
		she might be a bookstore clerk	++	intonation and speech rate are really natural.
23	0.23			
		or		
24	0.16			
		maybe		
		(pause)		
		her friend	+	
25	1.33			

국 문 초 록

유창성은 제 2언어 평가에서 중요한 평가 기준의 하나이다. 제 2언어 유창성 평가에 영향을 미치는 언어적 특성과 요인들을 밝히기 위해 많은 양적 연구들이 시행되어 왔다. 유창성을 구성하는 언어적 특성들을 분석하고 정리하는 선행 연구들을 통해 말하기 평가의 준거로써 유창성의 개념을 이해하고, 평가 기준을 정립할 수 있다. 그러나 공인 영어 인증 시험이나 수업에서의 말하기 수행평가처럼 현재 대부분의 영어 말하기 시험의 평가는 아직은 채점자들에 의해 이루어지는 경우가 많다. 따라서 채점자들이 학습자들의 유창성을 어떻게 인식하고 있으며 실제 시험 평가 상황에서 학습자들의 유창성 평가가 어떻게 이루어지는지 채점자의 시각을 분석해보는 것은 유의미하다. 본 연구는 유창성 평가에 대한 객관적인 언어적 특성을 분석하기보다는 채점자들의 주관적인 인식과 해석에 초점을 두고, 채점자들이 시험 상황에서 학습자들의 유창성 영역의 평가를 어떻게 이해하고 적용하는 지를 구체적으로 파악하고자 하였다. 유창성에 영향을 미치는 요인들과 평가 점수를 비교하는 기존 연구들과는 달리, 본 연구에서는 평가자들의 유창성에 대한 인식이 실시간으로 어떻게 변화하는 지를 Idio-dynamic software와 stimulated interview를 통해 심도 있게 살펴보았다. 이를 통해 채점자들이 제 2언어 유창성을 평가할 때 어떤 관점으로 접근하는지 살펴보고, 유창성 평가에 영향을 미치는 요인들이 무엇 이라고 채점자들이 인식하고 있는지, 그리고 이런 채점자들의 특성이 말하기 시험의 유창성 점수에 어떤 영향을 미치는지를 분석하였다.

본 연구는 채점자들의 실시간 평가 변화를 추적하여 제 2언어 학습자들의 발화에서 어떤 언어적 특징들이 유창성 평가에 영향을 미치는지 분석하였다. 7명의 채점자가

영어 말하기 능력이 각기 다른 6명의 학습자의 영어 말하기 수행을 듣고 유창성을 평가하였다. 학습자들은 그림 묘사하기와 주제에 대한 의견 말하기 과제를 수행하였고, 채점자들은 이 학습자들의 발화를 들으면서 Idio-dynamic Software를 사용하여 실시간으로 유창성이 증가하는지 감소하는지 표시하였다. 채점자들은 각각의 과업에 유창성 점수를 부여한 직후(1-9), 연구자와 면담을 통해 어느 부분에서 유창성이 증가했다고 혹은 감소했다고 평가했는지 그 이유에 대해 설명하였다. 채점자들은 각자가 고수하는 유창성에 대한 정의에 따라 채점기준을 세우고 적용하였다. 채점자들의 유창성 판단은 학습자의 발화를 듣는 동안 고정되어 있거나 한 번의 총체적인 평가로 이루어지기 보다는, 계속적으로 변화하는 유창성의 정도를 종합하여 결정된다고 볼 수 있다. 채점자들은 학습자의 발화를 듣는 동안 여러가지 요소를 고려하여 실시간으로 그리고 계속적으로 유창성을 평가하는 모습을 보였다. 유창성 평가에 영향을 미치는 요인들은 발화속도, 휴지, 자기수정, 문법과 발음의 정확성, 그리고 발화의 자동성으로 정리할 수 있다. 이 요인들은 독립적으로 유창성 평가에 적용되기 보다는 몇 개의 요소가 한꺼번에 언급되거나 이 요소들 사이의 적절한 균형이 강조되는 경우가 많았다. 다시 말해 채점자들이 유창성 평가에서 고려하는 발화 특징들은 서로 연결되고 결합되어 있어서 각 특징들이 명확하게 분리되어 채점기준으로 적용되고 있다고 보기 어렵다.

채점자들이 학습자들의 발화를 소리로만 듣고 평가를 하는 경우와 학습자가 발화하는 장면을 비디오로 보고 평가하는 방식을 비교했을 때, 유창성을 인식하는 데에 어떤 차이점이 있는지 함께 분석하였다. 비디오와 오디오를 이용한 채점 방식의 차이가 유창성 점수에 통계적으로 유의미한 차이를 가져오지는 않았지만, 유창성 평가에 영향을 미쳤다고 채점자들이 언급한 언어적 특징은 평가 방식에 따라 달랐다. 채점자들은 소리로만 듣고 채점할 때 문법적인 오류나 발음 오류에 더 민감하였고,

화면을 보면서 채점할 때 발화의 자동성이나 휴지 현상에 더 많은 주의를 기울였다. 그리고 소리로만 듣고 평가할 때 문장의 문법적인 오류에 대해 세세하게 설명하던 것과는 반대로 화면을 보고 평가를 할 때는 발화에 대한 전반적인 인상에 대해 더 많이 언급하였다. 그러나 유창성 평가의 기준으로 언급된 발화의 특성들은 그 테마(theme)에 있어서 유사하였고, 독립적으로 인식되기 보다는 서로 결합되어 평가에 적용되었다는 점에서, 본 연구는 두 평가방식의 차이점 보다는 '유창성 평가'에서 채점자들이 보여주는 평가 과정의 특징들에 더 초점을 두었다.

채점자들의 점수는 채점자간 신뢰도를 유지할 정도로 일관되었던 것에 반해, 평가 과정에서 언급되는 평가의 이유는 매우 다양하였다. 이는 비원어민 채점자들이 영어 말하기 시험에서 유창성을 인식하고 평가하는 것이 얼마나 복잡한 과정을 거쳐 이루어지는 것인지를 보여준다. 게다가 채점자들이 얼마나 다양하게 유창성이라는 개념에 접근하는지를 보여주고, 채점 과정에서 유창성을 독립적으로 평가하는 것에 대한 한계를 보여주고 있다. 그러므로 영어 말하기 평가에서 유창성 평가에 영향을 미치는 주요 변인이 포함된 채점 기준을 찾는 방안에 대한 논의가 필요하고, 구체적인 채점 기준을 적용하도록 하는 채점자 안내와 연수가 필요하다.

핵심어 : 영어 말하기 유창성, 유창성 평가, 유창성 인식, 실시간 평가, 평가 방법

학번 : 2007-30391