교육학박사학위논문

Korean High School Students' English Speaking Fluency: Focusing on Rater Variability and Phonetic Correlates

한국인 고등학생의 영어 말하기 유창성: 채점자 변동성과 음성적 특성을 중심으로

2022년 8월

서울대학교 대학원

외국어교육과 영어전공

윤 근 식

Korean High School Students' English Speaking Fluency: Focusing on Rater Variability and Phonetic Correlates

한국인 고등학생의 영어 말하기 유창성: 채점자 변동성과 음성적 특성을 중심으로

by Kunsik Yoon

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

Korean High School Students' English Speaking Fluency: Focusing on Rater Variability and Phonetic Correlates

한국인 고등학생의 영어 말하기 유창성:
채점자 변동성과 음성적 특성을 중심으로

지도교수 안 현 기

이 논문을 교육학박사 학위논문으로 제출함
2022년 5월

서울대학교 대학원
외국어교육과 영어전공
윤 근 식

윤근식의 박사 학위논문을 인준함
2022년 7월

위 원 장 _____ (인)

부위원장 _____ (인)

위 원 _____ (인)

위 원 _____ (인)

위 원 _____ (인)

Korean High School Students' English Speaking Fluency: Focusing on Rater Variability and Phonetic Correlates

by

Kunsik Yoon

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, Chair

SUN-YOUNG OH, Vice Chair

IN YOUNG YANG, Member

SEOKHAN KANG, Member

HYUNKEE AHN, Member

ABSTRACT

Korean High School Students' English Speaking Fluency: Focusing on Rater Variability and Phonetic Correlates

Kunsik Yoon
Department of Foreign Language Education (English Major)
Graduate School of Seoul National University

Fluency constitutes a crucial aspect of understanding second language (L2) performance and proficiency, and attaining high levels of fluency is one essential goal for many L2 learners. However, fluency has not been well understood, and the term has not been used consistently by L2 researchers and EFL educators. In addition, there is a paucity of studies concerning how raters in the EFL context perceive and evaluate fluency. To fill the academic gap and deepen understanding of the multidimensional construct of fluency, the current dissertation investigated how Korean English teachers, native English teachers, and peer students perceive and rate Korean high school students' speaking fluency in terms of perceived fluency and utterance fluency.

Study 1 investigated the differences in perceived fluency by three rater groups, employing a mix-method approach. Overall fluency ratings across two task types (picture narration, spontaneous speech) at speakers' different oral proficiency levels (low, mid, high) were analyzed quantitatively, and raters' written comments were examined qualitatively. The native and non-native teacher groups showed comparable severity patterns, but the peer group provided significantly lower fluency rating scores than the two EFL teacher

groups on both tasks across all proficiency levels. The following qualitative analyses confirmed the discrepancy between the two EFL teacher groups and the peer group. In addition, it was revealed that the three rater groups' evaluations for low-level learners were significantly affected by task types, with the spontaneous speech task scoring higher than the picture narration task.

The disparities in the three groups' perceptions of fluency reported in Study 1 were further supported and accounted for in Study 2. Study 2 examined the relationship between utterance fluency and perceived fluency to determine which acoustic model best predicted the three listener groups' perceived fluency and which acoustic features were associated with the three groups' decision-making of speakers' fluency levels. Two speed features (i.e., mean length of run, articulation rate) and two breakdown measures (i.e., silent pause rate within a clause, mean length of silent pauses) were found to be most strongly correlated with their perceived fluency. The regression analysis indicated that the mean length of run and the mean length of silent pauses were the two strongest predictors for the three rater groups, explaining most of the variance in the three regression models. However, the data further revealed that the regression models for native and non-native teachers were identical regarding the four entered variables and their relative contribution rankings, while the best regression model for the peer group showed some disparities. In addition, it was found that breakdown measures, such as the mean length of silent pauses, helped to distinguish the low-level from higher level (mid, high) groups, while speed measures, such as articulation rate, discriminated the high-level group from lower level (low, mid) groups.

These findings served as a foundation for a discussion of native versus non-native English teachers as fluency assessors on the one hand and the validity and reliability of peer assessment on the other. Based on empirical evidence drawn from Study 1 and 2, it can be concluded that native and non-native English teachers perceived and rated L2 fluency in a similar way, confirming that non-native teachers are as equally capable of serving as fluency raters as native teachers are. However, the peer group displayed rating patterns distinct from those of the teacher group, implying that much pedagogical effort is required to prepare peer students to serve as competent fluency raters in the Korean EFL context.

The current dissertation contributes to establishing a valid and reliable fluency assessment in the Korean EFL context by systematically analyzing how various groups of raters perceive and evaluate students' English speaking fluency. In addition, the study provides direct evidence regarding the possibility and limitations of peer assessment by comparing the peer group's judgments with those of the teacher groups. Regarding research methodology, the study contributes to illuminating the multidimensional constructs of fluency by combining two facets of fluency, like perceived fluency and utterance fluency. It is also shown that a comprehensive understanding of rating patterns drawn by different raters can be achieved by combining quantitative and qualitative research methods.

Keywords: speaking fluency, fluency rating, perceived fluency, utterance fluency, rater variability, peer assessment
Student Number: 2015-30489

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

The present dissertation aimed to investigate EFL students' speaking fluency by comparing fluency perceptions made by native English teachers, non-native English teachers, and peer students and correlating their fluency judgments with objectively observed acoustic features to trace the underpinning mechanism of rater disparities. This chapter begins with a statement of the study's purpose in Section 1.1. The background for the current study is discussed in Section 1.2. The research questions are presented in Section 1.3, and Section 1.4 outlines the organization of this dissertation.

1.1 Aims of Study

The study examines how various raters in the Korean EFL classroom, including native English teachers, non-native English teachers, and peer students, perceive and rate Korean high school students' English speaking fluency differently. Specifically, the study investigates fluency perception among raters across L2 learners' speaking proficiency levels and task types. Furthermore, the current study identifies acoustic features which contribute to raters' fluency ratings and differentiate learners' fluency levels. The research's ultimate objective is to investigate a multifaceted aspect of fluency and provide insight into the valid assessment of speaking fluency in

the Korean EFL context, where three groups of raters (native English teachers, non-native English teachers, and peers) serve as L2 learners' fluency raters. Additionally, the study provides pedagogical insight into the design of fluency-focused instruction aimed at improving L2 English speaking fluency effectively.

1.2 Background of Study

"Is Fluency, Like Beauty, in the Eyes (and Ears) of the Beholder?"
(Freed, 2000, p. 243)

The current research began with a single question: "Are L2 classroom evaluators using the same idea when they discuss fluency?" This question has to be answered because if fluency is in the EARS of the beholder, then diverse evaluators will operationalize the term fluency differently, which significantly impairs the validity of L2 fluency evaluation. Following that, a new question occurred. "If so (or if not), what causes raters to perceive and evaluate L2 fluency differently (or similarly) and what underlying mechanisms account for the disparities (or similarities)?" To begin answering these questions, two fundamental questions must be addressed in advance: "What is fluency?" and "Why does fluency matter?"

What is fluency? What do we mean by the term fluency? Koponen and Riggenbach (2000) pointed out a conceptual metaphor

underlying the meaning of fluency, namely that "language is motion" (p. 7). The metaphor focuses on those aspects of speech having to do with its fluidity or flowing quality (Segalowitz, 2010), indicating fluidity is the predominant underlying idea when people discuss fluency. In this sense, fluency is often understood to refer to the flow and smoothness of delivery (Chambers, 1997). Moreover, fluency, in the broadest sense, can be equated with overall language proficiency, including accuracy and complexity (Lennon, 1990). It appears in phrases such as "He had earlier spent several years in America and spoke fluent English." However, this nontechnical use of the term contrasts with the more restricted linguistic sense of fluency, which refers to one of several identifiable components of language abilities that can be evaluated independently (Freed, 2000). For example, language practitioners often distinguish between fluency and accuracy, implying that to be fluent is not necessarily to be accurate in certain circumstances. It occurs in the phrase, "Joan knows French grammar perfectly, but she does not speak the language fluently" (Freed, 2000, p. 244). Meanwhile, fluency is restricted to temporal measures in the narrowest sense, such as length and number of pauses and the number of hesitations and repetitions (De Jong & Perfetti, 2011).

The evidence presented demonstrates that fluency is not a concept used consistently, either globally or componentially, and there is no single definition of fluency (Koponen & Riggenbach, 2000). Consequently, definitions of different kinds of fluencies or various fluency components must be unambiguously expressed in linguistic

terms to ensure consistency before implementing definitions within the L2 classroom (Koponen & Riggenbach, 2000).

Next, why does fluency matter? Although the concept is difficult to define and has not been well understood (Kormos & Dénes, 2004; Segalowitz, 2010), being fluent in the target language is a primary goal for many L2 learners (De Jong & Perfetti, 2011), and fluency is a significant construct for assessing L2 proficiency (Bosker et al., 2013; Préfontaine et al., 2016). According to Bosker et al. (2013), oral fluency is a critical indicator of a person's language competency. They noted that fluency is frequently assessed in various professional examinations (e.g., the TOEFL), which might have long-lasting effects on a person's life, such as job-seeking and university admission. Additionally, Préfontaine et al. (2016) also showed that fluency is a significant construct in evaluating language proficiency. It can be found on a number of rating scales in high-stakes tests and descriptors of L2 competency levels.

Meanwhile, in the Korean EFL context, being fluent in English is often regarded as one of the biggest goals in learning English (Lim & Hwang, 2019). The English curriculum in the national curriculum for primary and secondary schools (Ministry of Education, 2015) emphasized cultivating and evaluating English oral fluency, noting to "instruct students to develop fluency and accuracy through various meaningful communicative activities and cultivate the ability to communicate and exchange meanings in practical situations" (p. 42). It also emphasized fluency over accuracy in terms of speaking

evaluation by putting "evaluate (students' speech) by emphasizing fluency rather than accuracy, and focusing on being able to confidently speak appropriate content" (p. 42).

In this context, it is challenging but necessary to investigate perceptions of L2 fluency by examining whether evaluation raters share the same underlying concept of fluency and by tracking how fluency is perceived and judged by various raters across different task types and proficiency levels throughout the L2 assessment process because this must be the first step toward ensuring the validity of fluency assessment. Taking into account task types and L2 speakers' proficiency levels on raters' perceptions of fluency is particularly important since it reflects a real-world L2 classroom in which students with varying proficiency levels perform various tasks. However, previous researches on rater variability have primarily focused on the effect of raters' linguistic (e.g., Brown, 1995; Fayer & Krasinski, 1987; Gui, 2012; Kim, 2009; Zhang & Elder, 2011) and professional (e.g., Chalboub-Deville, 1995; Hadden, 1991) backgrounds on their rating patterns, without considering tasks or speaker proficiency levels. Additionally, there is a lack of research that comprehensively integrates raters from diverse backgrounds, such as peer students, into the L2 classroom context. Given that peer assessment, which fosters student-centered learning and develops students' higher level of reasoning and cognitive thought (Birdsong & Sharplin, 1986), is becoming more critical as an alternative assessment method, it is necessary to incorporate peer students into the speaking

assessment process and investigate the possibility of peers' serving as competent fluency raters. To the best of my knowledge, no study has examined comprehensive aspects of L2 fluency perceptions that included various evaluators (e.g., native English teachers, non-native English teachers, and peer students). Thus, there are academic gaps in the research about the extent to which three groups of judges (native English teacher group, non-native English teacher group, peer group) perceive and rate L2 fluency across task types and speakers' proficiency levels in the Korean EFL setting.

Historically, research on L2 fluency started in the 1970s and 1980s. From the 1990s, L2 fluency research saw considerable growth in SLA and language testing (Kahng, 2022). The overarching purpose of previous L2 fluency research has been identifying speech features that function as reliable indicators of L2 fluency (De Jong, 2018; Kahng, 2022). Various approaches have been used to achieve this goal, depending on how fluency is conceptualized. The majority of previous research (e.g., Bosker et al., 2013; Cucchiarini et al., 2002; Derwing et al., 2004; Kormos & Dénes, 2004; Rossiter, 2009) treated fluency as a construct to be defined by listeners (Kahng, 2022). Typically, these studies explored the link between objective measures of fluency (i.e., utterance fluency) and subjective ratings of fluency (i.e., perceived fluency) (Segalowitz, 2010). On the other hand, there have been studies that viewed fluency from the speaker's perspective by exploring the relationship between objective measures of utterance fluency to speakers' overall proficiency (e.g., Ginther et al., 2010;

Iwashita et al., 2008; Kahng, 2014) or by tracking speakers' gains in utterance fluency (e.g., Lennon, 1990, Towell et al., 1996). Meanwhile, to explore the mechanisms of fluent speech from the speakers' point of view, few studies (e.g., De Jong et al., 2013; Kahng, 2014) have related objective measures of fluency to cognitive fluency, the underlying cognitive process responsible for the utterance fluency (Segalowitz, 2010). The current study, based on Segalowitz's (2010) three facets of fluency (i.e., perceived, utterance, and cognitive fluency), attempts to investigate 1) perceived fluency by examining how three groups of judges perceive speaking fluency and 2) utterance fluency by tracing the underlying mechanisms that account for the disparities observed among the three rater groups, and ultimately aims to understand L2 fluency comprehensively.

1.3 Research Questions

The purpose of this study is to determine how native teachers, non-native teachers and peers perceive L2 speaking fluency differently and to establish which acoustic properties influence each rater group's perception of fluency best, with the goal of identifying intertwined and multidimensional aspects of L2 fluency perception in the Korean EFL context. However, to date, there have been no studies that compare the judgments of two EFL teacher groups and a peer group concurrently, taking task types and speakers' proficiency levels into account, or investigate the in-depth impressions reported

to have influenced each group's ratings of L2 learners' oral fluency. Additionally, a majority of the previous studies have concentrated on rater variability (e.g., Brown, 1995; Fayer & Krasinski, 1987; Kim, 2009) or the relationship between raters' subjective impressions and objectively measured attributes (e.g., Kormos & Dénes, 2004; Rossiter, 2009), with few studies attempting to combine these two academic foci and observe multifaceted characteristics of fluency perception in diverse raters holistically. In light of these research needs, the following research questions will be addressed in the present study:

1. How does perceived fluency of native teachers, non-native teachers and peers differ?

1-1. Do fluency ratings by native teachers, non-native teachers and peers differ across groups?

1-2. How do fluency ratings of the three groups of judges differ across two task types at speakers' different oral proficiency levels?

1-3. Which impressions of L2 speech affect the three groups of listeners' fluency judgments, and how do they differ?

2. How does utterance fluency of native teachers, non-native teachers and peers differ?

2-1. Which measures of utterance fluency are most related to native teachers', non-native teachers', and peers' fluency ratings?

2-2. Which utterance fluency model, represented by sets of temporal features, best explains native teachers', non-native teachers',

and peers' fluency ratings?

2-3. Which measures of utterance fluency differentiate students with low, intermediate, and high levels of perceived fluency among the three types of raters?

These two research questions are not entirely separate. Employing three facets of fluency (Segalowitz, 2010), the two research questions together will trace the differences in fluency perceptions among the three groups and lead to a comprehensive understanding of L2 fluency assessment in the Korean EFL context. Answers to the first research question will reveal how the three groups of raters perceive and rate L2 fluency differently. The fluency perception discrepancies which will be found in the first research question will be further supported and accounted for by identifying objectively measured acoustic features related to each raters' fluency ratings in the second research question. In other words, with the two research questions, the current study attempts to investigate the relationship between (1) people's perceptions of speakers' fluency (as reflected in ratings on the fluency assessment scale) and (2) actual fluency-related phonetic features of speech (Ejzenberg, 2000). As mentioned, recognizing rater variability in fluency perceptions and understanding how the disparities in perceived fluency relate to acoustic correlates will be the first step toward valid assessment of speaking fluency and will be a necessary component in developing fluency in the EFL classroom.

1.4 Organization of the Dissertation

The dissertation is comprised of five chapters. Chapter 1 describes the purpose of the current study and states the rationale and research questions. Chapter 2 reviews the previous literature related to the current study. Chapter 3 (Study 1) reports the quantitative and qualitative studies that addressed the first research question by examining the rater differences in perceived fluency. Chapter 4 (Study 2) relates fluency perceptions with acoustic features, finding attributes that most accurately predict the fluency rating of each rater group and distinguish speakers' fluency levels. Chapter 5 summarizes the findings of the two studies (Study 1, 2), along with pedagogical implications and suggestions for future research.

CHAPTER 2. LITERATURE REVIEW

This chapter provides the theoretical background for the study. Section 2.1 gives an overview of the definitions of fluency and an introduction to the various methods for measuring fluency. The findings of research on perceived fluency, focusing on rater and task variables, are presented in Section 2.2. Lastly, section 2.3 discusses in detail the findings of research on utterance fluency and its relationship to perceived fluency, focusing on the reliable predictors of fluency.

2.1 Defining and Measuring Fluency

Several researchers have pointed out that fluency constitutes a crucial aspect of understanding L2 performance and proficiency (Kahng, 2022), and attaining high levels of fluency is an essential goal for many language learners (Tavakoli, 2011). However, as fluency is not a concept used consistently, either globally or componentially (Koponen & Riggenbach, 2000), it is crucial to define fluency unambiguously in a linguistic term and to explore reliable ways to measure fluency with a solid theoretical framework. In addition, to better understand the term, it is also necessary to examine how fluency fits into the overall picture of oral production in connection

with the other linguistic components.

Lennon (1990) distinguished between fluency in the broad and narrow senses. In the broad sense, fluency is often regarded as someone's overall proficiency (Chambers, 1997) or "impression on the listener's part that the psycholinguistic processes of speech planning and speech production are functioning easily and efficiently" (Lennon, 1990. p.391). Fillmore (1979) provided one of the broadest conceptualizations of fluency as defining four different kinds of fluency that individuals may consider while making fluency judgments. One is the ability to talk for an extended period of time with few pauses, and the second is the ability to package a message into semantically dense sentences. The third is the ability to communicate effectively in various social circumstances, and the fourth is the ability to use the language creatively and imaginatively (Segalowitz, 2010). In short, in the broad sense, fluency appears to function as a cover term for oral proficiency and represents the highest point on a scale that measures spoken command of a foreign language (Lennon, 1990).

However, fluency in the narrow sense is often considered one component of speaking proficiency, which Lennon (1990) defined as the "rapid, smooth, accurate, lucid, and efficient translation of thought or communicative intention under the temporal constraints of online processing" (p. 26). As one of the distinguishable aspects of language proficiency, fluency can be complemented by other linguistic components, such as accuracy and complexity (Housen et al., 2012), and is frequently understood in relation to these components.

Complexity, accuracy, and fluency (CAF) have been identified as significant research variables in L2 linguistic research and used both as performance descriptors as well as indicators of L2 learners' proficiency underlying their performance (Housen & Kuiken, 2009). Complexity has often been defined as "the extent to which the language produced in performing a task is elaborate and varied" (Ellis 2003, p. 340), accuracy as "the ability to produce error-free speech" (Housen & Kuiken, 2009, p. 461), and fluency as "the ability to process the L2 with the extent to which the language produced in performing a task manifests pausing, hesitation, or reformulation" (Ellis 2003, p. 342). These three dimensions are interdependent in L2 performance (Skehan, 1998), and fluency has also been understood in connection with the other two constructs. For instance, L2 researchers have observed that an increase in fluency in L2 acquisition may occur at the expense of developing accuracy and complexity. Similarly, fluency has been considered an aspect of L2 production that competes for attentional resources with accuracy, while accuracy competes with complexity (Housen & Kuiken, 2009).

Meanwhile, fluency has often been explored with Munro and Derwing's (1995) tridimensional model of oral production. Munro and Derwing (1995) proposed and established three critical constructs of L2 oral production, which included intelligibility (the degree to which a listener understands a speaker's intended message), comprehensibility (the degree of effort required for a listener to comprehend L2 speech), and accentedness (the degree to which the speaker's accent deviates

from the expected accent) (Levis, 2020). The three constructs have been invoked repeatedly in researches on L2 fluency and are often used to make connections to fluency judgments (e.g., Derwing et al., 2008). Previous studies (e.g., Crowther et al., 2015, Derwing et al., 2004; Derwing et al., 2008) have consistently demonstrated that comprehensibility seemed to be closely tied to learners' fluency, suggesting both fluency and comprehensibility were interrelated and thus developed together over time. However, some researchers (e.g., Derwing & Munro, 1997; Munro & Derwing, 1999) repeatedly reported that the degree of accentedness had little influence on listeners' perception of fluency and comprehensibility (French et al., 2020). According to Derwing et al. (2004), it appeared that increased fluency was less likely to lead to a perception of reduced accentedness, possibly because accentedness judgments were based more heavily on linguistic phenomena such as segments and prosodic elements.

Lastly, in the narrowest sense, fluency is further restricted to temporal measures, such as length and number of pauses, and the number of hesitations and repetitions (De Jong & Perfetti, 2011). The majority of the previous researches on L2 fluency (e.g., Bosker et al., 2013; Chamber, 1997; Derwing et al., 2004; Kormos & Dénes, 2004; O'Brien et al., 2007; Rossiter, 2009) also conceptualized fluency as a temporal performance phenomenon, manifested primarily as speed and hesitation. The current research concerned this definition of fluency (restricted sense of fluency) so that the concept and components of fluency can be unambiguously described in a linguistic term. Along

with the definition of the term fluency, the following introduces a theoretical framework used in the current dissertation.

It has been often claimed that a significant barrier to the systematic investigation of L2 fluency is a lack of a comprehensive theoretical framework to understand fluency (Segalowitz, 2010). In recent years, however, Segalowitz's (2010) model of fluency and Skehan's (2003) framework for measuring fluency have successfully expanded our conceptual understanding of fluency, providing the discipline with more valid and reliable indices of fluency (Tavakoli et al., 2020). Segalowitz (2010) identified fluency from a cognitive perspective (Bosker et al., 2013), introducing three fluency aspects: cognitive fluency, utterance fluency, and perceived fluency. Cognitive fluency refers to "the efficiency of operation of the underlying processes responsible for the production of utterances" (Segalowitz, 2010, p. 165). It is commonly understood that producing an utterance requires managing several separate, interacting cognitive processes, and the coordination must be done rapidly and effectively. Second, utterance fluency refers to "the features of utterances that reflect the speaker's cognitive fluency" (p. 165), which can be acoustically measured. It has to do with the acoustic features of an utterance. It refers to the temporal, pausing, hesitation, and repair characteristics, which are actual properties of the utterance, not just impressions a listener might have had. The last, perceived fluency, is "the inferences listeners make about speakers' cognitive fluency based on their perceptions of their utterance fluency" (p. 165). In other words,

perceived fluency is a judgment made about speakers based on impressions drawn from their speech.

In addition, according to Skehan (2003), utterance fluency can be consistently measured using indices related to three key aspects of fluency: breakdown fluency, speed fluency, and repair fluency. Breakdown fluency is concerned with the degree to which a continuous speech signal is interrupted; speed fluency is described as the rate and density of speed delivery; and repair fluency is concerned with the number of corrections and repetitions present in speech (Skehan, 2003).

Segalowitz's (2010) three facets of fluency were employed as a theoretical framework for the current dissertation because they accurately describe how speakers generate L2 speech, how listeners perceive fluency, and how listeners' perception is related to acoustic features. Furthermore, L2 speakers' utterance fluency features were systematically classified and rigorously measured based on Skehan's (2003) taxonomy.

Thus far, the definitions of fluency and the theoretical framework have been discussed. Just as there are various definitions of fluency, there are several ways of measuring fluency. Previous studies have employed three methods to determine valid utterance fluency attributes and assess fluency reliably. First, multiple studies of L2 fluency have looked at how L2 learners at various proficiency levels developed their oral fluency longitudinally in different educational contexts (e.g., Lennon, 1990; Towell et al., 1996). Second, many researchers have conducted comparison experiments of fluent

and non-fluent speakers' speech (of L1 and L2 speech), identifying what components make them different (Cucchiarini et al., 2002; Iwashita et al., 2008; Kahng, 2014, Riggenbach, 1991; Saito et al., 2018). Finally, a large number of studies have investigated the association between listeners' impressions of fluency (perceived fluency) on L2 speech samples and utterance fluency by correlating subjective fluency ratings with objectively measured temporal features (e.g., Bosker et al., 2013; Derwing et al., 2004; Kormos & Dénes, 2004; O'Brien et al., 2007; Rossiter, 2009).

In line with previous research methodology, the current study investigates which utterance fluency measures most significantly affect listeners' perceptions of fluency by correlating perceived fluency to utterance fluency. Examining the relationship between utterance fluency and perceived fluency will provide an answer to one of the most commonly asked questions in L2 literature: What makes speech sound more fluent? Fluent speakers are often thought to produce words faster, with longer runs, less hesitation, and fewer lengthy and disruptive pauses (Tavakoli, 2011). Similarly, Kormos and Dénes (2004) suggested that fluency could be best described as a fast, smooth, and effortless performance. Pieces of evidence suggest that fluency is essentially a temporal phenomenon, with utterance features affecting fluency judgments (Tavakoli, 2011). Thus, it is crucial to study the extent to which utterance fluency attributes are connected with perceived fluency. In the following, literature on perceived fluency and utterance fluency will be reviewed.

2.2 Perceived Fluency

2.2.1 Rater Variables on Fluency Ratings

Rating speaking fluency in the L2 classroom primarily requires human raters. Due to the inherent nature of human rating, it is plausible that the final scores assigned by a human rater may be influenced by variables inherent to that rater (McNamara, 1996). Thus, it is unavoidable that rater variables such as L1 background (i.e., native or non-native speaker of English) and linguistic training (i.e., language expert or naïve rater) pose issues about test reliability. In most Korean EFL educational settings, a team of one non-native teacher and one native teacher evaluates L2 fluency jointly, or one non-native EFL teacher serves as the sole source of English fluency ratings. As a result, rater variability has long been a significant concern among Korean EFL language practitioners questioning the validity and reliability of language performance tests (Yu, 2010).

Meanwhile, the growing interest in rater variability has also raised another issue: questions of eligibility. In particular, the controversy over whether native speakers should be the sole norm makers in language assessment has inspired heated discussion among language experts (Kim, 2009). Reflecting the current status of English as a global language of communication, language professionals doubt whether native speakers should be the only standard accepted

(Taylor, 2006). It has also been argued that non-native teachers can be considered more compelling or sensitive assessors than native teachers in certain circumstances, such as in "expanding circle countries" (Kachru, 1985). In the meantime, another school of thought that is growing in popularity recently argues that both learners and teachers must be involved in the assessment methods, procedures, and outcomes, emphasizing the benefits of including peer assessment into traditional assessment procedures (Cheng & Warren, 2005). Under these circumstances, it is getting more critical for EFL language experts to investigate the extent to which non-native EFL teachers or other assessment participants, such as peer students, can serve as credible fluency evaluators in the same manner that native teachers do.

While reviewing the previous research on rater variability, it was found that most of the researches focused on rater differences in English speaking proficiency, including fluency as a linguistic criterion, rather than on assessing fluency independently. Thus, the relationship between different raters and oral proficiency will be investigated along with oral fluency. There has been a great deal of previous research that has focused on raters' linguistic background (e.g., Brown, 1995; Fayer & Krasinski, 1987; Gui, 2012; Kim, 2009; Zhang & Elder, 2011) and professional backgrounds (e.g., Chalboub-Deville, 1995; Hadden, 1991). A few studies examined raters' language and professional backgrounds simultaneously (Barnwell, 1989; Galloway, 1980; Rossiter, 2009). Another line of research has examined the validity of peer evaluation on speaking

performance (e.g., Cheng & Warren, 2005; Patri, 2002). In the following, each branch of the previous researches will be reviewed.

Firstly, extensive research has explored raters' rating patterns with different linguistic backgrounds (e.g., native vs. non-native raters). Despite some findings being inconsistent and occasionally contradictory, the two rating groups appeared to have a similar rating pattern. The majority of the previous researches reported that native and non-native raters agreed on overall quantitative results but disagreed on specific criteria. For example, Fayer and Krasinski (1987) examined the differences in how seven Puerto Rican students' English oral proficiency was evaluated by English native speakers and Spanish native speakers (non-native English speakers). Both groups of listeners were asked to rate the speaker on a five-point scale for overall intelligibility and discrete components such as grammar, pronunciation, and hesitations. They discovered that both groups of raters assigned similar ratings to overall intelligibility, with no statistically significant mean difference in that area. However, non-native raters tended to be more critical of students' linguistic forms, with pronunciation and hesitation being the most distracting elements for both native and non-native rater groups. Brown (1995) investigated the effect of raters' occupational and linguistic backgrounds on their assessments of an occupation-specific oral language test, the Japanese Language Test for Tour Guides. Thirty-three assessors participated, including native and non-native Japanese speakers with experience in teaching Japanese as a foreign

language or guiding tours in Japanese. The results revealed that native speakers were generally harsher than non-native speakers, although there were no statistically significant differences in the overall grades. There were, however, considerable disparities in the grades assigned to specific categories. Non-native speakers were shown to be significantly harsher on politeness and pronunciation than native speakers.

Similarly, Kim (2009) used a mixed-method approach to study how native speakers and non-native speakers evaluated students' oral English performance. Twelve native English speakers from Canada and twelve Korean teachers were asked to rate the spoken English performance of ten Korean university students. Both native and non-native groups maintained acceptable internal consistency and severity levels, as evidenced by the quantitative data. However, the qualitative analysis revealed that the native teacher group provided more detailed and elaborate written comments on pronunciation, grammar use, and the accuracy of transferred information than non-native teacher group did. Similarly, Zhang and Elder (2011) examined how Chinese native speakers' judgments of Chinese test takers' oral English proficiency differed from those of English native speakers. The findings indicated no significant difference in the two rater groups' holistic judgments on the speech samples and a high degree of agreement on the construct components of oral English proficiency between the native and non-native rater groups. However, an analysis of their comments revealed that they differed on

constructs of oral proficiency they thought to be more critical. Non-native speakers, in particular, placed a higher premium on general linguistic resources than on other variables, whereas native speakers placed a higher value on demeanor, interaction, and compensation strategy. To summarize, in terms of linguistic backgrounds, researchers have mainly investigated whether native and non-native speakers evaluate L2 students' oral proficiency similarly or not, and it appears that both rater groups have similar quantitative rating patterns. However, it was noted that there were some discrepancies in specific scoring criteria and how they reacted to the oral performance of L2 students.

Secondly, investigations with raters from various professional backgrounds (e.g., experts vs. naïve raters) have been conducted. However, the findings from this research did not exhibit consistency. Several studies have found that linguistically trained raters provided harsher ratings than naïve raters (e.g., Galloway, 1980; Hadden, 1991), while other studies have found the opposite (e.g., Barnwell, 1989). Galloway (1980), for example, examined whether oral communicative proficiency judgments differed between communities of native and non-native Spanish speakers with and without teaching experience. The findings indicated that native Spanish-speaking teachers assigned harsher scores than non-teaching native speakers. In contrast, Barnwell (1989) discovered that when naïve native Spanish speakers were compared to raters trained by the American Council for the Teaching of Foreign Languages (ACTFL), untrained native Spanish

speakers provided more severe assessments than trained Spanish raters, which contradicted Galloway's (1980) findings.

Rossiter (2009), aiming explicitly to compare the fluency judgments of different raters, found no significant effect of rater differences in terms of professional background. She asked six expert native speakers, 15 novice native speakers (non-expert), and 15 advanced non-native speakers to judge 24 ESL learners who narrated picture stories at Time 1 and again ten weeks later at Time 2. The results indicated that even though novice native speakers had much higher fluency ratings than advanced non-native speakers, the ratings of expert native speakers were not significantly different from those of novice native speakers or non-native speakers. In sum, various researches suggest contrasting results, and more convincing evidence is needed to determine how various raters with different professional backgrounds rate L2 speech.

Lastly, a few studies have been conducted on the reliability of peer assessment compared to the evaluations given by teachers when measuring language proficiency (e.g., Cheng & Warren, 2005) and oral skills (e.g., Patri, 2002). Comparing peer and teacher assessments, Cheng and Warren (2005) investigated students' attitudes towards assessing English language proficiency and other aspects of the performance of their peers (e.g., preparation, content, organization, and delivery). The results showed that although peer students showed a less favorable attitude toward judging their peers' language proficiency, they did not evaluate their peers significantly differently

than they did for the other assessment criteria. Additionally, for oral presentation, which was one of the three projects used (i.e., seminar, oral presentation, and report writing), the data indicated no statistically significant differences in the language proficiency criterion between teachers and peers. However, considering all projects and follow-up interviews with students, the researchers concluded that students and teachers marked and interpreted oral and written language proficiency differently, indicating that students did not assess the same elements as their teachers.

Patri (2002), in a similar vein, examined the agreement between teacher and peer assessments of students in the presence of peer feedback. Fifty-six undergraduate students with an ethnic Chinese background were requested to give an oral presentation and participated in a training and practice session on peer assessment. The findings indicated that when assessment criteria were clearly defined and peer feedback was provided, peer students could make oral presentation judgments comparable to those made by teachers. The researchers concluded that peer feedback aided in increasing the correlation between teacher and peer assessments, implying that peer assessments might supplement teacher assessments in the context of oral skills under specific circumstances. In sum, some researchers have generally suggested that peer students and teachers often rated and interpreted speaking proficiency differently, indicating that students did not assess the same elements as their class teachers. However, peer assessment might be validated if certain conditions are met.

While the previous studies provide evidence for the effects of rater variables (e.g., linguistic and professional backgrounds) on L2 speaking assessments, additional studies are required for the following reasons. To begin, except for Rossiter's (2009) work, most previous researches have focused on overall English proficiency as judged by native and non-native raters rather than on fluency in particular. As a result, it is challenging to determine how individual raters with varying EFL backgrounds judge L2 fluency differently. Secondly, previous researches have primarily examined whether native raters' fluency ratings differed from those awarded by non-native raters, with only a few studies (e.g., Cheng & Warren, 2005; Patri, 2002) comparing peer students' rating patterns to those of both teacher groups. Thus, the current study intends to broaden the range of rater variability by considering peer student judgments of fluency and comparing them with EFL teacher groups (native and non-native teachers). Thirdly, few studies have been undertaken to determine the effect of L2 speakers' overall speaking proficiency on native, non-native, and peer raters' judgments of fluency. For example, questions such as "Are non-native teachers' fluency ratings comparable to those of native raters when evaluating lower-level L2 speakers?" or "How are peer students' fluency judgments different from those of native and non-native teacher raters when evaluating high proficiency level students?" were not fully addressed. In addition, it is still unclear whether non-native teachers and peer raters will award ratings equivalent to native raters at each proficiency level

(low, mid, high). Thus, the current study attempts to fill the gap by examining the relationship between three judges' fluency ratings and by identifying whether there is a particular level of oral proficiency for which native teachers, non-native teachers, and peer students exhibit different rating patterns.

2.2.2 Task Variables on Fluency Ratings

To better understand the three rater groups' perceptions of L2 fluency, the current study examines the effect of task on raters' fluency perceptions by comparing the fluency ratings of three distinct rater groups across two different tasks. The majority of previous literature on fluency ratings and task differences revealed that there were certain differences in fluency that may be directly attributed to task's features (e.g., Chalhoub-Deville, 1995; Derwing et al., 2004; Ejzenberg, 2000).

Derwing et al. (2004) investigated whether there were variations in fluency ratings among three task types (i.e., picture narration, monologue, dialogue task). They obtained L2 speech samples from twenty beginner Mandarin English learners and asked twenty-eight untrained judges to rate fluency and prosody. The results demonstrated that perceptions of L2 speakers' fluency varied between tasks, as ratings on the picture narration task were considerably lower than ratings on the monologue or dialogue tasks, which were not significantly different from one another. They

attributed the differences among tasks to task-dependent variability in the speaker's degree of freedom, such as selecting lexical items, structures, and content. The picture narration task inherently imposed restrictions that could not be entirely avoided. For example, the L2 students were required to describe the sequence of specific pictures, which many found difficult, while the monologue and conversation tasks may have enabled them to rely on scaffolding and formulaic sequences. Furthermore, L2 speakers could control the content in monologue and conversation tasks by telling familiar stories and avoiding possible trouble points that might lead to a communication breakdown.

In a similar vein, Ejzenberg (2000) investigated whether oral fluency was affected by the speaking context, such as task types. The forty-six young adults, classified into three proficiency levels, completed four different speaking tasks (i.e., cued dialogue, uncued dialogue, cued monologue, and uncued monologue). Four trained raters were asked to provide a holistic assessment of each task's fluency for the forty-six participants. The percentages of participants who received high fluency ratings for different tasks were as follows: uncued dialogue (50%), cued dialogue (37%), uncued monologue (30.4%), and cued monologue (21.7%), suggesting that L2 learners would be perceived as more fluent in an interaction with a native speaker (e.g., dialogue situations). It was found that in an interaction context, L2 learners were able to scaffold on the interlocutor's productions, but monologue tasks imposed much greater cognitive

demands, resulting in lower fluency scores. By demonstrating that different tasks might have a differential effect on a subject's oral fluency display, she emphasized the significance of utilizing a variety of tasks for evaluation purposes.

Several other researchers attempted to examine the effects of tasks on L2 speaking proficiency in general, not simply fluency. Chalhoub-Deville's study (1995), for example, contrasted three frequently used L2 oral task types (oral interview, narration, and read-aloud) to elucidate the criteria underlying learners' L2 oral abilities across the three task types. Three native speaker judges scored each of the 18 speech samples holistically. Three distinct dimensions underpinning L2 holistic oral scores across tasks were identified: grammar-pronunciation, creativity in presenting information, and amount of detail provided. The researcher highlighted variability in language performance across tasks, revealing that subjects performed differently across the three tasks. It was seen that performance variability could be attributed to the task's varying demands on the subjects' linguistic and cognitive processes. The study concluded that language dimensions underlying L2 oral abilities could be presented differently depending on the task, proposing that L2 researchers consider employing context-specific rating scales.

To summarize, the previous researches on fluency evaluations and task differences consistently demonstrate that tasks have a systematic effect on fluency scores. The inherent task-dependent variability possibly imposed varying degrees of linguistic and

cognitive load on L2 speakers, which appeared to affect both their speech production and raters' assessments of fluency. While the existing studies provide evidence for the effect of task types on L2 speaking assessments, additional studies are necessary for the following reasons. First, little research has been conducted to determine task effects on fluency judgments across different speakers' proficiency levels. For instance, some questions such as "How do different task types affect fluency judgments of native teachers, non-native teachers and peers when they evaluate lower-level speakers?" remain unanswered. Second, the previous researchers have only examined a subset of the listener population. Derwing et al. (2004), for example, used twenty-eight untrained native speaker raters, while Ejzenberg (2000) used only four trained native speaker raters. Due to the scarcity of research involving diverse judge groups, the effect of task types on fluency ratings of different rater groups has not been well examined. Additionally, the previous investigations used a small sample size of judges. Thus, this dissertation recruits a wide variety of listener groups with a relatively large number of judges and considers speakers' oral proficiency levels to gain a holistic understanding of task effects on various rater groups' fluency ratings.

2.3 Utterance Fluency

2.3.1 Predictors of Utterance Fluency

To date, with exploring a wide variety of contexts, participants, tasks, and judges, several previous pieces of literature attempted to find which temporal features best predict listeners' perceptions of fluency. As a first step, several utterance fluency features that have been proven to influence perceived fluency significantly will be introduced (see Table 2.1), and they were examined based on three notions of fluency: speed fluency, breakdown fluency, and repair fluency (Skehan, 2003).

Table 2.1 Measures of Utterance Fluency[1)]

	Measure	Formula
Speed fluency	Speech rate (unpruned)	Number of syllables / total time
	Speech rate (pruned)	(Number of syllables - number of filled pauses and repairs) / total time
	Articulation rate	Number of syllables / (total time - silent pausing time)
	Mean length of run	Mean number of syllables between silent pauses
	Phonation time ratio	Speaking time[a] / total time

1) Most of the variables are from Kahng (2022). I have divided speech rate into unpruned and pruned speech rate, and added pause rate variables (within a clause / at a clause boundary).

Breakdown fluency	Mean length of silent pauses	Pausing time / number of silent pauses
	Mean length of filled pauses	Pausing time / number of filled pauses
	Number of silent pauses (per minute)	Number of silent pauses / total time
	Number of filled pauses (per minute)	Number of filled pauses / total time
	Pause rate within a clause	Number of pauses within a clause / number of clauses
	Pause rate at a clause boundary	Number of pauses at a clause boundary / number of clause boundaries
Repair fluency	Number of repetitions (per minute)	Number of repetitions / total time
	Number of corrections (per minute)	Number of corrections / total time

[a]speaking time is equal to total time minus silent pausing time

2.3.1.1 Speed Fluency

A large body of L2 fluency literature has shown clear correlations between speed fluency and perceived fluency. All rate-related variables are based on the number of syllables or words produced by a certain amount of time or segment, while they differ with respect to their representation of rate either with or without filled and / or silent pauses (Ginther et al., 2010). The majority of studies found that unpruned speech rate (i.e., number of syllables / total time, including pause time), pruned speech rate (i.e., number of syllables - number

of filled pauses and repairs / total time), mean length of run (i.e., mean number of syllables between two silent pauses), and phonation time ratio (i.e., speaking time / total time) were consistently related to L2 perceived fluency. Additionally, articulation rate (number of syllables / total time - silent pause time) is commonly regarded as a significant predictor of fluency in a large number of L2 fluency studies with a few exceptions (e.g., Kormos & Dénes, 2004), which found no correlation between articulation rate and fluency score.

Lennon (1990) asked ten native English teachers to judge four German advanced EFL learners' fluency at the start and end of a six-month residence in Britain. According to the findings, the perceived gains in fluency over the six months were due to an increased speech rate. In particular, the mean length of run increased over time in the productions of three participants. Furthermore, Cucchiarini et al. (2002) found that speed fluency features significantly impact listeners' perceived fluency at different proficiency levels. They had ten teachers of Dutch rate fluency of the spontaneous speech from 57 beginner and intermediate Dutch learners. The researchers compared the subjective ratings of fluency to objective fluency indicators and discovered that speech rate and articulation rate were the best indicators of perceived fluency for beginner learners. In contrast, for intermediate learners, the mean length of run was more predictive of fluency. In a similar vein, in a study performed by Kormos and Dénes (2004), three native and three non-native English teachers rated sixteen Hungarian speakers'

fluency. The researchers related the six raters' fluency judgments to ten temporal variables. The best predictors of fluency, according to correlation analyses, were speech rate, mean length of run (250ms pause), phonation-time ratio, and the number of stressed words emitted per minute (pace), which could explain between 60 percent to 80 percent of the variance in the fluency scores. Interestingly, the rank order correlation results indicated that articulation rate was not related to fluency scores.

Recently, Préfontaine et al. (2016) used mixed-effects modeling to examine the relationship between raters' perceptions of L2 fluency in French and temporal features. Forty adult learners of French with varying proficiency levels were asked to complete three different types of narrative tasks, and four utterance measures (i.e., articulation rate, mean length of run, pause frequency, and average pausing time) were extracted from each performance. Eleven untrained judges assessed the learners' fluency and investigated which utterance fluency measures best predicted the raters' scores. In line with the results of the previous research, the mean length of run was found to be the most significant feature in raters' fluency judgments. However, contrary to Kormos and Dénes' (2004) findings, it was discovered that articulation rate was one of the essential factors in predicting perceived fluency.

To sum up, until now, a large body of research has investigated the relationship between measures of utterance fluency and perceived fluency to find which features of L2 utterance are

strong predictors of L2 fluency judgment. The majority of the previous studies indicate that speed fluency measures, such as unpruned speech rate (e.g., Cucchiarini et al., 2002; Kormos & Dénes, 2004; Lennon, 1990; Riggenbach, 1991), pruned speech rate (e.g., Derwing et al., 2004; Lennon, 1990; Rossiter, 2009), mean length of run (e.g., Cucchiarini et al., 2002 (only for intermediate learners); Kormos & Dénes, 2004, Lennon, 1990; Préfontaine et al., 2016), articulation rate (e.g., Cucchiarini et al., 2002 (only for beginner learners), Préfontaine et al., 2016), and phonation time ratio (e.g., Cucchiarini et al., 2002; Kormos & Dénes, 2004) consistently play a crucial role in predicting L2 perceived fluency. Despite the fact that speed measures are the most powerful predictors of fluency, little is known about the extent to which each speed measure is related to three different types of raters in the Korean EFL setting. Thus, the current research attempts to discover the speed measures that are most closely related to each rater group's fluency judgments and to determine the extent to which they influence each rater group. In the next section, the literature on breakdown fluency will be followed.

2.3.1.2 Breakdown Fluency

A number of L2 studies have linked breakdown fluency (e.g., pause phenomena) with perceived fluency (e.g., Bosker et al., 2013; Cucchiarini et al., 2002; Ginther et al., 2010; Kormos & Dénes, 2004). The majority of previous research investigated pause phenomena

focusing on three aspects: pause frequency (number of pauses), pause length (duration of pauses), and pause distribution. However, not only have researchers not been consistent in the way they have operationalized breakdown variables, such as silent pause cut-off (ranging from 200 ms to 1 sec) (Segalowitz, 2010) but also findings on pause frequency and length are mixed, which makes studies on pause phenomena more complicated. The following examine the previous research concerning pause frequency and length.

Previous studies found that both pause frequency and pause length were significant, and they were all negatively associated with fluency ratings (e.g., Bosker et al., 2013) and holistic English proficiency scores (e.g., Ginther et al., 2010). Ginther et al. (2010) examined the relationships selected temporal fluency measures, including number and duration of pauses and holistic scores on a semi-direct measure of oral English proficiency. The spoken responses of 150 respondents to one item on the Oral English Proficiency Test (OEPT) were analyzed. As expected, the findings indicated strong and moderately strong correlations between OEPT scores and speech variables, such as speech rate (r = 0.72), articulation rate (r = 0.61), and mean syllable per run (r = 0.72). They also reported that there were a moderate and weak negative correlation between breakdown variables, including silent pause time (r = -0.52), silent pause ratio (r = -0.59), and the number of silent pauses (r = -0.38). Simply put, examinees' OEPT scores increase when participants talk faster, pause less frequently, and pause for a

shorter duration, and OEPT scores are influenced by both the number and length of pauses. Surprisingly, no significant associations were found between the OEPT scores and any of the filled pause measures.

On the other hand, several other studies (e.g., Kormos & Dénes, 2004) found that fluency ratings related to pause length, not to pause frequency. The researchers calculated three native speakers' and three non-native speakers' composite fluency scores and found correlations with temporal breakdown variables in their study. The correlation results showed that mean length of pauses significantly correlated to the composite scores of native (r = -0.58) and non-native teachers (r = -0.62), whereas the number of silent pauses (for NS r = -0.10, for NNS r = -0.09) and the number of filled pauses (for NS r = -0.08, for NNS r = -0.16) were not related to fluency scores of either raters. However, Cucchiarini et al. (2002) discovered the opposite phenomenon, which indicated that pause frequency was considered more critical than pause length when rating L2 fluency. Their results indicated that the number of silent pauses per minute exhibited statistically significant correlations with fluency rating. However, the mean length of silent pauses seems to have almost no relation with perceived fluency. To summarize, the findings regarding pause frequency and length were inconsistent, and the relative importance of both pause phenomena varied among investigations.

Meanwhile, in order to understand the role and effects of

breakdown phenomena on L2 fluency perception in-depth and associate pause phenomena with speakers' underlying cognitive process of producing speech, several previous pieces of research focused on pause distribution (e.g., Kahng, 2014; Saito et al., 2018; Tavakoli, 2011), comparing breakdown fluency patterns made by native speakers (or fluent L2 speakers) and L2 speakers (or non-fluent L2 speakers). For example, Tavakoli (2011) compared the number and total amount of pauses in different locations (in the middle and at the end of clauses) made by forty native English speakers and forty L2 English learners while narrating picture stories. The quantitative analysis revealed that L2 learners paused more often and for extended periods of time than native speakers. Furthermore, the study discovered that the pausing pattern of L2 learners was distinct in that they often paused in the middle of clauses rather than at the end. Qualitative research was conducted to better understand the mid-clause pause. First, it was found that some L2 learners' mid-clause pauses were followed by repetition and replacement. In other words, L2 learners often paused before repeating a vocabulary item or substituting one word or phrase for another, implying that L2 learners' information processing load was increased. Second, the researcher found that when L2 learners reformulated a language structure, they seemed to pause. These observations indicated that breakdown and repair fluency could interact and have an effect on one another. Lastly, it was seen that the L2 students paused while formulating their thoughts (i.e., planning what to say online). The

study overall found that the critical difference between the pausing patterns of L2 learners and native English speakers was not in the number of pauses, but instead in where these pauses occurred. Additionally, the findings suggested that non-native speakers with lower proficiency levels might pause more frequently in the mid-clause position than speakers with higher proficiency levels, as mid-clause pauses were frequently associated with disruptions in linguistic and underlying cognitive processes.

In a similar vein, Kahng's (2014) research revealed the significance of silent pause locations in determining fluency levels. In her research, forty-six participants, including 31 Korean English learners and 15 English native speakers, completed a spontaneous speech task, and their speech productions were compared in terms of three aspects of fluency to see how L2 speakers' fluency differs from fluent speech output in L1. The results, which focused on pause phenomena, showed that the L1 and L2 groups differed more in the frequency than in the length of pauses, and they were also different in the use of silent pauses compared to filled pauses. The subsequent analyses showed that L1 and L2 speakers had a significant gap in the silent pause rate within clauses, implying that L2 speakers tended to pause in the middle of the clause more often, while L1 speakers paused within syntactic boundaries. Correlations between utterance fluency measures and speaking proficiency scores have revealed that silent pause rate within a clause (r = -.535) was crucial in predicting speakers' overall oral proficiency level. These findings were

compatible with other studies (e.g., Lennon, 1990; Tavakoli, 2011; Towell et al., 1996) and were consistent with the hypothesis that pauses within clauses represented processing difficulties in speech production, which was a typical fluency vulnerable point in L2 speech (e.g., Pawley & Syder, 2000; Wood, 2010).

In conclusion, a broad body of previous literature on L2 breakdown fluency has yielded mixed results, with inconclusive findings (Kormos, 2006), further complicating the pause phenomenon. Recently, pause distribution has received more attention than pause frequency and length as a more predictive measure concerning perceived fluency. However, little is known about whether students in the Korean EFL background exhibit the consistent pictures seen in the studies mentioned above and whether pause distribution can be used as a robust measure to judge fluency. Thus, the current study focuses on intensively examining pause phenomena by studying Korean EFL students' pause distribution along with pause frequency and length.

2.3.1.3 Repair Fluency

While a variety of earlier researches examined the association between repair fluency (i.e., the number of corrections and repetitions per minute) and perceived fluency, findings were frequently conflicting, making it the most controversial aspect of the fluency triad (Tavakoli et al., 2020). Several early studies (e.g., Cucchiarini et

al., 2002; Kormos & Dénes, 2004) indicated that, whereas speed and breakdown measures were strong predictors of fluency, repair fluency was not. On the other hand, in Bosker et al.'s (2013) study, repairs were found to add a small but substantial amount of explanatory power to perceived fluency and were reported to have a complex relationship with perceived fluency. Below is the previous literature on two widely used approaches to measuring repair fluency: self-corrections and repetitions.

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. Similarly, Kahng (2014) discovered that L2 speakers used more self-repetitions than L1 speakers, with the findings revealing a weak negative correlation between self-repetitions and overall speaking scores. Concerning self-correction, however, she argued that individual differences seemed to affect repair behavior, with personality and L2 learning experience playing a role in deciding whether or not to correct their errors based on the results of in-depth, introspective interviews. In the meantime, Tavakoli et al. (2020) highlighted the drawbacks of the frequently used approach for measuring repair fluency. To begin, repair measures often overlap with one another or with other facets of performance. Additionally, it has been claimed that verbatim repetitions do not necessarily reflect repair behaviors but rather breakdowns, as a speaker may employ repetitions to buy time. In

addition, several emerging kinds of research indicate that repair fluency seems to be more strongly associated with an individual's speaking style than L2 proficiency, demonstrating that fluency is stable throughout L1 and L2 production and across L2 proficiency levels (Suzuki et al., 2021). To summarize, prior researches on repair fluency have been inconsistent, and there has been little agreement on the extent to which repair measures accurately capture the fluency of L2 speakers. Thus, the current study intends to examine whether repair measures affect the three types of raters' perceptions of fluency and their best predictive models.

Thus far, the existing literature has extensively worked to demonstrate which subconstructs of L2 speech fluency (speed, breakdown, or repair fluency) determine native speakers' perception of fluency. It has been generally shown that native speakers' fluency judgments can be mainly related to speed and breakdown fluency measures, and to a much lesser degree, linked to repair fluency measures (Saito et al., 2018). However, to the best of my knowledge, only a paucity of studies have compared subjective fluency judgments from various types of raters in L2 contexts, such as native EFL teachers, non-native EFL teachers, and, in particular, peer students, and linked their fluency ratings to utterance fluency features in the EFL context. Thus, the current study aims to fill the knowledge gap regarding how various raters perceive fluency differently and which utterance features best predict each rater group's perceived fluency, explaining disparities of fluency perception among the three kinds of raters in the EFL context.

2.3.2 Utterance Fluency Model

While several previous pieces of literature have investigated how breakdown, speed, and repair fluency measures correlate with listeners' judgments of fluency, a few researchers have attempted to determine the best predictive utterance fluency model (as represented by sets of utterance fluency measures), which explained the most variance in L2 speaking fluency ratings (e.g., Bosker et al., 2013; Cucchiarini et al., 2002; Kahng, 2014; Saito et al., 2018).

Bosker et al. (2013) investigated the relationship between utterance fluency measures and perceived fluency and analyzed rater susceptibility to breakdown fluency: number of silent pauses (NSP), number of filled pauses (NFP), and mean length of pauses (MLP), speed fluency: mean length of syllables (MLS), and repair fluency: number of repetitions (NR) and number of corrections (NC). They created three models using predictors from just one fluency aspect and investigated how much a collection of objectively acoustic measurements could account for the variance in fluency ratings, as calculated by the adjusted R^2. Model (1) contained three measurements of breakdown fluency (NFP * NSP * MLP), while models (2) and (3) were built on speed (MLS) and repair fluency (NC + NR), respectively. The multiple linear regression analysis results showed that model (1) produced an adjusted R^2 of 0.5917, Model (2) an adjusted R^2 of 0.5449, and Model (3) an adjusted R^2 of 0.1583. The findings revealed that the collection of breakdown fluency

measures (number of filled / silent pauses, length of pauses) explained the most significant part of the variance in fluency ratings, while repair fluency measures (number of repetitions / corrections) were thought to have insufficient predictive capacity.

In a similar vein, Saito et al. (2018) conducted stepwise multiple regression to determine the relative weights of five utterance variables (i.e., number of pauses within / between clauses, articulation rate, and frequency of repetitions / self-correction) in perceived fluency scores. The findings revealed that three utterance fluency variables (articulation rate, mid-clause pauses, and final-clause pauses) accounted for 57 percent of the variance in perceived fluency ratings. According to the model, native listeners used speed fluency (articulation rate) as a primary cue for perceived fluency judgments and breakdown fluency (mid- and final-clause pauses) as a secondary cue, which contrasts with Bosker et al.'s (2013) findings, in which breakdown fluency measures explained the majority of variance in fluency ratings.

Similarly, Cucchiarini et al. (2002) performed multiple regression measures to examine whether a combination of temporal variables allows for more accurate fluency predictions. The results indicated that the variable that best describes the variance in spontaneous speech is speech rate (r = 0.57) for the beginner level and mean length of run (r = 0.65) for the intermediate level. The second variable included in the stepwise multiple regression analysis was the number of silent pauses per minute. However, at both levels,

the increase in the explained variance was minimal in both cases (for the beginner level: R rises to 0.63; for the intermediate level: R rises to 0.70).

Meanwhile, focusing solely on pause phenomena, Kahng (2014) examined the relative contributions of pause frequency, duration, and distribution to the perception of fluency, arguing that pause phenomena appear to play a significant role in fluency perception. Forty-six native speaker raters assessed the fluency of 80 speech excerpts (74 L2 samples from 37 Korean speakers and 6 L1 samples from three native English speakers), and a multiple linear regression analysis was used to determine the extent to which each aspect of the pause phenomenon could account for the variance in the raters' fluency ratings. Firstly, in a hierarchical multiple regression analysis, the frequency and duration of pauses were entered first, followed by the distribution of pauses. The result revealed that pause frequency explained 29 percent of the variance in fluency ratings and that when pause length was included, the explanation increased to 42 percent. Finally, the addition of the pause distribution explained an extra 10 percent of the variance. Together, the three silent pause measures explained approximately 52 percent of the variance in fluency assessments. To compare the findings of the hierarchical multiple regression, a stepwise multiple regression analysis was conducted next. According to the stepwise multiple regression analysis results, the pause distribution was initially included in the model and was found to account for almost 45 percent of the

variance in the fluency scores. The pause duration was entered after that, accounting for an extra 4 percent of the variance. However, pause frequency was excluded from the model because it did not significantly explain more variance. She concluded that pause distribution was critical in determining perceived fluency, demonstrating that it had the most vital link with fluency ratings and could account for 45 percent of the variance in fluency judgments.

In summary, several previous studies sought to identify the best utterance fluency model that explained the most significant amount of variance in native speaker judgments of L2 speaking fluency, employing sets of utterance fluency variables. The findings generally suggested that breakdown fluency measures (i.e., pause phenomena) and speed fluency measures (i.e., articulation rate) explained most of the variance in fluency ratings, whereas repair fluency measures (i.e., numbers of repetitions and corrections per minute) were deemed to have insufficient predictive capacity. A summary of the studies reviewed above is shown in Appendix A. The appendix contains information about the participants, L2 levels, raters, tasks, utterance fluency measurements, and main findings.

Despite the amount of research on the best predictive fluency model, little is known about the extent of what utterance fluency features can explain the variance of fluency ratings from various raters in an L2 context. Especially, the extent to which utterance fluency attributes influence native and non-native teachers' perceptions of fluency has been understudied, along with whether

their best prediction model as represented by sets of utterance fluency features is different or not. Thus, the current study attempts to determine which of the utterance fluency models best describe the perceived fluency of native teachers, non-native teachers and peers and analyze any discrepancies that may exist.

2.3.3 Utterance Fluency Features and Fluency Levels

Until now, substantial previous research has been conducted to determine which utterance fluency features influenced raters' perceptions of fluency. However, little research has been conducted about the acoustic correlates of perceived fluency at different fluency levels. Additionally, the majority of the previous studies have relied on particular groups of L2 learners with relatively homogeneous proficiency levels (Saito et al., 2018). It is crucial to ascertain which fluency measures are substantially associated with various listener groups' judgments of fluency levels in L2 classrooms, as these findings have pedagogical implications for fluency development curricula and teacher feedback. For example, if acoustic correlates distinguishing low and intermediate fluency levels are identified, EFL teachers can prioritize specific acoustic aspects while addressing them to low fluency level learners. Several studies (e.g., Cucchiarini et al., 2002; Saito et al., 2018; Tavakoli et al., 2020) have investigated which auditory variables were associated with low-, mid-, and high-level L2 fluency.

Cucchiarini et al. (2002) compared the acoustic characteristics of the speech of two different groups of L2 Dutch learners (30 beginner and 30 intermediate learners). The speech materials were scored for fluency by ten teachers and were analyzed through measures such as speech rate, articulation rate, phonation-time ratio, mean length of run, number and length of pauses, number of dysfluencies. Their findings indicated that perceived ratings were found to be best predicted by speech rate for the beginner learners and by mean length of run for the intermediate learners.

Saito et al. (2018) examined the linguistic characteristics and learner profiles of low-, mid-, and high-level fluency performance. Ten native speaker judges were asked to assess speaking fluency of spontaneous speech from 90 adult Japanese English learners and 10 native speakers. Following that, the participants were classified into four proficiency groups using cluster analysis, and their data set was assessed for the number of pauses within / between clauses, articulation rate, and frequency of repetitions / self-correction. The results indicated that the final-clause pause ratio (i.e., the number of filled and unfilled pauses at the end of the clause divided by the total number of words) distinguished low- and mid-level fluency performance, whereas the mid-clause ratio (i.e., the number of filled and unfilled pauses in the middle of the clause divided by the total number of words) distinguished mid- and high-level fluency performance. Additionally, they found that articulation rate distinguished between high-level and native-like performances.

Similarly, Tavakoli et al. (2020) recently explored which acoustic measures best-represented fluency at each assessed level of proficiency and could consistently distinguish one level from another. Thirty-two speakers completed four tasks (British Council's Aptis Speaking Test) and were categorized into four competency levels. Various utterance fluency features were investigated across the four different proficiency levels. The results indicated that breakdown measures differentiated between the lowest level and the rest, while the speed and composite measures (e.g., mean length of run) consistently distinguished fluency from the lowest to upper-intermediate levels.

In summary, the previous studies generally suggested that the acoustic correlates of perceived fluency vary according to proficiency level, with breakdown fluency being a relatively strong predictor of beginners' L2 fluency and speed fluency being a relatively strong predictor of more advanced learners' L2 fluency (Saito et al., 2018). However, due to a lack of previous research, it remains unknown to what extent different listeners employed breakdown, speed, and repair information when rating different levels of speech fluency. Thus, the current research attempts to determine which measures of utterance fluency are closely associated with different rater groups' decision-making concerning fluency levels, especially in the EFL setting.

CHAPTER 3.
STUDY 1: PERCEIVED FLUENCY

This chapter presents a series of experiments and analyses to investigate perceived fluency (overall impression) in the EFL context by examining how oral fluency is perceived by three groups of judges (native teachers, non-native teachers, and peer students). To begin, Section 3.1 details the experiment's methodology. Section 3.2 reports the findings. Finally, Section 3.3 concludes the chapter by discussing the results.

3.1 Methodology

The study examines the differences in perceived fluency by three groups of judges under the mixed-method framework. First, a set of ANOVAs were conducted in order to investigate group differences quantitatively in terms of overall fluency ratings across task types. Additionally, by examining listeners' written comments on their overall perceptions of fluency, linguistic elements affecting the three groups of judges' fluency perceptions were compared and scrutinized qualitatively. Given that many previous studies were only conducted quantitatively, and their findings frequently contradicted one another due to methodological differences, it is critical to use both

quantitative and qualitative methods to obtain a comprehensive picture of L2 fluency. The following section describes the procedure in detail.

3.1.1 Participants

In the current study, one hundred twenty-four individuals participated. 30 Korean high school students were recruited to perform speaking tasks, and 94 listeners, including 26 native English teachers, 29 non-native English teachers, and 39 peer students, participated in the research.

The volunteer speakers were all from the same high school in Korea, and they were all in 11th grade (during the year of 2021), ranging in age from 17 to 18 (*M* years = 17.9, *SD* = 0.3). The high school where the researcher obtained the speech samples was for gifted students in math and science. Most of them were predominantly oriented to math and science, highly motivated, and intellectually mature. Speakers reported that they had studied English for an average of 11 years (*SD* = 1.8), with the majority beginning in pre-school. None of them had any other language spoken at home as a child other than Korean. At the time the research was conducted, all of the speakers had completed the same courses: English Conversation I & II. These courses were mandatory for 10th grade students and designed to improve students' basic communicative competencies. Their syllabi were mainly comprised of listening and speaking activities focusing on lowering students'

inhibition toward speaking English and encouraging them to use English in various contexts. Students were expected to have taken English Conversation I in the first semester and English Conversation II in the second semester of 10th grade. Since speech samples were taken from students with a highly homogeneous background and the majority of speakers were male (29 men, 1 woman), the researcher was able to minimize several methodological variables that have been known to influence fluency scores, such as age, gender, and scholarly environment.

Then, speakers were classified into three categories based on their speaking proficiency. The speakers' one-year English speaking performance test scores from English Conversation I and II were used to divide them into three groups (high, mid, and low proficiency groups). All speaking performance test scores, ranging from interviews and individual presentations to discussion tasks, were added, but written test results were intentionally removed because they were deemed to tap into different linguistic abilities other than speaking ability. Table 3.1 shows the grouping information of the speaker participants.

Table 3.1 Speakers' Grouping Information

Group	*N*	*Min. Score*	*Max. Score*	*Mean Score*	*SD*
Low	10	273.60	324.10	303.90	16.85
Mid	10	324.95	337.65	331.53	4.38
High	10	338.65	348.55	342.93	3.06

Note. a total score: 350

Additionally, a total of ninety-four listeners participated in the experiment as raters. Twenty-six native English speaker teachers (NS), 29 non-native Korean English teachers (NNS), and 39 peer students (Peer) were recruited as listeners who rated the L2 speech fluency of the participants.

The native English teachers, comprising of 21 women and 5 men (*M* age = 28.3, *SD* = 6.1, *Min* = 22, *Max* = 43), had varying degrees of teaching experience, ranging from 0.5 to 12 years (*M* length of teaching = 4.1, *SD* = 3.6). The researcher purposefully recruited native English teachers with experience teaching English as a foreign language in Korean middle and high schools, as they were frequently involved in judging the speech of L2 speakers in educational and testing contexts (e.g., speaking performance tests). Moreover, as they had various experiences in judging the oral fluency of L2 speech in Korean English classes, they were more likely to provide a consistent and accurate assessment of fluency (Préfontaine et al., 2016). The entirety of the 29 non-native Korean English teachers (25 women and 4 men) were working in middle and high schools in Korea when the research was conducted. The mean age was 37.7 years (*SD* = 7.6), with a range of 26 to 56 years. They had a range of teaching experiences spanning 1 to 30 years (*M* length of teaching = 11.6, *SD* = 6.8). The 39 peer-group students (36 men, 3 women) were enlisted as a final listening group. All peer students were in the 10th grade (*M* age = 16.6, *SD* = 0.6) and were recruited

from the same high school. They were enrolled in the aforementioned 10th grade English Conversation I course at the time of recruiting. They reported in the demographic questionnaire that they began learning English around the age of 7 (*SD* = 1.9). The researcher guaranteed that the students serving as raters and those being rated had never met and had no prior knowledge of one another in order to avoid familiarity effects. Below, Table 3.2 is the summary of the listeners' information.

Table 3.2 Listeners' Information

Group	*N (gender)*	*Min. Age*	*Max. Age*	*Mean Age*	*SD*
NS	26 (5m, 21f)	22	43	28.3	6.1
NNS	29 (4m, 25f)	26	56	37.7	7.6
Peer	39 (36m, 3f)	16	17	16.6	0.6

Note. NS: Native English teachers, NNS: Non-native Korean English teachers

3.1.2 Instruments

Two speaking tasks that required different linguistic and cognitive efforts were employed in the study. As the first speaking task, one picture narrative task, in which speakers narrate a story according to the sequence of events depicted in a provided picture (Préfontaine et al. 2016), was chosen in line with previous researches (e.g., Derwing et al., 2004; Lennon, 1990; Rossiter, 2009). The present study made use of an eight-frame picture (Derwing et al., 2009) depicting two strangers who bump into each other on a street corner and

accidentally swap identical suitcases (see Appendix B). This task involved encoding new, visual information into linguistic form and required some degree of imagination, which required complex language and great cognitive effort (Foster & Skehan, 1996).

Spontaneous speech samples were obtained as a second speaking task. The second task was a single question designed to elicit spontaneous speech from participants, in which they talked openly about a given subject (Kahng, 2014). The question was about their future careers, which participants were familiar with and could discuss openly (see Appendix C). As it involved accessing information well known to the speaker, it was seen as requiring less cognitive effort and relatively simple linguistic forms to be used (Foster & Skehan, 1996).

Then, each L2 speaker's speech response was recorded individually in a quiet room with SONY PCM-A10 and saved as 44KHz (32-bit resolution). Later recorded responses were edited using the Audacity software (Audacity Team, 2021), as suggested by Munro and Derwing (2020). Following previous study methods (Derwing et al., 2004; Rossiter, 2009), 30 seconds of each narrative were included in the samples, with initial dysfluencies, false starts, and hesitations removed. Each fragment started and ended at a phrase boundary (Kahng, 2014).

3.1.3 Procedures

The experiment started with the collection of L2 speech samples. The thirty high school students were initially asked to complete a demographic questionnaire (see Appendix D), after which they were given spoken instructions on the two tasks and the recording and were directed to complete the two speaking tasks (i.e., one picture narrative and one spontaneous speech task). Before recording the picture narration task, the participating students were given one minute to examine the eight pictures provided to them and plan their narration. After completing the first task (picture narration), a question sheet for the second task (spontaneous speech) was provided. The participants spent 20 seconds reading the question before answering it.

Then, fluency ratings, the dependent variable of the current study, were collected from the three groups of judges (26 NS, 29 NNS, and 39 Peer: a total of 94 listeners). The evaluation process for NS and NNS teachers was completed online via a Google survey form, while peer group listeners listened to and rated samples in a quiet setting with the researcher present. Prior to initiating the fluency judgments, the researcher briefly instructed listeners on what they should focus on (i.e., how easily and smoothly speech is delivered).

In terms of the instructions given to raters, the majority of studies have explicitly encouraged raters to look for temporal aspects of fluency in the excerpts. For example, in the previous researches,

such as Derwing et al. (2008), Kahng (2014), and Rossiter (2009), the researchers provided raters with a list of temporal variables (e.g., speech rate, silent and filled pauses, self-corrections, self-repetition, as well as the overall flow of speech), associated with speaking fluency, and trained listeners with a few sample excerpts to ensure that raters did not confuse fluency with proficiency. On the other hand, some other studies (e.g., Saito et al., 2018) instructed raters to focus on global fluency in the absence of specific L2 speech sub-constructs (i.e., utterance fluency features). The current dissertation struck a balance between two versions of instructions by providing raters with a brief definition of perceived global fluency and areas of focus (e.g., the flow of the language – does the speaker have problems finding words, hesitating and pausing often, or do the words come quickly?) but not exhaustive lists of fluency's temporal aspects. Thus, the researcher guaranteed that each listener focused on fluency rather than overall proficiency and was capable of exploring the three groups of raters' perceived fluency without limiting their basis of fluency judgments too much (see Appendix E, F for the instructions).

After instructions, listeners received two speech examples and the pictures used in the speaking task to avoid familiarity bias. As Rossiter (2009) noted, it was crucial to avoid listeners evaluating the first few samples differently due to their unfamiliarity with the material. In the following section, listeners were asked to listen to the audio file to the end and rate the fluency score of 60 speech samples (30 picture narration and 30 spontaneous speech samples) using a

nine-point Likert scale, with one indicating very dysfluent and nine marking very fluent. The order of the speech samples delivered in the experiment was randomized. After rating the fluency of each sample, listeners were instructed to write comments about their overall impressions of the fluency of each speech sample in the area provided below the scale. The entire rating session took about one and a half hours.

3.1.4 Data Analysis

As the purpose of Study 1 is to elucidate the underlying mechanisms of perceived fluency, a series of analyses of how native teachers, non-native teachers and peer students perceived oral fluency were conducted. The study employed a mixed-method approach as a research framework, which combined quantitative and qualitative research methods enabling the researcher to gain a more profound knowledge of the raters' rating practices. Following Kim's (2009) research, the study adopted an expansion design (Green et al., 1989) within a mixed approach. The expansion design offered a comprehensive and diverse illustration of rating behaviors by examining both product (i.e., fluency scores assigned to L2 samples) and process (i.e., evaluative comments) by which the three types of raters assessed students' oral fluency.

As a first step, a one-way ANOVA was conducted to reveal how the three groups of judges rated oral fluency. Then, a series of

two-way ANOVAs were conducted to trace the group differences across the two different speaking tasks (picture narration task, spontaneous speech task) at different students' oral proficiency levels (low, mid, high). All statistical analyses were carried out using SPSS version 21, and the level of significance was set at $p < .05$. The findings were expected to quantitatively explain how the two EFL teacher groups' and the peer group's fluency perceptions varied according to different tasks at different oral proficiency levels.

Next, a qualitative study was conducted on the overall fluency impressions reported by the three listener groups to determine which evaluation criteria influenced their fluency ratings and how they differed. Following Rossiter's (2009) methodology, the fluency impressions of each group were first classed as positive, negative, and neutral. Then, the three groups' frequency of comments and frequency distribution of evaluation criteria were examined. Following that, the written comments were analyzed by the evaluation criteria. As Kim (2009) suggested, comments that contained only evaluative adjectives but no evaluation substance (e.g., fluent, clear, rapid, mumble, and so on) were excluded from the analysis in order to avoid misjudging the evaluative intent. Employing typology development and data transformation (Caracelli & Greene, 1993), a total of 4,456 written comments from the three rater groups were open-coded so that the fluency criteria that the three listener groups drew upon emerged. In the following part, the findings from both the quantitative and qualitative approaches were integrated and interpreted comprehensively.

3.2 Results

To address the first research question, 'How does perceived fluency of native teachers, non-native teachers and peers differ?' quantitative and qualitative data were collected and analyzed. The quantitative data set included 5,640 valid ratings given to 60 speech samples on two tasks by 26 native teachers, 29 non-native teachers, and 39 peer students. The qualitative data included 4,456 written comments. The following part begins the quantitative study by comparing the three groups of raters.

3.2.1 A Quantitative Study

3.2.1.1 Comparison of the Three Rater Groups

To begin, the consistency of the ratings for each group was determined using the Cronbach's alpha coefficient, which indicated the degree of agreement within each group. Table 3.3 summarizes the findings. Each group had a high degree of reliability in the range of .988 to .989 for task 1 and .978 to .986 for task 2. The overall consistency of all evaluations was between .988 and .991. Before conducting statistical analyses, the Shapiro-Wilk normality test (see Table 3.4) was used to ensure whether the rating data for the three

groups were normally distributed (*Sig.* > 0.05).

Table 3.3 Interrater Reliability by Task Type

	NS	NNS	Peer
Task 1	.988	.989	.988
Task 2	.978	.986	.982
Total	.988	.991	.989

Table 3.4 Test of Normality

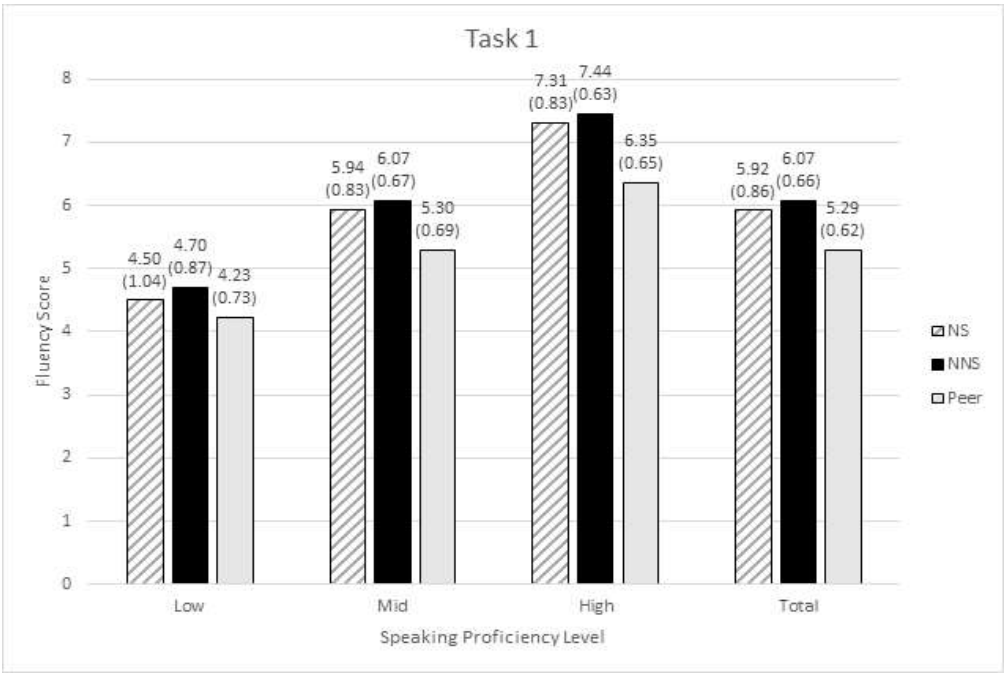
	Kolmogorov-Smirnova			Shapiro-Wilk		
	Statistic	*df*	*Sig.*	*Statistic*	*df*	*Sig.*
NS ratings	.106	30	.200	.952	30	.190
NNS ratings	.139	30	.147	.954	30	.212
Peer ratings	.096	30	.200	.969	30	.516

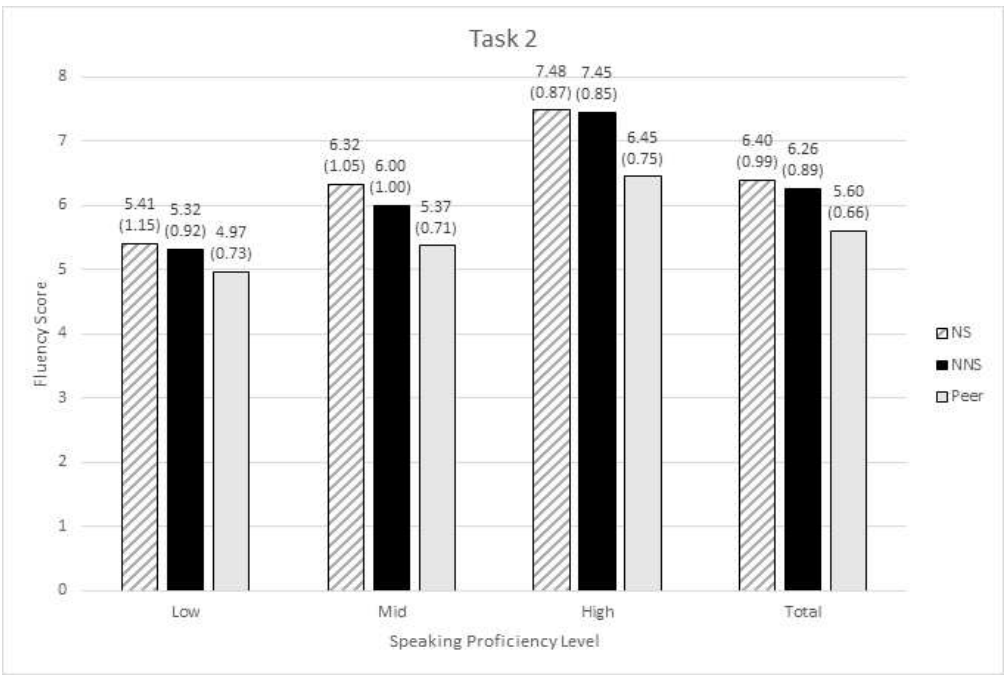
Following that, to examine how native teachers', non-native teachers', and peers' judgments of fluency differ among groups, the ratings of twenty-six native EFL teachers were statistically compared to those of 29 non-native EFL teachers and 39 peer students. Figure 3.1 presents descriptive statistics that summarize each rater group's means and standard deviations according to the speakers' oral proficiency levels and task types. The peer group rated the two tasks more harshly than the native and non-native teacher groups at all levels, but native and non-native teachers had similar severity patterns across tasks. For example, the mean scores for the first task for native and non-native teacher groups were 4.50 and 4.70 at the

low level, whereas the mean score for the peer group was 4.23. Similarly, at the mid-level, native teachers gave mean scores of 5.94, non-native teachers 6.07, but peers gave a mean score of 5.30. At the high level, the same rating pattern was found. Native teachers scored 7.31, non-native teachers scored 7.44, but peers scored 6.35, lower than both teacher groups. For the second task, the same rating patterns were also observed. It is worth noting that when task 1 and task 2 were combined, the mean scores for native and non-native teachers were nearly the same (both M = 6.16), whereas the peer group awarded a lower score (M = 5.44). Additionally, all three groups gave task 2 a higher rating than task 1.

Meanwhile, it was observed that the standard deviations of the peer ratings were consistently lower than those of native and non-native teachers, except for two cases (mid and high level for task 1). In other words, the peer students in the current study generally scored their peers within a narrower range than teachers. This observation has been made in other studies on peer assessment (Cheng & Warren, 2005; Freeman, 1995), and according to Cheng and Warren (2005), it is usually attributed to students' reluctance to grade their peers up and down.

Figure 3.1 Descriptive Statistics of the Three Groups' Fluency Ratings by Oral Proficiency Levels and Task Types





Note. The mean and standard deviation are displayed on the bar graph, with the standard deviation in parenthesis.

Overall, examining the three groups' mean scores revealed that the native and non-native teacher groups' scoring patterns were very similar, but the peer group consistently rated lower on both tasks than native and non-native teacher groups across the three levels. In Table 3.5, the statistical results of a one-way ANOVA for group differences are presented. The results indicated that there were significant mean differences with a large effect size[2] ($\eta_p^2 > .14$) among the three groups on task 1 ($F = 11.836$, $p = .000$, $\eta_p^2 = .206$), task 2 ($F = 8.991$, $p = .000$, $\eta_p^2 = .165$), and the tasks combined ($F = 10.977$, $p = .000$, $\eta_p^2 = .194$), which suggested that raters' linguistic and professional background influenced the fluency ratings across the two tasks.

Table 3.5 One-Way ANOVA and Post-hoc Results of the Group Difference

Task	Group	*N*	*M*	*SD*	*F*	*p*	η_p^2	Scheffe
Task 1	NS	26	5.92	0.86	11.836	.000**	.206	Peer < NS, NNS
	NNS	29	6.07	0.66				
	Peer	39	5.29	0.62				
Task 2	NS	26	6.40	0.99	8.991	.000**	.165	Peer < NS, NNS
	NNS	29	6.26	0.89				
	Peer	39	5.60	0.66				
Task 1 & 2	NS	26	6.16	0.89	10.977	.000**	.194	Peer < NS, NNS
	NNS	29	6.16	0.75				
	Peer	39	5.44	0.60				

Note. * = $p < .05$; ** = $p < .01$

2) Cohen (1988) has provided benchmarks for interpreting effect size: small effect size, .06: medium effect size, .14 or higher: large effect size. Nevertheless, the benchmark must be interpreted with caution in L2 research as it is arbitrary (Cohen, 1988) and never intended as a prescription but rather as a general guide (Plonsky & Oswald, 2014).

Table 3.5 also shows the post-hoc comparison results. As the homogeneity of variances assumption was met, the Scheffe test was used. The results indicated that the difference between the native teacher group and the peer group and the difference between the non-native teacher group and the peer group contributed to the group difference on tasks 1 and 2. The native and non-native teachers groups did not vary substantially on either task. These findings statistically indicated that the EFL teacher groups, which included both native and non-native teachers, had similar patterns across tasks.

3.2.1.2 Effects of Raters and Task Types on Fluency Ratings

The preceding experiment established the discrepancy in severity patterns among the three groups of judges. Following that, the question of how the three groups of judges differ in their judgments of fluency across two different task types at different speakers' oral proficiency levels must be addressed to determine the effect of rater and task variability on fluency ratings. Thus, follow-up two-way ANOVA measures were conducted to verify the main effect of the rater group and task types on fluency ratings and the interaction effects between rater groups and tasks at the three different proficiency levels (low, mid, high proficiency levels).

Table 3.6 summarizes the results of the two-way ANOVA performed on the participants with low proficiency. The findings indicated that the main effects of the rater group and task on fluency

ratings were substantial, demonstrating that significant disparities in fluency ratings existed between rater groups and tasks at the low competence level (F = 4.244, p = .016, η_p^2 = .045 for Group; F = 32.869, p = .000, η_p^2 = .153 for task). However, the comparison of the effect size indicated that a proportion of variance accounted for task types was greater than that of rater groups. Lastly, it was found that the interaction effects between group and task were not statistically significant (F = .365, p = .694, η_p^2 = .004).

Table 3.6 Two-Way ANOVA Results of the Group Difference at Low Proficiency Level

Source	Type III Sum of Squares	*df*	*F*	*Sig.*	η_p^2
Rater Group	6.812	2	4.244	.016*	.045
Task	26.378	1	32.869	.000**	.153
Rater Group * Task	.586	2	.365	.694	.004
Error	146.055	182			

Table 3.7 presents the post-hoc comparison results of the group difference by raters. At the low proficiency level, the difference between the non-native teacher group and the peer group was significant (p = .027), and the difference between the native teacher group and the peer group approached significance (p = .084), but the difference between the two teacher groups was not significant (p = 1.000). The findings confirmed that when the three groups of raters

judged the fluency of low proficiency level students, native and non-native teachers rated with similar severity, but peer students rated speakers far harsher than non-native teachers.

Table 3.7 Post-hoc Comparison Results of the Group Difference by Raters at Low Proficiency Level

(I) Group	(J) Group	Mean Difference (I-J)	*S.E.*	*Sig.*[b]	95% CI	
					Lower Bound	Upper Bound
Native	Non-native	-.055	.171	1.000	-.468	.359
	Peer	.355	.160	.084	-.032	.743
Non-native	Peer	.410	.155	.027*	.035	.785

b. Adjustment for multiple comparisons: Bonferroni.

The post-hoc comparison results of the group difference by task types is shown in Table 3.8. The main effect on task types was statistically confirmed at the low proficiency level, demonstrating that the low-level L2 speakers would receive considerably higher fluency scores from the three groups of raters when performing task 2 (spontaneous speech) compared to task 1 (picture narration task).

Table 3.8 Post-hoc Comparison Results of the Group Difference by Task Types at Low Proficiency Level

(I) Group	(J) Group	Mean Difference (I-J)	*S.E.*	*Sig.*[b]	95% CI	
					Lower Bound	Upper Bound
Task 1	Task 2	-.760	.133	.000**	-1.022	-.499

b. Adjustment for multiple comparisons: Bonferroni.

The two findings from the prior analyses are also depicted in Figure 3.2. First, at the low proficiency level, the peer group assigned significantly lower ratings than the EFL teacher groups on both tasks, and the fluency ratings between the two EFL teacher groups did not differ significantly. Second, task 2 received higher rating scores than task 1 from all the three group raters. While the gap between tasks 1 and 2 appeared to be the narrowest at non-native teachers' ratings, the task-group interaction effect was not verified.

Figure 3.2 Interaction Effects of Rater Group and Task on Fluency Ratings at Low Proficiency Level

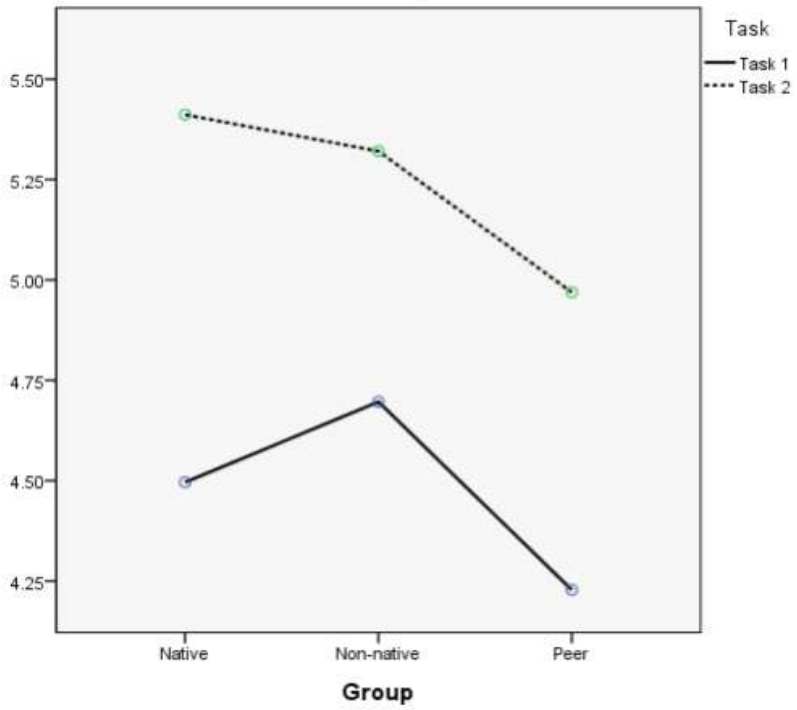


Table 3.9 demonstrates the result of the two-way ANOVA done on the mid-level L2 speakers. The findings indicated that the

main effect of the rater group on fluency ratings was substantial with a large effect size, demonstrating that significant differences in fluency ratings existed between rater groups at the mid proficiency level (F = 19.026, p = .000, η_p^2 = .173). However, contrary to the low-level case, the main effect of the task was not significant (F = 1.091, p = .298, η_p^2 = .006), indicating that the fluency judgments of the three groups on mid proficiency level speakers did not differ between tasks 1 and 2. As with the low-level statistical conclusion, the rater group and task interaction effects were not statistically significant (F = 1.043, p = .354, η_p^2 = .011).

Table 3.9 Two-Way ANOVA Results of the Group Difference at Mid Proficiency Level

Source	Type III Sum of Squares	*df*	*F*	*Sig.*	η_p^2
Rater Group	25.627	2	19.026	.000**	.173
Task	.735	1	1.091	.298	.006
Rater Group * Task	1.405	2	1.043	.354	.011
Error	122.575	182			

The results of the post-hoc comparison of the group difference are presented in Table 3.10. While the difference between the two EFL teacher groups and the peer group was significant at the mid proficiency level (p = .000 for NS, p =.000 for NNS), the difference between the two EFL teacher groups was not significant

(p = 1.000). The data indicated that when the three groups of raters assessed a student's fluency at mid proficiency level, native and non-native teachers rated similarly, but peer students rated speakers significantly more harshly than EFL teachers.

Table 3.10 Post-hoc Comparison Results of the Group Difference by Raters at Mid Proficiency Level

(I) Group	(J) Group	Mean Difference (I-J)	*S.E.*	*Sig.*[b]	95% CI	
					Lower Bound	Upper Bound
Native	Non-native	.087	.157	1.000	-.291	.466
	Peer	.792	.147	.000**	.437	1.147
Non-native	Peer	.705	.142	.000**	.361	1.049

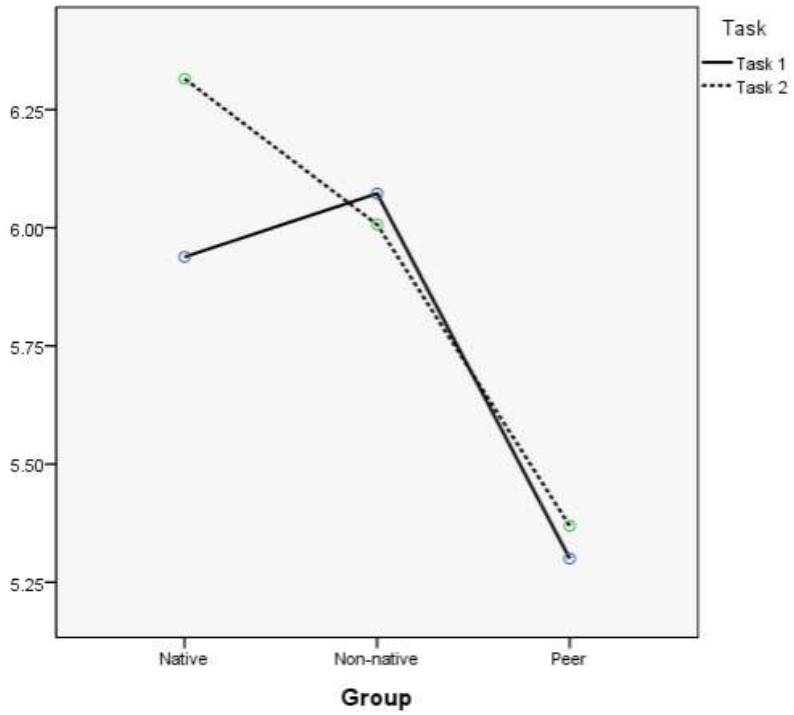
b. Adjustment for multiple comparisons: Bonferroni.

The results of the post-hoc comparison of the task difference revealed that the main effect on task type was not verified at the mid proficiency level (p = 0.298), indicating that the ratings given to task 1 and 2 by the three rater groups were not significantly different. In other words, the results suggested that L2 students at the mid proficiency level were graded similarly on both tasks.

In sum, as described in Figure 3.3, the peer group assigned lower fluency rating scores to mid-level L2 speakers than the two EFL teacher groups on both tasks, and the fluency rating scores between the two EFL teacher groups were not statistically different. In contrast to the low-level results, there was no task effect, indicating that the three rater groups did not differ in their fluency

assessment between the two task types. It is worth mentioning that native teachers rated much higher on task 2 than non-native teachers, while non-native teachers scored marginally higher on task 1. However, no interaction effect of task and group was seen.

Figure 3.3 Interaction Effects of Rater Group and Task on Fluency Ratings at Mid Proficiency Level



Finally, a two-way ANOVA was conducted to examine L2 speakers with a high level of proficiency (see Table 3.11). As with the low- and mid-level results, the findings indicated that the main effect of the rater group on fluency ratings was significant with a large effect size, showing that significant differences in fluency ratings existed between rater groups at the high proficiency level (F

= 41.658, p = .000, η_p^2 = .314). Additionally, contrary to the low-level analysis result, the task's main effect was not significant (F = .679, p = .411, η_p^2 = .004), indicating that the fluency judgments of the three groups on speakers with a high proficiency level did not differ between tasks 1 and 2. Group and task interaction effects were not statistically significant (F = .155, p = .856, η_p^2 = .002) as with the statistical results for the low- and mid-levels.

Table 3.11 Two-Way ANOVA Results of the Group Difference at High Proficiency Level

Source	Type III Sum of Squares	*df*	*F*	*Sig.*	η_p^2
Rater Group	47.916	2	41.658	.000**	.314
Task	.390	1	.679	.411	.004
Rater Group * Task	.178	2	.155	.856	.002
Error	104.671	182			

Table 3.12 demonstrates the results of the post-hoc comparison of the group difference by raters at a high L2 proficiency level. Similar to the findings at the low and mid-levels, no significant differences between the two EFL teacher groups were found (p = 1.000), but discrepancies between teacher groups and the peer group were found to be significant at high proficiency level (p = .000 for NS, p = .000 for NNS). The findings consistently showed that the peer group assigned significantly lower scores to high-level students

than both teacher groups, while the native and non-native teacher groups assigned similar scores.

Table 3.12 Post-hoc Comparison Results of the Group Difference by Raters at High Proficiency Level

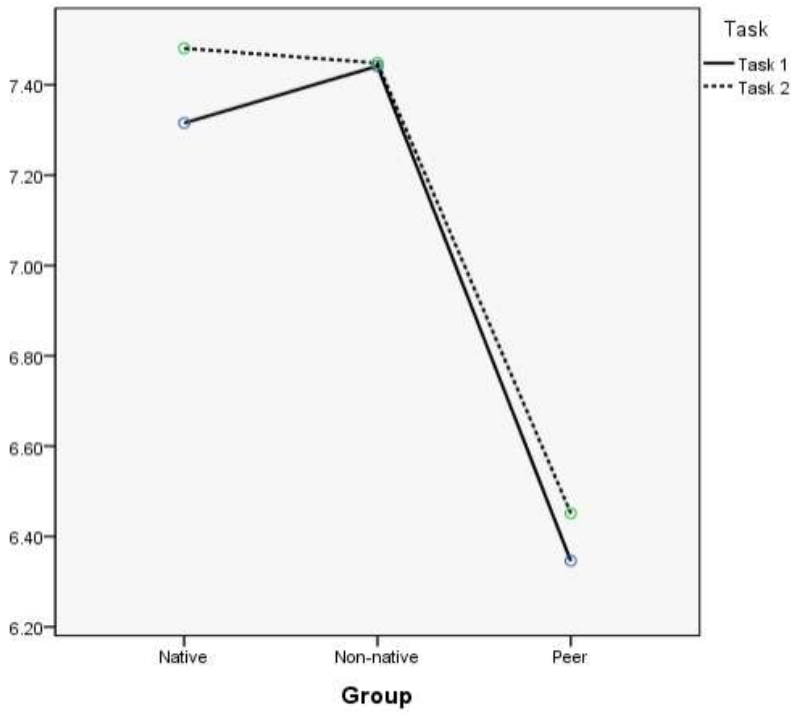
(I) Group	(J) Group	Mean Difference (I-J)	*S.E.*	*Sig.*[b]	95% CI	
					Lower Bound	Upper Bound
Native	Non-native	-.047	.145	1.000	-.397	.303
	Peer	.999	.136	.000**	.671	1.327
Non-native	Peer	1.046	.131	.000**	.728	1.364

b. Adjustment for multiple comparisons: Bonferroni.

The results examining the group differences by task type indicated that the main effect on task type was not significant at a high proficiency level (p = 0.411), similar to the case at mid-level. The data demonstrated that high-level speakers received equally high grades from the three rater groups regardless of task types.

In summary, as illustrated in Figure 3.3, the peer group consistently awarded lower fluency ratings to high-level L2 speakers than the two EFL teacher groups on both tasks, and the difference in fluency ratings between both EFL teacher groups was not statistically significant. As with the result at the mid-level, the task effect was not significant, indicating that ratings on task 1 and task 2 awarded by the three rater groups did not differ for high-level proficiency speakers. Additionally, the interaction effect of task and group was not significantly meaningful.

Figure 3.4 Interaction Effects of Rater Group and Task on Fluency Ratings at High Proficiency Level



3.2.2 A Qualitative Study

Following the statistical analyses, the three groups' written comments were qualitatively analyzed to illustrate their rating patterns. While the quantitative approach to their evaluations provided initial insight into the raters' evaluation patterns, a qualitative method was expected to provide a more comprehensive and enriched understanding of the three rater groups' rating behaviors (Kim, 2009).

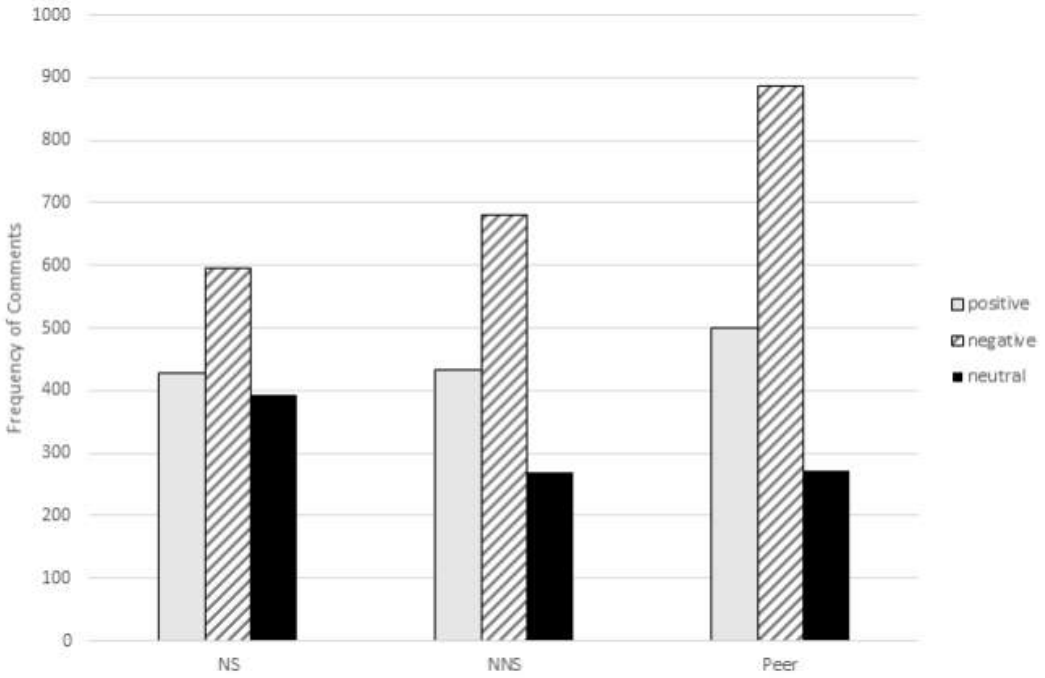
First, the number of comments received by the three groups was compared, which was found to vary significantly. The native

teacher group made 1,417 comments (54.5 comments per person), and the non-native teacher group made 1,380 comments (47.6 comments per person), peer group members contributed only 1,659 comments (approximately 42.5 comments per participant). This could be because peer students were not used to writing detailed evaluations on their peers' speaking performance. Several peer students remarked afterward that it was their first time listening to many of their peers' speech samples and grading their fluency with written comments, or that it was not a frequent occurrence to them. It is also worth noting that native teachers made slightly more comments than non-native teachers. This could be explained by the notion that, as Kim (2009) suggested, providing students with detailed evaluative comments might not be as widely used in an EFL context as the traditional fixed response assessment that has been used.

Second, the written comments made by the three groups of listeners were analyzed in terms of mood (good, negative, and neutral) to see whether their mood affected their fluency judgments (see Figure 3.5). The proportion of positive comments was consistent in all three rater groups (30.2% for the native teachers, 30.2% for the non-native teachers, and 31.3% for the peers), indicating that approximately 30 percent of highly fluent L2 speakers were consistently judged positively across all the listener groups. The discrepancies, however, occurred in the proportion of negative comments. Whereas the proportion of negative responses of the native and non-native teacher groups was 42.0 percent and 49.3 percent,

respectively, the peer group produced a relatively large portion of negative remarks at 53 percent. The findings indicated that peer students might be more critical of their peers' fluency than their teacher groups, and these negative moods possibly explained peer students' tendency to award more harsh fluency ratings.

Figure 3.5 Mood Distribution of the Comments by Native Teachers, Non-Native Teachers and Peers

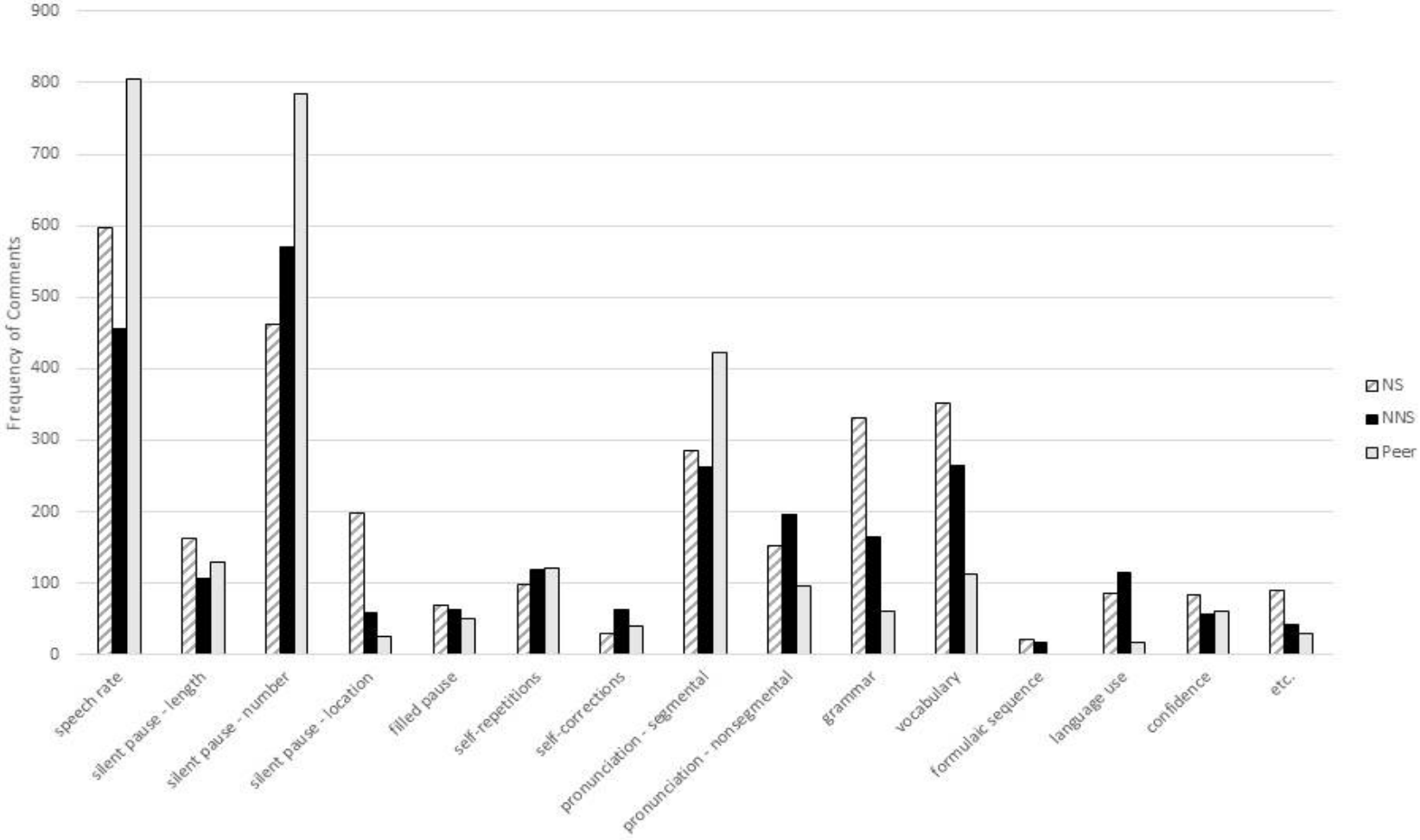


Following that, the comments were examined using fluency criteria. While classifying comments provided by the three groups, several common themes emerged that constitute fluency perception. There were fifteen fluency criteria identified: seven temporal (speech rate, the length, number, and distribution of silent pauses, filled

pauses, self-repetition, self-correction) and seven non-temporal variables (pronunciation, grammar, vocabulary, formulaic sequence, language use, confidence). As guided by Rossiter (2009), pronunciation was divided into segmental (e.g., pronunciation problems such as l/r, f/p) and non-segmental (e.g., monotone, no linking, bad rhythm, natural intonation) groups. The comments that did not fit into either category were classified as 'etc.'. The frequency of comments provided by the three groups on the 15 evaluation criteria is presented in Figure 3.6.

As shown in Figure 3.6, the native teacher group emphasized speech rate (19.83% of all remarks), the number of silent pauses (15.29%), vocabulary (11.64%), grammar (10.98%), and segmental pronunciation (9.45%). The group of non-native teachers placed more importance on the number of silent pauses (22.39%) than speech rate (17.84%), followed by vocabulary (10.39%), segmental pronunciation (10.31%), and non-segmental pronunciation (7.69%). Surprisingly, four out of the top-five fluency criteria were matched between native and non-native teachers (i.e., speech rate, the number of silent pauses, vocabulary, and segmental pronunciation) when they evaluated L2 speakers' fluency. These trends suggest that the two teacher groups had similar views on how students' speaking fluency should be perceived and evaluated. Similarly, the peer group placed the most important emphasis on speech rate (29.27%), followed by the frequency of silent pauses (28.47%), segmental pronunciation (15.32%), the length of the silent pauses (4.72%), and self-repetition (4.36%).

Figure 3.6 Frequency Distribution of the Comments by Native Teachers, Non-Native teachers and Peers



It is worth noting that, in contrast to the EFL teacher groups, the peer group placed the most emphasis on the top three criteria (speech rate, the number of silent pauses, and segmental pronunciation), which accounted for more than 73 percent of the total. It appeared that the peer group concentrated mainly on speech rate, silent pauses, and pronunciation of L2 speech samples, with little regard for other factors. It could be explained that the peer group was not an expert and could not identify several linguistic factors affecting fluency compared to expert groups such as EFL teacher groups. However, it is interesting to observe that all the three rater groups placed a high value on silent pause features, including the number, length, and location. The total silent pause criteria (number, length, and location in total) accounted for nearly a third of written comments by the three groups. Overall, a comparable consistency was discovered among the raters' written comments. All three rater groups regarded silent pause and speed and pronunciation of the L2 speech samples as critical components in determining fluency.

While examining the distribution of comments, it was further discovered that the expert groups (native and non-native teacher groups) took care of all the categories, such as non-segmental pronunciation (e.g., intonation, rhythm, stress), vocabulary, and formulaic sequence, possibly due to their superior linguistic training and more extensive explicit knowledge of these phenomena (Rossiter, 2009). For instance, the peer group made no mention of formulaic or multiword sequences, even though multiword sequences are tightly

correlated to oral fluency (Tavakoli & Uchihara, 2020).

In addition, disparities between the native teacher group and the other two non-native groups were also found. The native teacher group seemed to emphasize the distribution of silent pauses and grammar, making significantly more comments than the non-native teacher group or the peer group. For example, it was revealed that the native teacher group made twice as many comments on grammar than the non-native teacher group and five times more than the peer group (NS: 331, NNS: 165, Peer: 60). The relatively small amount of comments on grammar by non-native teachers raised the possibility that they were not as attentive to specific features of linguistic errors as native teachers were, as long as the speech was understandable. It could be possible that the non-native teachers considered linguistic errors a trade-off for students' speaking fluency, given that rapid speech output might contain more grammatical mistakes due to limited processing time. Another possibility might be that those non-native teachers were more familiar with Korean students' common grammatical mistakes (e.g., tense, article) than native teachers because they shared the same first language background and were more tolerant of mistakes students made.

The responses of the three groups to the distribution of silent pause followed the same pattern. It was discovered that the native teacher group provided over three times the comments on the distribution of silent pauses compared to the non-native teacher group and seven times more than the peer group (NS: 197, NNS: 59, Peer:

26). This disparity could be attributed to the different sensitivity of the distribution of silent pauses among the three rater groups. The native teachers' large quantity of comments may reflect that they were more sensitive or stringent about silent pause location than non-native teachers and peers. This could also mean that native teachers were less tolerant of or more readily distracted by unusual silent pause distributions, such as extended silent pauses within clauses, than non-native teachers and peers.

Lastly, when the written comments were examined, it was discovered that native teachers generally provided more detailed and elaborated comments than the non-native teachers and the peers, which was consistent with findings of previous researches (e.g., Gui, 2012; Kim, 2009; Shi, 2001). For instance, when the distribution of silent pauses was examined, it was discovered that not only did the native teacher group provide significantly more comments than the other groups, but they also provided more detailed feedback, at times highlighting specific locations and possible explanations for problematic areas on both positive and negative comments.

1. '… many unnatural … pauses were in the middle of sentences, showing that, more often than not, the speaker had trouble coming up with the right words to use next for parts of the story …'
(Native teacher #21)

2. '… unlike some of the previous subjects that seemed choppy by words this seems choppy by sentence level. The individual breaks between words are not uniform but are frequent noting the subjects struggle and or hesitation in production …'

(Native teacher #24)

3. '… there were a few moments they paused to come up with the right words / sentences to say. But I appreciated the detail included, and most individual clauses were spoken in one go (without pauses) …'

(Native teacher #21)

4. '… they only paused in between clauses. What I mean by that is, the speaker was not getting hung up on the right words to use, but the direction of the story/which sentences to use next …'

(Native teacher #13)

The native teacher group occasionally selected a word or phrase from L2 speech samples to illustrate a problem area and support their judging comments.

5. 'I'd say this speaker's fluency is mediocre; I got the main idea, but their flow was choppy, especially in between words (e.g. "I want to... teach them...").'

(Native teacher #2)

6. '… Particularly because of the speakers breaks not only in grammar retrieval but also pauses in the middle of words such as "building" and "wake up," the overall fluency seems to be low-beginner.'

(Native teacher #24)

When non-native teachers' comments on the distribution of silent pauses were examined, they were found to be quite different. The non-native teachers' evaluation comments were more general and relatively more straightforward than those of native speakers. While several non-native teachers referred to specific cases where difficulties occurred, they mainly focused on the overall impression of speech, stressing the unnaturalness of pause.

7. 'He seems to be a very fluent speaker because there are not many pauses in between the sentences'

(Non-native teacher #15)

8. 'Difficulty in completing sentences and delivering meaning as there are many spaces between words'

(Non-native teacher #20)

9. 'Pauses occur very often in one sentence, and as a result, the pace is very slow.'

(Non-native teacher #26)

Contrary to the teacher groups, the peer group's comments did not refer to the position of silent pauses, neither responsively nor carefully. A total of twenty-six remarks on silent pause distribution was frequently too brief to interpret their judgment. The majority of peer comments did not directly address the silent pause issue. Instead, they tend to express a general perception of communication that includes silent pauses.

10. 'The space between the words is too long.'
(Peer student #13, translated into English by the researcher)

11. 'Unnatural pause occurs between words.'
(Peer student #7, translated into English by the researcher)

The three rater groups' reactions to speech rate followed a similar pattern. The native teacher group provided elaborated comments on speech rate, paying close attention to the speed change and often stating how the L2 speaker's speed and flow had altered throughout the sample. Furthermore, many native teachers described a speaker's rate or flow in relation to other fluency criteria, such as silent pauses, vocabulary, or grammar, rather than referring to the speaker's speech rate in isolation.

12. '… The speaker seems to hesitate and fumble more towards the beginning showing more fluency towards the end of the recording, which is why I chose a high-intermediate rating for this speaker.'

(Native teacher #24)

13. 'The beginning started off well; good pace and the phrases were executed pretty fluently was nice. but then he hit a wall. What to say next? there was a long pause and after that, the flow had been broken and the end was choppy.'

(Native teacher #22)

14. 'This speaker spoke somewhat slowly, with the first sentence being the most fluent, however, as the participant produces the next two sentences fluency drops to with somewhat uniform breaks in between. The first sentence may generally be produced more fluently as this is a question frequently asked from a lower level of language study.'

(Native teacher #23)

Although some non-native teachers highlighted the spots where speed issues arose and related speed problems to other fluency indicators, like the native teachers, non-native teachers generally provided less detailed and elaborated comments than native teachers.

15. *'I think the speaker is not fluent because it takes too long to think of words.'*

(Non-native teacher #3, translated into English by the researcher)

16. *'The speaker chose right vocabularies to explain the situation and spoke quickly and naturally, hesitating only a little bit.'*

(Non-native teacher #2)

Unlike the two teacher groups, the peer group did not point out problem areas or explain why speed problems arose. Although speech rate was the most often reported criteria, accounting for nearly 30 percent, the peer group's responses were more generic. The majority of the peer group's remarks focused on the speech rate by itself or described overall flow in terms of speech rate.

17. 'The talker talked fast, and he didn't mumble.'

(Peer student #16)

18. 'The speaker stops talking often and doesn't speak fast. There are some parts where the speaker stops for a long time and talks again.'

(Peer student #27, translated into English by the researcher)

The tendency of the native teachers to provide more detailed,

elaborated comments compared to the non-native teacher group on specific criteria needs careful interpretation (Kim, 2009). According to Kim (2009), non-native teachers in the EFL context could be poorly informed about how to evaluate students' language performance without depending on numeric scores and traditional fixed response assessment. She also emphasized the importance of different evaluation cultures, arguing that the culture may have contributed to the different evaluation behaviors. In addition, the discrepancy in comments between both groups of teachers and peers was supported by several previous studies. For example, Cheng and Warren's (2005) findings suggested that students and teachers differed in their marking behaviors and interpretations of oral language proficiency.

3.3 Summary and Discussion

Study 1 investigated the differences in perceived fluency by the three groups of judges across two task types and different oral proficiency levels. Additionally, by analyzing listeners' written comments on their overall perceptions of fluency, linguistic elements affecting the three groups of judges' fluency perceptions were also examined. The following summarizes and discusses findings, with answering the first research question.

To begin, according to the one-way ANOVA results, the fluency ratings of the native and non-native teacher groups were found to be very similar, but the peer group consistently rated lower on both tasks than the two EFL teacher groups. The descriptive statistics also indicated substantial mean differences among the three groups on task 1 (picture narration), task 2 (spontaneous speech), and task 1 and 2 combined. This suggested that the raters' linguistic and professional backgrounds influenced the fluency ratings across the two tasks.

After initially confirming the discrepancy among the three groups of judges, the questions of how the three groups of judges differ in their judgments of fluency across the two task types for speakers with different levels of oral proficiency were addressed. After dividing the L2 speech samples into the three oral proficiency levels, two-way ANOVAs were performed at each proficiency level (low, mid, and high) to determine the main effects of rater group and

task on fluency ratings and their interaction effects at the three proficiency levels. The results indicated that the main effect of the rater groups was verified at all three levels, revealing that the peer group assigned significantly lower fluency scores than both EFL teacher groups on both tasks, and the fluency scores between the two EFL teacher groups did not differ significantly. However, the main effect of task types was confirmed only at the low proficiency level, demonstrating that students with low proficiency levels received higher ratings from task 2 (spontaneous speech) than task 1 (picture narration). The interaction effects of task and group were not found to be significant for the three levels of students.

Returning to the first research question, the results of a series of statistical analyses answered research question 1-1 by revealing that the two EFL teacher groups had comparable severity patterns but peers assigned considerably lower fluency scores on the two tasks across all proficiency levels. Simply put, fluency ratings were similar between native and non-native teacher groups but differed between the two teacher groups and the peer group. Then, research question 1-2 could also be answered by stating that the three rater groups usually assigned a lower fluency score to the picture narration task than the spontaneous speech task, but the difference was significant only for students with low proficiency levels.

The findings are consistent with previous researches (e.g., Kim, 2009; Zhang & Elder, 2011). Kim (2009) reported that both

native English teachers and non-native Korean English teachers maintained acceptable internal consistency and severity level, although the native teachers provided more elaborate written comments than the non-native teachers. Zhang and Elder (2011) also found comparable findings in their study of how Chinese English teachers' judgments differed from those of native English teachers. They discovered no significant difference in the judgments of the two rater groups, although the two groups did not completely agree on which notions of oral proficiency were more critical. Meanwhile, the differences discovered between the two teacher groups and the peer group were corroborated by Patri's (2002) research, which revealed that peer students and teachers frequently judged and perceived oral proficiency differently when assessment criteria were not explicitly stated.

The finding that the main effect of the task was only verified at the low proficiency level needs more discussion. It implies that fluency judgments of low-level students should be interpreted cautiously, as task type had a significant effect on low-level students' fluency ratings. The results that the low proficiency level learners did better on task 2 (spontaneous speech) than on task 1 (picture narration) might be attributed to the task-dependent variability, which required different linguistic, functional, and cognitive strategies (Kim, 2009). Given that picture narration requires a speaker's degree of freedom to be constrained, such as selecting lexical items or grammatical structures that fit well with the picture

narration (Derwing et al., 2004), it may be considerably more difficult for low-level students to complete picture narration tasks than for mid- and high-level students. Additionally, the spontaneous speech task may be perceived as relatively easy for low-level students because it involves accessing information well known to the speaker (Foster & Skehan, 1996) and may provide a range of communicative strategies for L2 speakers to compensate for their linguistic difficulties (e.g., avoiding complicated words or phrases and pronunciation). Therefore, it could reduce cognitive effort and increase the low-level speaker's autonomy.

Following statistical analyses, qualitative analyses of the three groups' written comments on the fluency ratings were conducted in order to gain a more complete and enriched understanding of the three rater groups' rating practices. First, the analysis of the frequency of comments revealed that the EFL teacher groups' and peer group's overall number of comments varied to a great degree (NS: 54.5 comments per person, NNS: 47.6 comments per person, Peer: 42.5 comments per participant). This could be explained by the fact that peer students were not experts and were unaccustomed to making written reviews of their peers' oral productions. Next, the written comments made by the three groups of listeners were analyzed in terms of mood (good, negative, and neutral). The proportion of positive comments was consistent across all the three rater groups (around 30%). In contrast, the proportion of negative responses from peer students was significantly higher than that from

native and non-native teachers, implying that the peer group may be more critical of their peers' fluency than both teacher groups and that these negative moods may contribute to peer students' low fluency ratings.

Then, the frequency distribution of the comments by the native teacher group, non-native teacher group and peer group was examined. Despite their disparate backgrounds, the findings confirmed that the three groups of judges appeared to be paying attention to the same features of oral production when assigning fluency scores. It was especially shown that the native teacher group and the non-native teacher group emphasized almost the same fluency criteria (i.e., speech rate, the number of silent pauses, vocabulary, and segmental pronunciation) when evaluating L2 speakers' fluency. These trends suggested that the two teacher groups had similar views on students' speaking fluency. The findings corroborated Kim's (2009) research, which discovered that native and non-native teachers had similar perspectives on assessing L2 oral proficiency, with vocabulary, pronunciation, and overall language usage serving as the critical evaluation criteria. Similarly, it was found that the peer group prioritized linguistic features that were similar to those of both EFL teacher groups. However, the difference between the two teacher groups and the peer group appeared to be that the latter group placed the most significant emphasis on the top three criteria (i.e., speech rate, the number of silent pauses, and segmental pronunciation), accounting for more than two-thirds of total responses. This

discrepancy might be due to differences in professional background between the peer group and two teacher groups. The peer group was often unable to identify various linguistic factors affecting fluency compared to the expert groups.

In addition, the frequency distribution of the comments also revealed discrepancies between the native teacher group and other groups. The native teacher group appeared to be more concerned with the distribution of silent pauses and grammar than the non-native teacher or peer group, in addition to providing many more comments. The discrepancy could be explained by the three rater groups' different sensitivity to the location of silent pauses and grammar errors in speech. Native teachers appeared to be more bothered than non-native teachers and peers by extended silent pauses within clauses and less tolerant of grammar errors when delivering messages.

Lastly, the qualitative analysis of the written comments revealed that native teachers were more specific and elaborate in their comments than the non-native teachers and peers. For example, when examining the distribution of silent pauses, the native teacher group provided more comments and detailed feedback, mentioning specific locations and possible explanations for problematic areas. By contrast, the non-native teachers' evaluation comments were comparatively more general and straightforward than those of native speakers. The peer group's comments did not refer to the position of silent pauses, neither responsively nor carefully.

Based on the findings of qualitative studies, research question 1-3 could be answered by stating that while similar evaluation criteria (e.g., speech rate, the number of silent pauses, and segmental pronunciation) influenced the three rater groups' fluency judgments, the native teachers' written comments were more detailed and elaborated than those of the non-native teachers and peer students. The findings were consistent with Kim's (2009) research, which revealed through qualitative analysis that native teachers provided far more detailed and elaborate comments than non-native teachers.

By integrating quantitative and qualitative research methodologies, the current study gained a holistic picture of the three rater groups' perceived fluency. The qualitative analysis, especially, provided insight into the different ways in which the three groups assessed fluency, beyond findings from the quantitative analysis alone (Kim, 2009). Three major findings in Study 1 are summarized below.

First, the quantitative analysis revealed that the native and non-native teachers had comparable severity patterns on both tasks across all proficiency levels. This was further reinforced by a qualitative study, which demonstrated that both native and non-native teacher groups appeared to be influenced by a similar general impression of temporal phenomena. These findings support the assertion that non-native teachers are equally capable of serving as fluency raters as native teachers. Although it was revealed in the qualitative analysis that the native teachers emphasized the distribution of silent pauses and grammar and delivered more

extensive and elaborated comments than the non-native teachers, the variation in fluency criteria and the degree of elaboration did not lead to significant discrepancies in perceived fluency between the two EFL teacher groups.

Second, the quantitative results found that the peer group assigned significantly lower fluency ratings than both EFL teacher groups on both tasks. This was also backed up by the qualitative analysis, which revealed that negative comments from peers were substantially higher than those of native and non-native teachers. However, the peer students' distinct grading pattern requires careful interpretation. Their tendency to assign harsh scores and provide unfavorable remarks on their peers' speech performance should not be used as a basis to rule out the possibility of peer students' participating as fluency raters in the EFL classroom. Since what variables contribute to the discrepancies between the teacher and peer groups has been understudied, there is a chance to bridge the gap by better understanding the underlying mechanisms students usually use when evaluating fluency and designing a more customized fluency curriculum. Furthermore, students can benefit from a peer assessment process, which allows them to study and monitor their learning process (Cheng & Warren, 2005), and this ultimately helps develop their speaking fluency.

Third, the quantitative analysis discovered that task types significantly affected low-level students' fluency ratings, with spontaneous speech scoring higher than that of picture narration.

Although the qualitative analysis was conducted using comments on both tasks rather than on individual tasks, several rater participants later stated that it appeared that low-level students were more troubled by picture narration than spontaneous speech due to their limited linguistic competence in describing and narrating pictures. Thus, fluency judgments of the low-level students made by EFL teachers and peer groups should be interpreted cautiously. Furthermore, as some researchers (e.g., Ejzenberg, 2000) suggested, multiple tasks rather than a single task may be more beneficial in reliably capturing L2 oral fluency, particularly among students with low proficiency levels.

The qualitative findings showed that non-temporal variables such as pronunciation, grammar, and vocabulary influenced perceived L2 fluency. However, it was temporal variables, such as pause and speech rate, that had the most influence on the fluency ratings of the three listener groups. Thus, in the next chapter, the research investigates to what extent temporal variables (utterance fluency measures) relate to the native teachers', non-native teachers' and peers' fluency ratings by relating their subjective judgments with objectively measured temporal features.

CHAPTER 4.
STUDY 2: UTTERANCE FLUENCY

This chapter reports the results from the experiment conducted on the relationship between utterance fluency (temporal features classified as breakdown, speed, and repair fluency measures) and perceived fluency (overall impression) by the native teacher, non-native teacher, and peer groups, exploring the underlying mechanism of the three groups' disparities in fluency perceptions. The methodology for the experiment is described in Section 4.1, and Section 4.2 reports the results. Finally, Section 4.3 summarizes the chapter by discussing the results.

4.1 Methodology

The previous chapter investigated how the perceived fluency of native teacher, non-native teacher and peer groups differ. While it was established that subjective judgments (i.e., fluency ratings) varied among the raters across the two tasks, it is critical to elucidate the relationship between their subjective judgments and objectively measurable features. Thus, chapter 4 attempts to determine which measures of utterance fluency are associated with the three groups of listeners' perceived fluency and decision-making concerning fluency levels.

4.1.1 Participants and Procedures

As with Study 1, ninety-four individuals, comprising of 26 native English teachers, 29 Korean English teachers, and 39 peer students, participated as raters for Study 2. Each native English teacher (*M* age = 28.3, *SD* = 6.1, *Min.* = 22, *Max.* = 43) had various experiences teaching English as a Foreign Language in Korean middle and high schools. The second raters were all non-native English teachers (*M* age = 37.7, *SD* = 7.6, *Min.* = 26, *Max.* = 56) who were working in middle and high schools at the time of conducting the research. The final judgment panel was composed of 39 10th grade peer students (*M* age = 16.6, *SD* = 0.6, *Min.* = 15, *Max.* = 17) enrolled in the English Conversation I course who were uninformed of the assessed participants. 30 Korean high school students (*M* age = 17.9, *SD* = 0.3, *Min.* = 17, *Max.* = 18) participated as speakers. They were all in the 11th grade at the time of the research and had finished the English Conversation I and II courses.

The speech samples utilized in Study 2 were responses to task 1 (picture narration task). Around 30-second excerpts were taken from the beginning for presentation to the raters, with initial dysfluencies, false starts, and hesitations eliminated (Derwing et al., 2004; Rossiter, 2009). Every fragment began and ended at a phrase boundary (Kahng, 2014).

The raters listened to 30 speech samples in random order and rated their level of fluency using a nine-point scale following a practice

session. After completing the scale, raters were asked to write about their overall impression of the L2 speech sample's fluency.

4.1.2 Temporal Measures

In line with previous researches (e.g., Bosker et al., 2013; Kahng, 2014; Préfontaine et al., 2016), a total of ten utterance temporal measures were purposefully selected and calculated. The ten chosen temporal measures and their operational definitions are listed in Table 4.1. The choice of measures was made so that the measures clearly represented each aspect of fluency (i.e., speed, breakdown, and repair fluency), and they have been found to demonstrate little intercollinearity (Kahng, 2014).

Table 4.1 Selected Utterance Fluency Features

Fluency aspect	Utterance features	Definitions
Speed fluency	*Articulation rate*	Total number of syllables / speech time excluding pause time
	Mean length of run	Average number of syllables produced between two silent pauses (250ms)
Breakdown fluency	*Mean length of silent pauses*	Total length of silent pause time / number of silent pauses
	Mean length of filled pauses	Total length of filled pause time / number of filled pauses
	Number of silent pauses per minute	Total number of silent pauses / total time

	Number of filled pauses per minute	Total number of filled pauses / total time
	Silent pause rate within a clause	Number of silent pauses within a clause / number of clauses
	Silent pause rate at a clause boundary	Number of silent pauses at a clause boundary / number of clause boundaries
Repair fluency	*Number of corrections per minute*	Number of corrections / total time
	Number of repetitions per minute	Number of repetitions / total time

For speed fluency, the articulation rate and the mean length of run were computed as they were revealed to be strong predictors of fluency in the previous researches. The articulation rate was intentionally chosen since it is a pure measure of speed that does not include pause time and has been widely used in recent studies (e.g., Bosker et al., 2013; De Jong et al., 2013; Kahng, 2014). Additionally, the mean length of run was included since it has been consistently associated with the development of L2 oral fluency and perceived fluency in prior studies. Further, it appeared closely related to automated speech production while long fluent runs facilitate L2 oral fluency (Kahng, 2014). Initially, along with the articulation rate and the mean length of run, the speech rate (pruned and unpruned) and the phonation-time ratio were computed as well. However, multicollinearity testing revealed that they had the highest variation inflation factor (VIF) values (> 10) which means they were highly associated and could be predicted from the others. Thus, speech rate

and phonation-time ratio were later excluded from the experiment.

To investigate the breakdown fluency, the number of silent and filled pauses per minute and the mean length of silent and filled pauses were used to quantify the frequency and duration of silent and filled pauses. Furthermore, to measure the distribution of silent pauses, this study used the silent pause rate in different locations (i.e., within a clause and at a clause boundary), which was devised by Kahng (2014). Contrary to pause measures used in the previous studies, such as the number of pauses per minute, the pause rate accurately computes how often a speaker pauses within a clause or at each clause boundary, which captures a more accurate picture of pause distribution. Kahng's (2014) study originally included both silent and filled pause rates. However, this study employed only the silent pause rate because it was discovered in her research that the silent pause rate was strongly associated with L2 oral fluency and L2 speaking scores and successfully separated L1 from L2 speakers. However, the filled pause rate did not appear to be connected to fluency or L2 speaking scores.

Finally, the amount to which an L2 speaker repairs their speech is often quantified by counting the number of reformulations, false starts, self-corrections, repetitions, replacements, or hesitations per minute (Tavakoli et al., 2020). Among the various repair measures, the number of corrections per minute and the number of repetitions per minute were calculated in accordance with the previous studies (Bosker et al., 2013; Kahng, 2014).

4.1.3 Acoustic Analysis

Acoustic analysis of L2 speech samples was conducted to investigate the relationship between subjective fluency ratings and utterance measures. To begin, all speech excerpts were meticulously transcribed, including information regarding silent pauses (250ms). The silent pause cut-off in the current study was set at 250ms in line with the previous researches (e.g., Bosker et al., 2013; Ginther et al., 2010; Kahng, 2014; Towell et al., 1996), as a cut-off of 250ms resulted in the highest correlation between the number of silent pauses and L2 proficiency scores (De Jong & Bosker, 2013). In addition, Kahng (2012) further supported a 250ms cut-off point by comparing the analysis results based on two cut-off points (250, 400ms). She discovered that the 400ms missed 12 percent of the pauses identified by 250ms. More notably, 77 percent of the pauses which 400ms missed were pauses within clauses, one of the essential focuses of the present study. Then, silent pauses equal to or longer than 250 ms were identified as silent pauses, and shorter than 250ms were classified as micro-pauses (Riggenbach, 1991) in the current dissertation.

For the speed fluency, the articulation rate was computed by dividing the total number of syllables by speech time, excluding pause time (silent and filled pauses). The number of syllables was counted using an online program called Syllable Counter

(syllablecount.com, n.d.), which was created using an English syllable dictionary. After entering transcripts of speech recordings into the software's window, it generated a result table containing the number of words and syllables in the given text (Kahng, 2014). Additionally, the mean length of run was calculated to determine the average number of syllables produced between two silent intervals (250ms).

For the breakdown fluency, the duration of silent and filled pauses was measured in milliseconds (ms) by listening to each speech excerpt and examining the waveform and spectrogram via Praat (Boerma & Weenink, 2012); the duration of silent pause was marked in parentheses (in milliseconds), and the duration of filled pauses was marked right next to each filled pause without parentheses. Next, the frequency of silent and filled pauses was then calculated by counting the number of filled and silent pauses per minute. Lastly, the distribution of silent pauses was operationalized as silent pause rates within a clause and at a clause boundary. In the transcript, a clause boundary was marked by a double slash .//... The number of silent pauses within a clause and at a clause boundary were counted and divided into the number of clauses and clause boundaries.

The repair fluency, such as self-repetitions and self-corrections, was put inside brackets {...}. Then, the number of syllables used in repairs was counted. The following is an example of a transcript with information about clause boundaries, repairs, and the duration of silent and filled pauses.

there was a (751) {two} uh455 one woman and a man (615) // and they walked (1190) // and they hit (438) // after that (1184) uh489 (1019) um222 they said sorry to each other // and they take (973) their bag // and (1087) {go} go each other (711) //

4.1.4 Statistical Analysis

Firstly, a Pearson correlation analysis was conducted to investigate the relationship between ten utterance fluency features and each group of judges' (NS, NNS, peer group) fluency ratings, revealing which temporal features have the most impact on each appraiser's fluency ratings.

Next, stepwise multiple regression analyses were used to determine which collection of objective acoustic measures best explains the variance in the predicted fluency ratings. Three different groups of judges' perceived fluency ratings served as dependent variables, whereas six pre-selected temporal aspects served as independent variables. Prior to conducting stepwise multiple regression analyses, the researcher ensured that the VIF values of the independent variables were all below 10 to avoid multicollinearity issues.

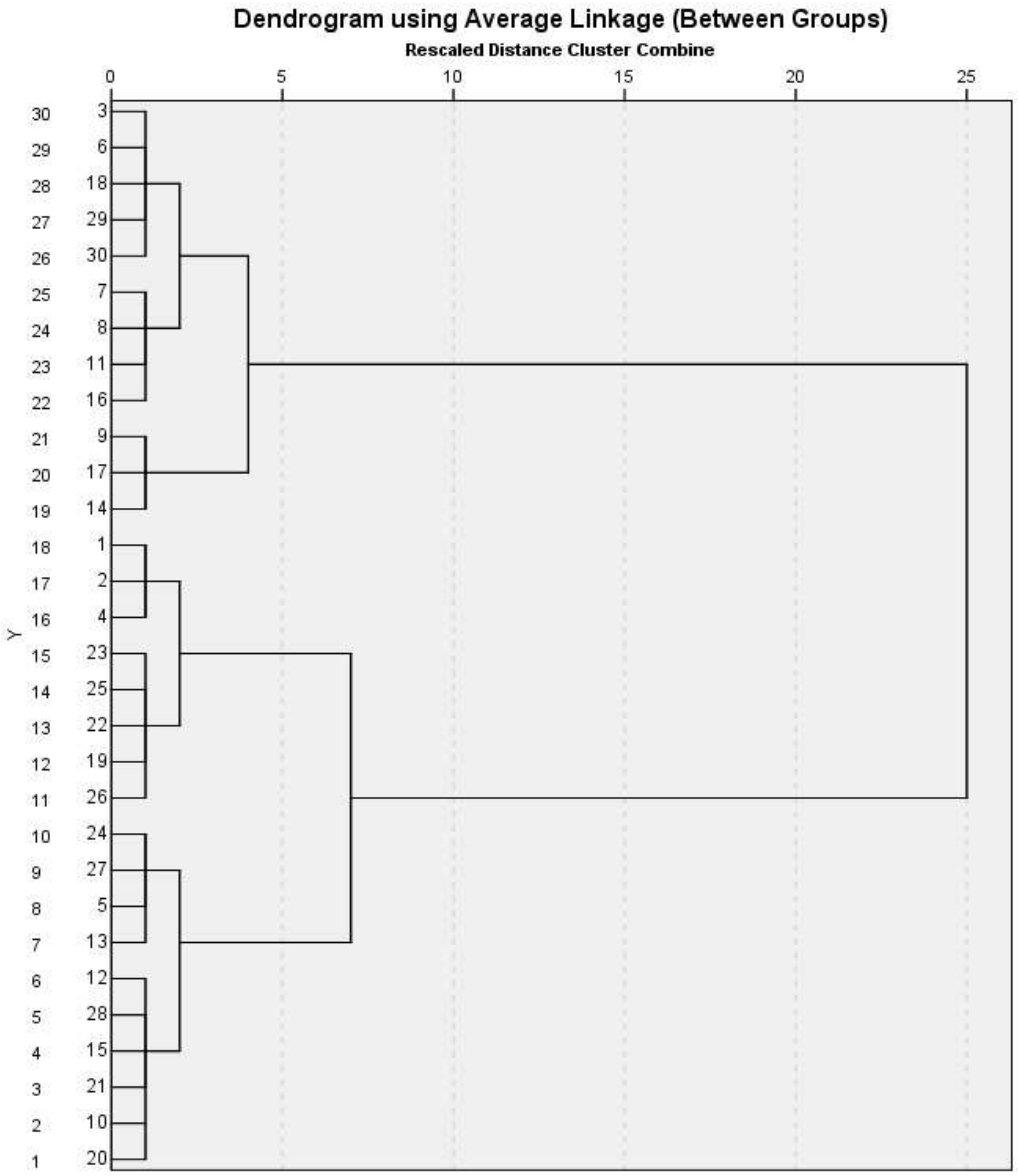
Lastly, a cluster analysis and a series of one-way ANOVAs were performed to address the commonly raised question of which utterance measures distinguish speakers' fluency levels by the three

groups of raters. First, a hierarchical cluster analysis was used to determine the number of homogenous groups of perceived fluency in 30 Korean speakers (see Figure 4.1) (Saito et al., 2018). Based on an inspection of the dendrogram, which visualized the clustering results, a three-factor method was chosen and split 30 L2 speakers into three groups (low, mid, and high fluency groups). Then, using K-mean cluster analysis, all speech samples (n = 30) were classified into three smaller homogenous groups (see Table 4.2). After classifying the 30 L2 speech samples into the three distinct fluency levels, three one-way ANOVAs and post-hoc comparisons were performed to identify which temporal characteristics influence the three rater groups' choice of fluency leveling. Investigating this relation can be essential to gain insight into how different groups of listeners arrive at judgments and what aspects of speech are considered when judging fluency.

Table 4.2 Numbers of Assigned Students by Fluency Levels

Group	NS	NNS	Peer
Low Fluency	9	12	11
Mid Fluency	10	8	10
High Fluency	11	10	9

Figure 4.1 Dendrogram Tree of Hierarchical Clusters Based on the Participants' Perceived Fluency Ratings



4.2 Results

4.2.1 Predictors of Three Rater Groups' Fluency Ratings

The twenty-six native, 29 non-native, and 39 peer raters rated the fluency of the 30 speech samples. First, the interrater reliability test was conducted to ascertain the degree of agreement between each group of raters. Cronbach's alpha coefficients showed .988 for native teachers, .989 for non-native teachers, and .988 for peers. According to the test results, each group of listeners indicated a high degree of agreement about their perceptions of L2 speech. As a result, the researcher could average perceived fluency ratings from each group of judges and utilize them as variables. Table 4.3 displays the descriptive statistics of ten selected utterance fluency features as well as the fluency ratings of the three groups of listeners. L2 speakers demonstrated a range of performance in terms of speed, breakdown, and repair, as expected from a wide range of overall oral proficiency.

Table 4.3 Descriptive Statistics of Ten Utterance Fluency Features and Fluency Ratings of L2 Speech

	Utterance fluency features	*N*	*M*	*SD*	*Min.*	*Max.*
Speed fluency	*articulation rate*	30	3.53	0.68	2.45	5.06
	mean length of run	30	5.83	3.13	2.18	14.00
Breakdown fluency	*mean length of silent pauses*	30	0.72	0.23	0.42	1.36
	mean length of filled pauses	30	0.21	0.22	0.00	0.96
	number of silent pauses per minute	30	0.43	0.11	0.24	0.65
	number of filled pauses per minute	30	0.06	0.07	0.00	0.27
	silent pause rate within a clause	30	1.27	0.82	0.11	3.20
	silent pause rate at a clause boundary	30	0.58	0.24	0.11	1.00
Repair fluency	*number of repetitions per minute*	30	0.04	0.05	0.00	0.16
	number of corrections per minute	30	0.04	0.03	0.00	0.10
Fluency Rating	NS fluency ratings	30	5.92	1.73	3.23	8.77
	NNS fluency ratings	30	6.07	1.70	3.62	8.86
	Peer fluency ratings	30	5.29	1.55	2.90	8.26

To begin, the speed fluency measurements such as the articulation rate and the mean length of run varied significantly among speakers. The average articulation rate was 3.53, with a low of 2.45 and a high of 5.06. The average number of syllables produced between silent pauses (≥ 250ms) demonstrated an even greater gap ranging from 2.18 to 14. The mean length and the number of silent pauses ranged from 0.42 to 1.36 and 0.24 to 0.65, respectively. The

mean length and the number of filled pauses varied between 0 to 0.96 and 0 to 0.27, respectively, indicating that the speakers in the study used fewer and shorter filled pauses than silent pauses. As for the silent pause rate, the silent pause rate within a clause varied between 0.11 and 3.20, while the silent pause rate at a clause boundary varied between 0.11 and 1.00. The participants demonstrated lower variation in the repair fluency measures, such as the number of repetitions (0 to 0.16) and corrections per minute (0 to 0.1). Finally, fluency ratings from the three rater groups were computed. As demonstrated by the ANOVA results in Study 1, the mean fluency rating of the peer group was significantly lower than those of the EFL teacher groups.

To determine intercorrelations between utterance fluency measures and to investigate the relationship between utterance fluency measures and the fluency ratings of each group of judges, a Pearson correlation analysis was performed on the L2 speakers' data on speed (articulation rate, mean length of run), pause phenomena (number, mean length of silent and filled pauses, distribution of silent pauses), and repairs (number of repetitions, corrections per minute) and perceived fluency of each group (see Table 4.4).

First, the articulation rate was shown to be positively connected with the mean length of run (r = .772**), negatively correlated with the mean length of silent pauses (r = -.597**), and the silent pause rate within a clause (r = -.591**). The mean length of run was found to be highly linked with several silent pause measures. It was moderately related to the mean length of silent

pauses (r = −.557**) and the silent pause rate at a clause boundary (r = −.416*), and strongly correlated with the number of silent pauses per minute (r = −.792**) and the silent pause rate within a clause (r = −.788**). As Kahng (2014) noted, the connections between mean length of run and silent pause features were expected as the mean length of run is calculated using the number of silent pauses. It is worth noting, however, that the mean length of run had a stronger link with the silent pause rate within a clause (r = −.788**) and a much lower correlation with the silent pause rate at a clause boundary (r = −.416*). Additionally, it is worth mentioning that the silent pause rate within a clause exhibited a moderately strong association with the articulation rate (r = −.591**), despite the fact that they were not mathematically related, as indicated in Kahng's study (2014). The number of repetitions per minute was unrelated to the number of corrections per minute or any of the other speed and breakdown variables. On the other hand, the number of corrections per minute was positively correlated with the number of silent pauses per minute (r = .523**) and the silent pause rate within a clause (r = .434*).

Table 4.4 Correlations Between Utterance Features and Fluency Ratings of the Three Groups of Judges

	AR	MLR	LngSP	LngFP	NumSP	NumFP	SPRw	SPRc	NR	NC	NS fluency ratings	NNS fluency ratings	Peer fluecy ratings
AR	1										.705**	.762**	.828**
MLR	.772**	1									.869**	.898**	.905**
LngSP	-.597**	-.557**	1								-.764**	-.751**	-.773**
LngFP	.078	.153	-.179	1							.133	.176	.102
NumSP	-.326	-.792**	.117	-.115	1						-.635**	-.654**	-.573**
NumFP	-.015	.039	-.173	.428*	-.061	1					.097	.055	.044
SPRw	-.591**	-.788**	.363*	-.162	.801**	-.032	1				-.786**	-.801**	-.735**
SPRc	-.183	-.416*	.395*	-.130	.343	-.297	.180	1			-.303	-.354	-.370*
NR	-.279	-.279	.126	-.189	.122	-.121	.169	.082	1		-.245	-.263	-.272
NC	-.065	-.359	.010	-.191	.523**	-.169	.434*	.034	.141	1	-.388*	-.395*	-.282

Note. * = $p < .05$; ** = $p < .01$. AR = articulation rate; MLR = mean length of run; LngSP = mean length of silent pauses; LngFP = mean filled pauses; NumSP = number of silent pauses per minute; NumFP = number of filled pauses per minute; SPRw = silent pause rate within a clause; SPRc = silent pause rate at a clause boundary; NR = number of repetitions per minute; NC = number of corrections per minute.

Finally, correlation analyses among the three groups of judges' fluency ratings and utterance fluency measures were examined in order to identify which utterance fluency features were most related to the native teacher, non-native teacher, and peer groups' fluency ratings (see Table 4.5). The native teacher group's (NS) fluency ratings had a strong correlation with the mean length of run (r = .869**), the silent pause rate within a clause (r = -.786**), the mean length of silent pauses (r = -.764**), and the articulation rate (r = .705**). They were moderately correlated with the number of silent pauses per minute (r = -.635**) and weakly related to the number of corrections (r = -.388*). In the case of the non-native teachers' (NNS) fluency ratings, it was discovered that the mean length of run (r =.898**) was the most closely related to NNS fluency ratings, followed by the silent pause rate within a clause (r = -.801**), which was the same as the native teachers' finding. Another speed measure, the articulation rate (r = .762**), showed a third strong correlation with fluency ratings. Both the mean length of silent pauses and the number of silent pauses per minute (r = -.751** and r = -.654**, respectively) revealed a negative relationship with the non-native teachers' fluency judgments. Correlations between the number of corrections per minute and the fluency evaluations of NNS were found to be weak (r = -.395*). It is worth noting that the top six utterance traits associated with the native and non-native teachers were identical, despite minor ranking variances. As for the peer group, the speed measures, the mean length of run (r = .905**)

and the articulation rate (r = .828**) were found to be the most correlated with peer group fluency ratings, followed by the silent pause measures – the mean length of silent pauses (r = -.773**), the silent pause rate within a clause (r = -.735**), and the number of silent pauses per minute (r = -.573**). Interestingly, in contrast to the results of both EFL teacher groups, the silent pause rate at a clause boundary demonstrated a weak negative association with fluency ratings (r = -.370*) for the peer group.

Table 4.5 Correlation of Utterance Features with the Three Groups of Judges' Fluency Ratings

NS's fluency ratings	*r*	NNS's fluency ratings	*r*	Peer's fluency ratings	*r*
Mean length of run	.869**	*Mean length of run*	.898**	*Mean length of run*	.905**
Silent pause rate within a clause	-.786**	*Silent pause rate within a clause*	-.801**	*Articulation rate*	.828**
Mean length of silent pauses	-.764**	*Articulation rate*	.762**	*Mean length of silent pauses*	-.773**
Articulation rate	.705**	*Mean length of silent pauses*	-.751**	*Silent pause rate within a clause*	-.735**
Number of silent pauses per minute	-.635**	*Number of silent pauses per minute*	-.654**	*Number of silent pauses per minute*	-.573**
Number of corrections per minute	-.388*	*Number of corrections per minute*	-.395*	*Silent pause rate at a clause boundary*	-.370*

Note. * = $p < .05$; ** = $p < .01$

To summarize the utterance fluency correlation findings, the speed measurements such as the articulation rate and the mean length of run were highly connected with the silent pause measures, such as the mean length of silent pauses and the number of silent pauses per minute, but not with the filled pause measures. These findings corroborated Kahng's (2014) findings. No statistically significant link between the two measures within the repair fluency was found, consistent with Bosker et al.'s (2013) findings. Finally, four utterance variables, including the mean length of run, the articulation rate, the silent pause rate within a clause, and the mean length of silent pauses, were found to be strongly linked with the fluency judgments of the three groups of judges in common. The four utterance features were not only substantially connected with perceived fluency ratings in the three groups of judges but also with one another within variables. Moreover, it is also worth mentioning that the fluency ratings of the three groups of raters were most closely related to the mean length of run.

4.2.2 A Best Prediction Model on L2 Speaking Fluency

The second objective of Study 2 was to identify the best predictive model which could explain the most variance in the L2 speaking fluency ratings by the native teacher, non-native teacher, and peer groups. In order to address this issue, stepwise multiple regression analyses were performed with six significant predictors, as shown in

Table 4.5. Table 4.6 presents the best regression model for the native teacher group.

Table 4.6 The Best Regression Model for the Native Teacher Group

Variables	B	β	t-value	sig. p	Step Entered	R^2 change	VIF
Mean length of run	.172	.310	3.171	.004	1	0.746	3.446
Mean length of silent pauses	-3.559	-.477	-7.260	.000	2	0.113	1.555
Silent pause rate within a clause	-.652	-.309	-3.465	.002	3	0.047	2.854
Numbers of corrections per minute	-7.690	-.138	-2.301	.030	4	0.013	1.294
Final model R^2 = .930, F = 83.656, p < .000, Adjusted R^2 = .919, Durbin Watson = 2.024							

The final model's F value of 83.656 and p <.000 demonstrated that both variables and the regression model were statistically meaningful. Additionally, the adjusted R^2 (= .919) indicated that the final model accounted for around 92 percent of the variance in fluency ratings assigned by the native teachers. The regression model also revealed that the mean length of run, entered first, explained the most variance (74.6%) in the native teachers' fluency judgments, followed by the mean length of silent pauses, which accounted for an additional 11.3 percent. The inclusion of the silent pause rate within a clause explained an extra 4.7 percent variance. Finally, the number of

corrections was demonstrated to be the least important factor in determining native teachers' fluency ratings, explaining only an additional 1.3 percent. Overall, L2 speakers were judged to be more fluent by native teachers when they used a high tempo of speech with shorter silent pauses and when they paused less within a clause and made fewer self-corrections.

Table 4.7 The Best Regression Model for the Non-Native Teacher Group

Variables	B	β	t-value	sig. p	Step Entered	R^2 change	VIF
Mean length of run	.213	.392	4.718	.000	1	.800	3.446
Mean length of silent pauses	-3.148	-.429	-7.698	.000	2	.090	1.555
Silent pause rate within a clause	-.585	-.281	-3.725	.001	3	.040	2.854
Numbers of corrections per minute	-7.011	-.128	-2.514	.019	4	.012	1.294
Final model R^2 = .975, F = 118.775, p < .000, Adjusted R^2 = .942, Durbin Watson = 2.324							

As demonstrated in Table 4.7, the utterance fluency features that influence fluency judgments for non-native teachers were identical to those for native speakers. The ratings of non-native teachers were mainly detected using silent pause and speed measures. The mean length of run was the strongest predictor, followed by the mean length of silent pauses, with the two variables accounting for

approximately 90 percent of the variance in non-native teachers' fluency judgments. The silent pause rate within a clause and the number of corrections added a minor variance. With four variables in the final model, about 94 percent of the variance can be explained. Lastly, Table 4.8 shows the best regression model for the peer students group.

Table 4.8 The Best Regression Model for the Peer Group

Variables	B	β	t-value	sig. p	Step Entered	R^2 change	VIF
Mean length of run	.286	.575	7.167	.000	1	.812	2.569
Mean length of silent pauses	-2.324	-.347	-5.464	.000	2	.106	1.612
Articulation rate	.401	.176	2.114	.044	3	.009	2.756
Final model R^2 = .935, F = 124.181, p < .000, Adjusted R^2 = .927, Durbin Watson = 1.882							

As shown in Table 4.8, the best regression model for the peer group differed from both EFL teacher groups in terms of entered variables. Compared to the native and non-native teacher groups, the peer group's final model excluded the number of corrections per minute and the silent pause rate within a clause but included the articulation rate, indicating that speed measures might be more related to peer fluency ratings. However, the most robust predictor, explaining about 82 percent of the variance in peer fluency

judgments, was the mean length of run, which was consistent with the native and non-native teacher groups. In addition, like both EFL teacher groups, the mean length of silent pauses was found to be the second most important variable in peer fluency judgments, explaining about an extra 10 percent of the variance. The articulation rate added only a small amount of the variance in the peer model. Together, the three variables explained approximately 93 percent of the variance in fluency assessments.

In summary, the results have demonstrated that native and non-native teacher groups and the peer group used a similar predictive model when assessing L2 fluency. The two strongest predictors, which explained most of the variance in the three regression models, were the same among three groups: the mean length of run and the mean length of silent pauses, showing that the three groups evaluated L2 fluency using the same phonetic features. However, the data further demonstrated that four entered utterance features and relative contribution rankings were the same in the native and non-native teacher groups. In contrast, the peer group's model differed from both EFL teacher groups regarding variables included in the regression model.

In addition, it is worth noting that the β values of the variables in the three final models also revealed some differences. The β coefficient is the degree of change in the dependent variable for every unit of change in the predictor variable (Meyers et al., 2016). According to Pedhazur (1997), in the research context, where

the independent variables are significantly correlated, it is possible to quantify the relative contribution of predictors using β weights (the absolute value) as the basis of the comparison. In other words, it may be possible to say that the predictors with larger β weights contribute more to the prediction of the dependent variable than those with smaller weights (Meyers et al., 2016). The comparison of β values of the four variables in both EFL teacher groups' models showed that the native and non-native teacher groups had the same relative rankings for β values. The mean length of silent pauses contributed the most to the prediction of the dependent variable (β value for NS: -.477, for NNS: -.429), followed by the mean length of run (β value for NS: .310, for NNS: .392), the silent pause rate within a clause (β value for NS: -.309, for NNS: -.281), and the number of corrections per minutes (β value for NS: -.138, for NNS: -.128). However, the β values of the three variables in the peer group's model revealed some disparities from those of both EFL teacher groups, showing that the mean length of run (β = .575) was most strongly related to the prediction of the dependent variable, followed by the mean length of silent pauses (β = -.347) and the articulation rate (β = .176). Overall, the regression results indicated that the mean length of run and the mean length of silent pauses explained most of the variance in the fluency judgments of the three groups of raters. However, the β coefficients of the variables, which informed us of how much a change in the fluency score was associated with a unit difference in the utterance features, showed

some disparities between the two EFL teacher groups and the peer group.

4.2.3 Utterance Measures Distinguising Fluency Levels

The final purpose of Study 2 was to determine which utterance fluency features distinguished the three distinct perceived fluency groups (low, mid, high) by native teachers, non-native teachers, and peer students. To this end, the three fluency groups were determined based on cluster analyses of the three rater groups' fluency ratings on 60 speech samples. Next, following Saito et al.'s (2018) methodology, a set of one-way ANOVAs was run with perceived fluency level as the grouping factor and each utterance fluency measure as the dependent variable. Then, follow-up Duncan's post-hoc analysis was checked to verify which utterance measures distinguished different L2 speakers' fluency levels. Table 4.9 illustrates the native teacher group's result.

Table 4.9 Summary of Group Differences for Low, Mid, High Levels of Native Teacher Group's Perceived Fluency

	ANOVA Results			
	$F(2, 27)$	p	η_p^2	Significant Group Differences
Articulation rate	12.127	.000	0.47	Low, Mid < High
Mean length of run	41.043	.000	0.75	Low, Mid < High
Mean length of silent pauses	13.594	.000	0.50	Low < Mid < High
Number of silent pauses per minute	10.769	.000	0.44	Low, Mid < High
Silent pause rate within a clause	15.928	.000	0.54	Low < Mid < High
Number of repetitions per minute	4.326	.023	0.24	Mid < High

To begin, as shown in Table 4.9, while speed measures such as the articulation rate and the mean length of run consistently distinguished high-level of perceived fluency from lower levels among native teachers, speed measures for the low- and mid-level of fluency appeared to be similar. The mean length of silent pauses and the silent pause rate within a clause distinguished not only between mid- and high-level perceived fluency but also distinguished between low- and mid-level perceived fluency. The number of repetitions per minute feature differed between mid- and high-level perceived fluency of native teachers. In summary, speed features (e.g., articulation rate, mean length of run) and pause measures directly

connected to the perception of delivery speed (e.g., the number of silent pauses per minute) proved to distinguish high-level fluency from other levels' fluency (mid and low) in the native teacher group. Moreover, silent pause measures, such as the mean length of silent pauses and the silent pause rate within a clause, were shown to distinguish low-level fluency from other levels' fluency (mid and high). Next, Table 4.10 demonstrates the non-native teacher group's results.

Table 4.10 Summary of Group Differences for Low, Mid, High Levels of Non-Native Teacher Group's Perceived Fluency

	ANOVA Results			
	$F(2, 27)$	p	η_p^2	Significant Group Differences
Articulation rate	17.738	.000	0.57	Low, Mid < High
Mean length of run	52.293	.000	0.79	Low < Mid < High
Mean length of silent pauses	13.205	.000	0.49	Low < Mid < High
Number of silent pauses per minute	9.946	.001	0.42	Low, Mid < High
Silent pause rate within a clause	16.561	.000	0.55	Low < Mid < High

As with the native teacher group, non-native teachers employed the articulation rate and the number of silent pauses per minute to distinguish high-level fluency from low- and mid-level fluency. Additionally, the non-native teachers used one speed measure, the mean length of run, and silent pause measures, such as

the mean length of silent pauses and the silent pause rate within a clause, to identify low-level fluency from other levels' fluency (mid and high). It is worth noting that the non-native teachers, like the native teachers, used speed features (e.g., the articulation rate) to distinguish high-level from other levels and silent pause measures (e.g., the mean length of silent pauses and the silent pause rate within a clause) to separate low-level fluency from other levels in this study. Lastly, peer students' utterance features used in judging different levels of fluency are presented in Table 4.11.

Table 4.11 Summary of Group Differences for Low, Mid, High Levels of Peer Group's Perceived Fluency

	ANOVA Results			
	$F(2, 27)$	p	η_p^2	Significant Group Differences
Articulation rate	16.475	.000	0.55	Low, Mid < High
Mean length of run	35.947	.000	0.73	Low < Mid < High
Mean length of silent pauses	20.946	.000	0.61	Low < Mid < High
Number of silent pauses per minute	6.766	.004	0.33	Low, Mid < High
Silent pause rate within a clause	9.656	.001	0.42	Low, Mid < High
Silent pause rate at a clause boundary	3.803	.035	0.22	Low < High
Number of corrections per minute	4.089	.028	0.23	Mid < High

As illustrated in Table 4.11, the articulation rate, the number of silent pauses per minute, and the silent pause rate within a clause were used to distinguish high-level fluency from mid- and low-level fluency. The mean length of run, the mean length of silent pauses, and the silent pause rate at a clause boundary were utilized to distinguish low-level fluency from other levels' fluency. It is worth noting that the peer group employed the number of corrections per minute to distinguish high-level from mid-level while the native teacher group used the number of repetitions to distinguish high-level from mid-level. Additionally, high- and low-level fluency groups differed in terms of the silent pause rate at a clause boundary that the EFL teacher groups did not use. However, the peer group, like the native and non-native teacher groups, demonstrated that they could distinguish high-level fluency from lower level fluency using speed measurements (e.g., the articulation rate) and low-level fluency from higher level fluency using pause measures (e.g., the mean length of silent pauses).

Indeed, as shown in Table 4.12, the articulation rate and the number of silent pauses per minute, which was directly connected to the perception of delivery speed, were used by all three groups of judges to distinguish L2 students with a high level of fluency, demonstrating that highly fluent L2 students' speech was described as incorporating fast delivery.

Table 4.12 Summary of Utterance Measures and Perceived Levels of Three Rater Groups

	NS	NNS	Peer
Distinguishing high from lower levels	**AR** MLR **NumSP**	**AR** **NumSP**	**AR** **NumSP**
Distinguishing low from higher levels	**LngSP** SPRw	MLR **LngSP** SPRw	MLR **LngSP** SPRc

Note. AR = articulation rate; MLR = mean length of run; LngSP = mean length of silent pauses; NumSP = number of silent pauses per minute; SPRw = silent pause rate within a clause; SPRc = silent pause rate at a clause boundary.

Additionally, for all the three groups of judges, one pause measure, the mean length of silent pauses, was utilized to distinguish between low-level and other levels, implying that L2 students with low-level fluency spoke with long silent pauses.

4.3 Summary and Discussion

Study 2 investigated 1) which temporal features had the greatest influence on each group of raters' fluency ratings, 2) which acoustic model best predicted the fluency ratings of the three different groups of judges, and 3) which utterance measures had an effect on the decision-making of the three different listener groups regarding fluency levels. The following summarizes and discusses findings, along with answering the second research question.

According to the correlation analysis results, two speed measures (i.e., the mean length of run and the articulation rate) and two pause measures (i.e., the silent pause rate within a clause and the mean length of silent pauses) were most strongly correlated with the perceived fluency of the three groups of judges. On the other hand, measures of filled pauses such as the mean length of filled pauses, the number of filled pauses per minute, and the number of repetitions per minute were not associated with the fluency judgments of the three listener groups. Then, the findings answered research question 2-1 by stating that the four utterance measures, including articulation rate, the mean length of run, the silent pause rate within a clause, and the mean length of silent pauses, were most related to the three rater groups' fluency ratings, and these features would serve as significant markers of perceived fluency.

The following further discusses the four utterance fluency variables (i.e., the mean length of run, the articulation rate, the silent

pause rate within a clause, and the mean length of silent pauses) that were found to be significant predictors for all three groups of judges in the current study in relation to previous studies.

To begin, it is worth noting that in the current study, the mean length of run had the strongest link with fluency ratings across all the three groups (NS: r = .869**, NNS: r = .898**, Peer: r = .905**). The findings were consistent with numerous other researches that established that the mean run length was a strong predictor of perceived fluency (e.g., Cucchiarini et al., 2002; Kormos & Dénes, 2004; Lennon, 1990; Préfontaine et al., 2016). Additionally, it supported Kahng's (2014) study, which established a strong correlation between the mean length of run and L2 fluency by revealing a significant difference between L1 and L2 speakers, as well as a significant correlation between the mean length of run and speaking proficiency scores. The probable explanation for the substantial relationship between mean length of run and fluency perception is that run length is believed to have a conceptual connection to the processing of automatic speech production (Kahng, 2014). According to Towell et al. (1996), the most important temporal variable contributing to fluency development is an increase in the mean length of run, as the increase in the run length is primarily attributable to the proceduralization of declarative knowledge. Additionally, long fluent run appears to be related to the usage of prefabricated language units and formulaic language, both of which have been reported to facilitate L2 oral fluency (Kahng, 2014).

Another speed feature, articulation rate, which is a pure measure of speed as it does not include pause time, showed a significant positive association with each group's fluency scores (NS: r =.705**, NNS: r =.762**, Peer: r = .828**). It is worth noting that the peer group appeared to place a higher value on articulation rate (ranked second) than the EFL teacher groups (ranked as fourth for NS and third for NNS) when judging L2 fluency. Previous studies showed inconsistent findings on articulation rate. According to Cucchiarini et al. (2002) and Kormos and Dénes (2004), articulation rate seems to have little relationship with perceived fluency ratings. However, recent research (e.g., De Jong et al., 2013; Préfontaine et al., 2016) supported the current study by demonstrating that articulation rate played a role as a significant predictor of L2 fluency. In Préfontaine et al.'s (2016) study, which used mixed-effects modeling to examine the relationship between raters' perceptions of L2 fluency in French and temporal features, articulation rate was discovered as one of the essential factors in predicting perceived fluency. It was further supported by De Jong et al.'s (2013) study suggesting that pure speed measures such as articulation rate and mean syllable duration (inverse of articulation rate) can be claimed to reflect L2 cognitive fluency and L2 proficiency.

Thirdly, silent pause rate within a clause appears to be crucial in fluency perception. The current study utilized Kahng's (2014) silent pause rate in different locations (i.e., within a clause and at a clause boundary) since it correctly computed how frequently a

speaker pauses within or at a clause boundary. According to the correlation results, the silent pause rate within a clause was strongly negatively associated with the three rater groups. Especially among the two EFL teacher groups, the silent rate within a clause showed the second-strongest correlation with their perceived fluency (NS: r = -.786**, NNS: r = -.801**), while the peer group placed it fourth (Peer: r = -.735**). The results indicate that native and non-native teachers appear to be more attentive to silent pauses in the midst of clauses than fellow learners when rating the fluency of L2 speakers. A large body of previous studies have also established the importance of silent pause rate within a clause as a significant predictor of oral fluency (e.g., Kahng, 2014; Tavakoli, 2011), which demonstrated that L2 speech frequently contained pauses in the middle of clauses, whereas L1 speech commonly contained pauses at clause boundaries. These findings were also supported by Lennon (1990) and Towell et al.'s (1996) findings that the rate of silent pauses within a clause appeared to be a powerful predictor of L2 fluency development.

Finally, the duration of silent pauses (i.e., mean length of silent pauses) seemed to be important in determining L2 fluency for the three groups of listeners (NS: r = -.764**, NNS: r = -.751**, Peer: r = -.773**). It is worth mentioning that for all three groups, the mean length of silent pauses was statistically more connected with fluency rating than the number of silent pauses per minute (NS: r = -.635**, NNS: r = -.654**, Peer: r = -.573**). The results were consistent with the findings in the previous literature that pause

duration was more associated with L2 oral fluency than pause frequency. In Kormos and Dénes's (2004) study, for example, the correlation results showed that the mean length of pauses significantly correlated to the composite scores of native (r = -0.58) and non-native teachers (r = -0.62), whereas the number of silent pauses (NS: r = -0.10; NNS: r = -0.09) and filled pauses (NS: r = -0.08; NNS: r = -0.16) were not related to fluency scores of either raters.

Next, in order to determine the relative contribution of various utterance features and find the optimal predictive model that best explained the variance in the three groups of judges' ratings of L2 speaking fluency, stepwise multiple regression analyses were conducted using significant utterance features identified by the correlation analysis. The regression analyses revealed that three aspects of fluency (speed, breakdown, repair fluency), including the mean length of run, the mean length of silent pauses, the silent pause rate within a clause, and the number of corrections per minute, all significantly contributed to native and non-native teachers' perceptions of L2 fluency, accounting for 91.9 percent of native teachers' ratings and 94.2 percent of non-native teachers' ratings. In terms of the peer group, the final model incorporated two aspects of fluency (speed, breakdown fluency), demonstrating that the mean length of run, the mean length of silent pauses, and the articulation rate all significantly contributed to fluency assessments, accounting for 92.7 percent of the variance. For the three groups of judges, it

was discovered that one speed fluency measure: the mean length of run and one breakdown fluency measure: the mean length of silent pauses accounted for the majority of variance, suggesting the three groups of raters shared comparable final regression models. However, when the β coefficients were examined, some disparities between the teacher and peer groups were discovered.

These findings answered research question 2-2, stating that the best utterance fluency models of two EFL teacher groups incorporated all three aspects of fluency (speed, breakdown, and repair fluency), while the best model for the peer group included two aspects of fluency (speed and breakdown fluency). In addition, the mean length of run and the mean length of silent pauses, which were speed and breakdown measures, respectively, accounted for the majority of the variance in the fluency ratings.

The findings that the best utterance fluency models of the two teacher groups incorporated all three aspects of fluency (speed, breakdown, and repair fluency) were consistent with those of Bosker et al.'s (2013) investigation. They examined the contributions of three aspects of fluency (breakdown, speed, and repair fluency) to perceived fluency ratings from 80 native speaker raters and found that all three aspects contributed to fluency perception, while breakdown fluency accounted for the largest share of variance in subjective fluency ratings, followed by speed fluency. Derwing et al. (2004) and Rossiter (2009) also supported the importance of breakdown and speed fluency by demonstrating high correlations between pause and speed

measures and fluency evaluations. Additionally, the findings in the current investigation that the mean length of run was the best predictor, accounting for the most significant variance among the three rater groups, were consistent with that of Cucchiarini et al. (2002). They compared subjective ratings of fluency to objective measurements of fluency and discovered that the mean length of run was the strongest predictor of perceived fluency for intermediate learners, whereas, for beginner learners, the articulation rate was the best predictor of perceived fluency.

Finally, to discover which utterance features distinguished the three distinct perceived fluency groups (low, mid, and high) by native teacher, non-native teacher, and peer groups, a series of one-way ANOVAs were conducted, and the post-hoc comparisons of the utterance features were examined. The results indicated that, for the native teacher group, speed measures, such as the articulation rate, the mean length of run, and pause measures directly related to the speed of delivery (e.g., the number of silent pauses per minute) demonstrated the ability to distinguish high-level from mid and low level speech. Meanwhile, silent pause measures such as the mean length of silent pauses and the silent pause rate within a clause demonstrated the ability to distinguish low-level from mid and high level speakers. Similarly, the non-native teachers also employed speed features (e.g., the articulation rate) to distinguish high fluency levels from mid and low levels and silent pause measures (e.g., the mean length of silent pauses and the silent pause rate within a clause) to

separate low-level from mid and high levels in this study. The peer group, similar to the native and non-native teacher groups, demonstrated that they could distinguish high-level fluency from lower level fluency using speed measures (e.g., the articulation rate) and low-level fluency from mid and high level fluency using pause measures (e.g., the mean length of silent pauses, the silent pause rate at a clause boundary). In short, the results confirmed that breakdown fluency (e.g., the mean length of silent pauses) generally distinguished low-fluency learners from mid and high fluency learners, whereas speed measures (e.g., the articulation rate) identified high-fluency learners from mid and low fluency learners.

These findings answered research question 2-3, reporting that the three groups' phonetic correlates of their perceived fluency differed by students' fluency levels, with breakdown fluency being a relatively strong predictor of beginners' L2 fluency and speed fluency being a stronger predictor of more advanced learners' fluency.

The findings were consistent with many previous researches (e.g., Cucchiarini et al., 2000; Saito et al., 2018). Saito et al. (2018) examined whether and to what degree the listeners differentially used breakdown, speed, and repair information while assessing different levels of speech fluency (i.e., low, mid, high, and native-like levels). According to the results of a series of ANOVAs, the frequency of final-clause pauses differentiated low- and mid-level fluency performance; the number of mid-clause pauses differentiated mid- and high-level performance; and the articulation rate differentiated

high-level and native-like performance. Given that Saito et al. (2018) selected low-level learners to represent the initial stage of Japanese learners' L2 fluency development, it was acceptable to presume that a mid-level learner in their study was comparable to a low-level learner in the current study. Then, Saito et al.'s (2018) findings were congruent with the current study's. It could be presumed that native listeners distinguished low- and mid-level fluency using mid-clause pauses, equivalent to the silent pauses within a clause in the current study, and differentiated mid- and high-level fluency using the articulation rate. Furthermore, Cucchiarini et al. (2000) investigated the auditory features associated with beginner, intermediate, and advanced L2 fluency. While analyzing the acoustic characteristics of two distinct groups of speakers (beginner and intermediate), they discovered that breakdown fluency influenced their perceived fluency for the beginner group and speed fluency for the intermediate group, which was consistent with the current investigation's findings as well.

CHAPTER 5. CONCLUSION

This chapter summarizes the major findings and concludes with pedagogical implications, limitations, and suggestions for future research.

5.1 Findings and Pedagogical Implications

Although fluency is a noticeable characteristic of speech and has been identified as an essential skill to assess (Préfontaine, 2013), it has not been well understood in the EFL context. Fluency is frequently defined as the ability to speak smoothly, accurately, and confidently (Lennon, 2000) in accordance with native-speaker norms, but there is no absolutely agreed upon consensus on what fluency is and what constitutes fluency. Moreover, there is a paucity of studies concerning how raters perceive fluency differently, especially in the EFL context. The present study, thus, investigated the differences in perceived fluency by native teacher, non-native teacher, and peer groups in the Korean EFL context (Study 1) and traced the underlying mechanism in different fluency perceptions by relating perceived fluency with utterance fluency (Study 2). The major findings of the present study are summarized as follows:

First, it has been demonstrated that native and non-native

English teachers have similar rating patterns when assessing L2 fluency. In Study 1, the statistical results suggested that both EFL teacher groups had comparable severity patterns on both picture narration and spontaneous speech tasks across all proficiency levels. It was further corroborated by the following qualitative investigation. The frequency distribution analysis of written comments revealed that the native and non-native teacher groups used nearly identical fluency criteria (i.e., speech rate, number of silent pauses, vocabulary, and segmental pronunciation) when they evaluated L2 speakers' fluency. The trend that the two teacher groups had similar views on students' speaking fluency was also reinforced by Study 2. In Study 2, correlation analysis revealed that the top six utterance features associated with native and non-native teachers were identical (mean length of run, silent pause rate within a clause, mean length of silent pauses, articulation rate, number of silent pauses per minute, and number of corrections per minute). In addition, the final predictive model that best explained the variance in perceived fluency for the three rater groups revealed that the four utterance attributes and their relative contribution rankings for native and non-native teachers were identical, implying that both native and non-native teachers evaluate L2 fluency using the same acoustic qualities. Among the four utterance features, the mean length of run was found to be the most robust predictor for the two EFL teacher groups, explaining most of the variance in fluency judgments. Examining the β values of these variables further revealed that the mean length of silent

pauses was found to be the most influential factor in both native and non-native teachers' judgments of fluency, followed by the mean length of run, the silent pause rate within a clause, and the number of corrections per minute. Based on this empirical evidence, it is reasonable to assume that non-native teachers are equally capable of serving as fluency raters as native teachers in the Korean EFL context. This finding also provides a compelling argument against the notion that native speakers should be the only acceptable norm for assessing fluency in the EFL classroom.

Second, it was discovered that the peer group's ratings of L2 fluency differed from those of both EFL teacher groups. In Study 1, two-way ANOVA analyses at each proficiency level revealed that the peer group provided significantly lower fluency ratings to picture narration and spontaneous tasks than both EFL teacher groups across all proficiency levels. The frequency analysis of written comments also corroborated this tendency, revealing that peers made far fewer comments than both EFL teacher groups and that their comments were often harsher and more negative than both teacher groups. Additionally, the findings from the frequency distribution of the comments verified the distinction between the instructor groups and the peer group. Although all three groups of judges seemed to be influenced by similar perceptions of fluency criteria, peers placed the most significant importance on the top three criteria (the speech rate, the number of silent pauses, and the segmental pronunciation), accounting for more than two-thirds of total responses. Peer

participants appeared to place a high value on speech rate, silent pauses, and pronunciation of L2 speech samples while paying little attention to non-segmental pronunciation (e.g., intonation, rhythm, stress), grammar, and formulaic sequence, which were considered just as important to teacher groups. Disparities in the quality of written feedback have also been discovered between the teacher and peer groups. According to the qualitative study, the teacher groups, particularly the native teachers, provided more extensive and elaborated feedback than peer students.

In addition, Study 2 also confirmed the differences between teacher groups and the peer group, accounting for the disparities in perceived fluency found in Study 1. The correlation analysis demonstrated that speed measures, such as the mean length of run and the articulation rate, were most correlated with peer group fluency ratings. It is worth noting that, aside from the mean length of run, both native and non-native teacher groups placed a higher premium on breakdown measurements, such as the silent pause rate within a clause, than speed measures, which demonstrated a different trend from peers. Although the best regression model for the peer group was not distinct from that for EFL teacher groups, in which two utterance features (i.e., the mean length of run and the mean length of silent pauses) explained most of the variance, the peer group's model differed from those of both EFL teacher groups regarding included variables and the β coefficient of the three variables. In contrast to both EFL teacher groups, the speed fluency

measures, including the articulation rate, and the mean length of run, explained more variance in fluency ratings for the peer group. Especially, the mean length of run was found to be most strongly related to the prediction of the dependent variable. Based on these findings, it is reasonable to assume that raters' professional backgrounds influenced their fluency ratings and that peer groups' fluency rating patterns differ from teacher groups.

However, the observed disparities in peer students' grading patterns should not be used to rule out the possibility of peer students serving as fluency raters in the EFL classroom. Not only do peer students gain from peer evaluation, but peer assessment can also serve as a model for assessment centered on learning (Saito, 2008). For instance, students could be involved in designing and implementing the fluency evaluation criteria for speaking performance tests, so assisting peer students in developing a structured understanding of fluency constructs. Particularly, its participatory and negotiable process would help peer students view the peer evaluation exercise more positively, resulting in more favorable attitudes about peer performance and a more supportive learning environment (Cheng & Warren, 2005). In addition, the gap between the teacher groups and the peer group could be bridged by peer feedback in the peer assessment process. As Patri (2002) demonstrated, when assessment criteria were clarified and exemplified by the teacher, peer feedback would enable peer students to make judgments of their peers similar to those of the teachers. It appeared that having students participate

in the peer feedback session with clarifying explicit criteria for fluency assessment would help achieve a greater agreement between teacher assessment and peer assessment.

Thirdly, despite the differences between the teacher and peer groups, there was some common ground among the three rater groups for evaluating L2 fluency. The correlation results indicated that despite some ranking variations, four utterance fluency features, including the mean length of run, the articulation rate, the silent pause rate within a clause, and the mean length of silent pauses, were significantly commonly linked with the fluency judgments of the three groups of judges. Furthermore, it was discovered that the best regression models for the three different groups of judges included the same four utterance fluency features, which accounted for the majority of the variance. Based on the statistical data gathered, it is plausible to assume that these four utterance fluency features serve as essential criteria for judging L2 fluency by the three rater groups. From a pedagogical standpoint, identifying reliable phonetic correlates of fluency has not only aided in the improvement of L2 learners' oral fluency (Kahng, 2014) but has also shed light on new potential for a more valid and accurate assessment tool to measure oral fluency: automated speech evaluation. Due to the fact that these four common utterance fluency features can be extracted and calculated more easily by computers than other linguistic features such as grammar, content, and discourse organization (Zechner & Evanini, 2020), the possibility of using an automatic scoring method to rate oral fluency is

enormous, and it could be utilized as a reliable rating tool in the classroom in conjunction with a manual scoring method.

Fourth, it was shown that certain common utterance fluency features could help the three groups of judges identify three distinct perceived fluency levels (low, mid, and high). According to the statistical analyses, speed measures such as the articulation rate and the mean length of run demonstrated the ability to discriminate high-level from mid and low level fluency speech. In contrast, breakdown measures such as the mean length of silent pauses and the silent pause rate within a clause demonstrated the ability to discriminate low-level fluency from mid and high level fluency. These findings have pedagogical implications for fluency enhancement and teacher feedback in the EFL classroom. For example, the data suggested that EFL teachers need to focus on pause measures more than speed measures when teaching L2 learners who have low fluency performance. In other words, learners with low fluency levels should first be taught how to minimize the length and frequency of silent pauses while speaking and should be reminded of the importance of not pausing within a clause. In this sense, it would be helpful for novice learners to learn how to use filled pauses strategically. Based on the observation that filled pauses from L1 learners were frequently preceded or followed by silent pauses, it could be claimed that appropriate use of filled pauses is an ability that L2 learners should acquire and hence is a part of their developing L2 fluency (Kahng, 2014). With guidance from EFL

teachers, students with low fluency levels can learn how to use filled pauses in order to break up long silent pauses and make their speech less dysfluent.

However, for learners with intermediate fluency, EFL teachers should focus more on increasing their speech rate because it was a speed measure that differentiated high-level fluency performance from lower level fluency performance. One way teachers can assist intermediate L2 learners in improving their L2 fluency is to provide ample opportunities for collocation and formulaic language, which can help learners produce longer runs (Kahng, 2014). A formulaic language sequence is a good place to start developing fluency (Chambers, 1998), especially in the EFL classroom. A formulaic sequence is defined by Wray (2002) as "a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated" (p. 9) and ranges from simple fillers and functions to collocations, idioms, proverbs, and lengthy standardized phrases (Schmitt & Underwood, 2004). Many previous studies (e.g., Boers et al., 2006; Skehan, 1998) confirmed that the mastery of formulaic sequences could help learners become fluent L2 speakers as formulaic sequences were stored and retrieved whole from memory at the time of use, rather than being subject to analysis by the language grammar (Wray, 2002). As a useful pedagogical task, dictogloss could be one way to improve fluency, facilitating formulaic language units' retrieval and promoting automatization (Onoda, 2014).

Additionally, it is highly suggested to provide practical

approaches for improving oral fluency, such as the 4/3/2 fluency training (De Jong & Perfetti, 2011), in which speakers deliver three speeches under increasing time constraints (4, 3, and 2 minutes). The 4/3/2 training, which can be easily adapted in the EFL classroom, enables learners to integrate previously encountered language items into an easily accessible and largely unconscious language system, helping to automate language units (Onoda, 2014). In discussing various pedagogical tasks promoting oral fluency, other researchers suggested employing similar tasks, such as shadowing, reading aloud, and summarizing, which can help students internalize language units.

Fifth, it was revealed that three rater groups' evaluations for low-level learners' fluency were significantly affected by task types, with the spontaneous speech task scoring higher than the picture narration task. This could be due to different linguistic and cognitive loads required for certain tasks. From a pedagogical perspective, EFL teachers should be aware that completing picture narrative tasks, a popular type of speaking performance test, may be significantly more difficult for low-level students than for mid- to high-level students. Thus, not only should fluency judgments of low-level students be taken with caution, but incorporating varied tasks rather than a single task is required to capture L2 oral fluency among low-level students accurately. As Ejzenberg (2000) noted, the speaking task should provide an opportunity for L2 speakers to demonstrate many aspects of their oral fluency, and multi-task or oral portfolio assessment may have this potential.

5.2 Limitations of the Study and Suggestions for Future Research

The present study has sought to fill gaps and extend the body of research on L2 fluency evaluation by providing empirical evidence on rating differences among three groups of raters and exploring the relationship between the three groups' perceived fluency and utterance fluency measures by thoroughly analyzing utterance fluency. The study also contributed insights into appropriate research methodology, including qualitative analysis of the three rater groups' perceived fluency, where quantitative analysis is dominant. However, the results of the current dissertation should be viewed with some limitations in mind.

The first limitation of the research is the study's limited number of samples and demographic composition. While all EFL teacher listeners were native and non-native English teachers with various EFL teaching experiences, the number of evaluators was still insufficient to appropriately reflect the Korean English classroom evaluation situation. Furthermore, because the speaker participants and peer raters were all from the same high school and most were male, the sample did not adequately represent the average Korean EFL classroom environment. Thus, future research should include a more diverse range of speakers and a larger sample size of listeners, considering possible gender differences when designing methodology

and interpreting findings.

Second, only two speaking tasks were used in the L2 speech samples. Although the two speaking tasks utilized in the study were the most frequently used types for assessing speaking performance in Korean EFL classroom and necessarily portrayed the different linguistic and cognitive loads of L2 speakers, it is still unclear how other types of tasks affect L2 speakers' speech and to what extent their impact is on speakers and the three distinct rater groups. Future studies will include additional speaking tasks and investigate their influences on speakers and raters.

Lastly, as Study 1 demonstrated, the peer group's assessments of L2 fluency differ markedly from those of EFL teacher groups. Although differences in perceived fluency partially explained the disparities in utterance fluency in Study 2, it remains uncertain what accounts for the rater discrepancies between teacher groups and the peer group and what conditions must be met for peer students to rate speaking fluency equivalent to that of teachers. Thus, additional research should be conducted to determine how peer groups' fluency ratings can be justified and what has to be done in the EFL classroom to integrate peer assessment into the conventional assessment process.

REFERENCES

Audacity Team (2021). Audacity(R): Free audio editor and recorder (Version 2.4.2) [Computer software]. https://audacityteam.org/

Barnwell, D. (1989). 'Naive' native speakers and judgements of oral proficiency in Spanish. *Language Testing, 6*(2), 152-163.

Birdsong, T. P., & Sharplin, W. (1986). Peer evaluation enhances students' critical judgment. *Highway One, 9*(1), 23-28.

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. *Language Teaching Research, 10*(3), 245-261.

Boersma, P., & Weenink, D. (2019). Praat: doing phonetics by computer (Version 6.1.07) [Computer program]. https://www.fon.hum.uva.nl/praat/

Bosker, H. R., Pinget, A. F., Quené, H., Sanders, T., & De Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing, 30*(2), 159-175.

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

Caracelli, V. J., & Greene, J. C. (1993). Data analysis strategies for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis, 15*(2), 195-207.

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

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

Cheng, W., & Warren, M. (2005). Peer assessment of language proficiency. *Language Testing, 22*(1), 93-121.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York: Routledge Academic.

Crowther, D., Trofimovich, P., Isaacs, T., & Saito, K. (2015). Does a speaking task affect second language comprehensibility?. *The Modern Language Journal, 99*(1), 80-95.

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*(6), 2862-2873.

De Jong, N., & 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., & Bosker, H. R. (2013). Choosing a threshold for silent pauses to measure second language fluency. In R. Eklund (Ed.), *Proceedings of the 6th Workshop on Disfluency in Spontaneous Speech (DiSS)* (pp. 17-20).

De Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. H. (2013). Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics, 34*(5), 893-916.

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

237-254.

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition, 19*(1), 1-16.

Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning, 54*(4), 655-679.

Derwing, T. M., Munro, M. J., & Thomson, R. I. (2008). A longitudinal study of ESL learners' fluency and comprehensibility development. *Applied Linguistics, 29*(3), 359-380.

Derwing, T. M., Munro, M. J., Thomson, R. I., & Rossiter, M. J. (2009). The relationship between L1 fluency and L2 fluency development. *Studies in Second Language Acquisition, 31*(4), 533-557.

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

Ellis, R. (2003). *Task-based language learning and teaching*. London: Oxford university press.

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

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: Academic Press.

Foster, P., & Skehan, P. (1996). The influence of planning and task type

on second language performance. *Studies in Second Language Acquisition, 18*(3), 299-323.

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

Freeman, M. (1995). Peer assessment by groups of group work. *Assessment & Evaluation in Higher Education, 20*(3), 289-300.

French, L. M., Gagné, N., & Collins, L. (2020). Long-term effects of intensive instruction on fluency, comprehensibility and accentedness. *Journal of Second Language Pronunciation, 6*(3), 380-401.

Galloway, V. B. (1980). Perceptions of the communicative efforts of American students of Spanish. *Modern Language Journal, 64*(4), 428-433.

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

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis, 11*(3), 255-274.

Gui, M. (2012). Exploring differences between Chinese and American EFL teachers' evaluations of speech performance. *Language Assessment Quarterly, 9*(2), 186-203.

Hadden, B. L. (1991). Teacher and nonteacher perceptions of second

language communication. *Language Learning, 41*(1), 1-20.

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

Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy and fluency. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 1-20). Philadelphia: John Benjamins.

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.

Kachru, B. B. (1985). Standards, codification and sociolinguistic realism: The English language in the outer circle. In R. Quirk & H. Widdowson (Eds.), *English in the world: Teaching and learning the language and literatures* (pp. 11-30). Cambridge: Cambridge University Press.

Kahng, J. (2012). *How long should a pause be? Effects of cut-off points of pause length on analyzing L2 utterance fluency*. Poster presented at Fluent Speech Workshop, Utrecht, The Netherlands.

Kahng, J. (2014). *Exploring the production and perception of second language fluency: Utterance, cognitive, and perceived fluency.* Unpublished doctoral dissertation, Michigan State University.

Kahng, J. (2022). Fluency. In T. M. Derwing, M. J. Munro, & R. I. Thomson (Eds.), *The routledge handbook on second language acquisition and speaking.* (pp. 188-200). New York: Routledge.

Koponen, M., & Riggenbach, H. (2000). Overview: Varying perspectives

on fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 5-24). Ann Arbor: University of Michigan.

Kim, Y. H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing, 26*(2), 187-217.

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

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

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning, 40*(3), 387-417.

Levis, J. (2020). Changes in L2 pronunciation: 25 years of intelligibility, comprehensibility, and accentedness. *Journal of Second Language Pronunciation, 6*(3), 277-282.

Lim, I., & Hwang, J. (2019). Korean adult English learners' perceptions of the common grammatical features of English as a lingua franca. *Journal of Asia TEFL, 16*(3), 876-893.

McNamara, T. F., & Macnamara, T. J. (1996). *Measuring second language performance*. New York: Longman Publishing Group.

Meyers, L. S., Gamst, G., & Guarino, A. J. (2016). *Applied multivariate research: Design and interpretation (3rd ed.)*. Thousand Oaks: Sage publications.

Ministry of Education. (2015). *English curriculum(supplement book #14) in General guideline of the national curriculum for the*

primary and secondary schools. Seoul: Ministry of Education.

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 45*(1), 73-97.

Munro, M. J., & Derwing, T. M. (1999). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 49*, 285-310.

Munro, M. J., & Derwing, T. M. (2020). Collecting data in L2 pronunciation research. In O. Kang, S. Staples, K. Yaw, & K. Hirschi(Eds.), *Proceedings of the 11th Pronunciation in Second Language Learning and Teaching Conference* (pp. 8 - 18). Flagstaff, Arizona: Northern Arizona University.

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*(4), 557-581.

Onoda, S. (2014). An exploration of effective teaching approaches for enhancing the oral fluency of EFL students. In T. Muller, J. Adamson, & P. S. Brown (Eds.), *Exploring EFL fluency in Asia* (pp. 120-142). London: Palgrave Macmillan.

Patri, M. (2002). The influence of peer feedback on self-and peer-assessment of oral skills. *Language Testing, 19*(2), 109-131.

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

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction (3rd ed.)*. Fort Worth, TX: Harcourt Brace.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878-912.

Préfontaine, Y., Kormos, J., & Johnson, D. E. (2016). How do utterance measures predict raters' perceptions of fluency in French as a second language?. *Language Testing, 33*(1), 53-73.

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

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

Saito, H. (2008). EFL classroom peer assessment: Training effects on rating and commenting. *Language Testing, 25*(4), 553-581.

Saito, K., Ilkan, M., Magne, V., Tran, M. N., & Suzuki, S. (2018). Acoustic characteristics and learner profiles of low-, mid-and high-level second language fluency. *Applied Psycholinguistics*, 39(3), 593-617.

Schmitt, N., & Underwood, G. (2004). Exploring the processing of formulaic sequences through a self-paced reading task. In N Schmitt (Ed.), *Formulaic sequences: Acquisition, processing and use* (pp. 173-189). Philadelphia: John Benjamins.

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

York: Routledge.

Shi, L. (2001). Native-and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing, 18*(3), 303-325.

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

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

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.

SyllableCounter.net. (n.d.). Syllable counter [Online software]. https://syllablecounter.net/

Tavakoli, P. (2011). Pausing patterns: Differences between L2 learners and native speakers. *ELT Journal, 65*(1), 71-79.

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.

Tavakoli, P., & Uchihara, T. (2020). To what extent are multiword sequences associated with oral fluency?. *Language Learning, 70*(2), 506-547.

Taylor, L. (2006). The changing landscape of English: Implications for language assessment. *ELT Journal, 60*(1), 51-60.

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

17(1), 84-119.

Wood, D. (2010). *Formulaic language and second language speech fluency: Background, evidence and classroom applications*. New York: Continuum International Publishing Group.

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

Yu, K. A. (2010). The effect of raters' language background on English-speaking test ratings across test-takers' oral proficiency levels. *Korea Journal of Applied Linguistics, 26*(4), 395-419.

Zechner, K., & Evanini, K. (Eds.). (2020). *Automated speaking assessment: Using language technologies to score spontaneous speech*. New York: Routledge.

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.

APPENDICES

Appendix A. Summary of Utterance Fluency Models in Literature Review

	Participants & Level	Rater	Task	Utterance Fluency Measures	Findings
Cucchiarini et al. (2002)	60 NNS (30 beginner level, 30 intermediate level)	10 NS	spontaneous speech in response to short and long prompts	Articulation rate, Speech rate, Phonation time ratio, Mean length of run, Mean length of silent pauses, Duration of silent pauses per minute, Number of silent pauses	Speech rate (for the beginner level) and mean length of run (for the intermediate level) were the best predictors, the number of silent pauses explained minimal variance
Bosker et al. (2013)	38 NNS (Intermediate to advanced)	80 NS	descriptive, simple and formal argumentative tasks	Number of silent pauses, Number of filled pauses, Mean length of pauses, Mean length of syllables, Number of repetitions, Number of corrections	Breakdown fluency measures explained the most of the variance in fluency ratings, while repair fluency measures were thought to have insufficient predictive capacity.
Kahng (2014)	37 NNS 6 NS	46 NS	spontaneous speech (two interview questions)	The pause frequency, length, and distribution	The pause distribution accounted for almost 45% of the variance and the pause duration added an extra 4%.
Saito et al. (2018)	90 NNS 10 NS	10 NS	timed picture description task	Articulation rate, Final-clause pause ratio, Mid-clause pause ratio, repetition ratio, Self-correction ratio	Articulation rate, Mid-clause pauses, and Final-clause pauses accounted for 57% of the variance

Appendix B. Picture Narration Task

THE SUITCASE STORY

The "Suitcase Story" may be used for research purposes only, provided that the user cites the following source:

Derwing, T. M., Munro, M. J., Thomson, R. I., & Rossiter, M. J. (2009). The relationship between L1 fluency and L2 fluency development. Studies in Second Language Acquisition, 31(4), 533-557.

Appendix C. Spontaneous Speech Task

Answer to the question below.

Q. What do you want to be in the future?

Appendix D. Demographic Questionnaire

IRB No. 2102/003-010 유효기간: 2022년 02월 16일

사전 설문지

Student Number | | | | |

1. How much English do you speak?

0	1	2	3	4
Not Fluent in English	Limited Fluency in English	Somewhat Fluent in English	Quite Fluent in English	Very Fluent in English
No understanding or speaking ability	Some understanding and can say short, simple sentences e.g. can answer the phone in English	Good understanding and can express myself on many topics e.g. can go to the doctor and explain what is wrong	Can understand and use English adequately for work and most other situations e.g. can communicate effectively with teachers; can follow movies or television shows	Understand almost everything. Very comfortable expressing myself in English in all situations

2. How do you feel when you have to use English?

	strongly disagree	disagree	slightly disagree	slightly agree	agree	strongly agree
I can **READ** and understand English text confidently without any serious difficulty.						
I can **WRITE** in English to deliver my message confidently without any serious difficulty.						
I can **LISTEN** and understand English speech confidently without any serious difficulty.						
I can **SPEAK** in English to deliver my message confidently without any serious difficulty.						
I feel unsure and worried when I have to **READ** English text.						
I feel unsure and worried when I have to **WRITE** in English.						
I feel unsure and worried when I have to **LISTEN** to English speech.						
I feel unsure and worried when I have to **SPEAK** in English.						

3. Which do you think is the most important (so you should care the most about) when you intend to improve your English speaking? (1: the most important / 6 the least important)

____ natural flow (less pause or correction)
____ native-like accent and pronunciation
____ using proper words
____ speaking correctly using proper grammar
____ getting the meaning across (even if roughly)
____ using higher-level vocabulary and structure

4. What is the biggest stumbling block in your English speaking?

_____

_____

_____

_____

5. Why do you think Speaking English fluently is important?

_____

_____

_____

_____

Thank you!

Appendix E. Instruction for Native and Non-native Teacher Groups

Part 1 - Instruction

- You will hear 30 Korean second language learners of English speaking on a narrative description task. Students were asked to narrate eight-frame pictures (below) making a single story.

- All of the speech samples are about 30-second long, excerpted from a 1-minute long audio file. I would like you to judge speaking fluency about each sample. (Please listen to the audio to the end.)

- Speaking fluency: the flow of the language – does the speaker have problems finding words, hesitating and pausing often, or do the words come quickly? So a speaker who is very fluent – that is, the words just flow with no struggle, would be at the top of the scale(9), while someone who has a hard time expressing him or herself would be closer to the right end of the scale(1).

- 2 samples will be given before starting the survey.

Part 2 - Instruction

- You will hear 30 Korean second language learners of English speaking on a spontaneous speaking task. Students were asked to answer one following question

- Question) What do you want to be in the future?

- All of the speech samples are about 30-second long, excerpted from a 1-minute long audio file. I would like you to judge speaking fluency about each sample. (Please listen to the audio to the end.)

- Speaking fluency: the flow of the language – does the speaker have problems finding words, hesitating and pausing often, or do the words come quickly? So a speaker who is very fluent – that is, the words just flow with no struggle, would be at the top of the scale(9), while someone who has a hard time expressing him or herself would be closer to the right end of the scale(1).

- 2 samples will be given before starting the survey.

Appendix F. Instruction for the Peer Group

Part 1 - Instruction

- 여러분은 앞으로 30명의 한국인 학생의 그림 묘사 과제에 대한 음성 파일을 들을 것입니다. 학생들은 아래의 8개 그림을 연결하여 하나의 이야기를 만드는 과제를 수행하였습니다.

- 모든 음성 파일은 30초 내외입니다. 여러분은 이 음성 파일을 끝까지 듣고 화자(speaker)의 유창성(fluency)을 평가하면 됩니다.

- 유창성은 말의 흐름(flow)과 관련이 있습니다. 화자(speaker)가 적당한 말을 하는데 어려움을 느끼며 말이 느려지는지, 말을 할 때 망설이거나 휴지 부분(pause)이 많은지를 보면 됩니다. 그래서 단어를 생각하거나 말을 하는데 전혀 망설임이 없는 가장 유창한 화자는 9점, 스스로 말을 하는데 많은 어려움을 겪는 비유창한 화자는 1점을 주면 됩니다.

- 본격적으로 설문에 앞서 두 개의 예시를 살펴보겠습니다.

Part 2 - Instruction

- 여러분은 앞으로 30명의 한국인 학생의 음성 파일을 들을 것입니다. 학생들은 아래의 인터뷰 질문에 대답하는 과제를 수행하였습니다.

- 인터뷰 질문: What do you want to be in the future?

- 모든 음성 파일은 30초 내외입니다. 여러분은 이 음성 파일을 끝까지 듣고 화자(speaker)의 유창성을 평가하면 됩니다.

- 유창성은 말의 흐름(flow)과 관련이 있습니다. 화자(speaker)가 적당한 말을 하는데 어려움을 느끼며 말이 느려지는지, 말을 할 때 망설이거나 휴지 부분(pause)이 많은지를 보면 됩니다. 그래서 단어를 생각하거나 말을 하는데 전혀 망설임이 없는 가장 유창한 화자는 9점, 스스로 말을 하는데 많은 어려움을 겪는 비유창한 화자는 1점을 주면 됩니다.

- 본격적으로 설문에 앞서 두 개의 예시를 살펴보겠습니다.

국 문 초 록

유창성은 제2언어 수행과 언어능력에서 매우 중요한 부분을 구성하고 있으며, 많은 제2언어 학습자가 유창성을 획득하고자 노력하고 있다. 그러나 제2언어 연구자나 EFL(외국어로서의 영어) 교육자들은 유창성에 대해 명확히 이해하지 못하고 있으며, 그 개념 또한 일관성 있게 정의하지 못하고 있다. 게다가 EFL 환경에서 다양한 평가자들이 어떻게 유창성을 인식하고 평가하는지에 대한 연구는 매우 부족하다. 이러한 학문적인 필요성을 충족하고, 유창성의 다차원적 구성에 대한 깊은 이해를 위해 본 연구는 원어민 교사, 비원어민 교사, 그리고 동료 학생들이 한국 고등학생의 영어 말하기 유창성을 어떻게 인식하고 평가하는지를 인식 유창성과 발화 유창성이라는 두 가지 관점에서 연구하였다.

연구 1은 혼합 연구 방법을 이용하여 세 평가 집단의 인식 유창성이 어떻게 다른지를 비교한 것이다. 세 집단의 평가자는 다른 말하기 능력(상, 중, 하)을 가진 발화자가 두 과제 유형(그림 이야기, 자유 발화)에 대해 수행한 샘플을 듣고 유창성을 평가하였고, 연구자는 세 집단의 전반적인 유창성 점수는 양적인 방법으로, 채점자의 서면 평가는 질적인 방법으로 분석하였다. 두 과제와 세 언어 능력 집단 모두에게서 원어민과 비원어민 교사 집단은 비슷한 수준의 엄격성 패턴을 보였지만 동료 학생 집단은 교사 집단에 비해 유의미하게 낮은 점수를 주었음이 드러났다. 이어진 질적 분석은 EFL 교사 집단과 동료 학생 집단의 차이를 다시 한번 확인시켜주었다. 그리고 세 평가 집단이 하 수준의 학생들을 평가할 때, 그림 이야기에는 낮은 점수를 준 반면 자유 발화에는 높은 점수를 주는 경향을 보여, 모든 평가 집단이 과제 유형에 상당한 영향을 받음이 드러났다.

연구 1에서 발견된 세 그룹의 유창성 인식의 차이는 연구 2에서 좀 더 뒷받침된다. 연구 2에서는 발화 유창성의 특성 중 어떤 요소나 어떤 음향 모델이 평가자들의 인식 유창성을 가장 잘 예측할 수 있는지, 또한 어떤 발화 유창성 특성이 평가자로 하여금 발화자의 유창성 수준을 판단하는 데 영향을 미치는지를 알아보기 위해 발화 유창성과 인식 유창성 간의 관계를 분석하였다. 그 결과 두 가지 속도와 관련된 특질(평균 발화 길이, 발화 속도)과 휴지와 관련된 특질(절 내 무음 휴지기 비율, 평균 무음 휴지 길이)이 세 평가 집단의 인식 유창성과 가장 큰 상관이 있는 것으로 나타났다. 회귀 분석 결과 평균 발화 길이와 평균 무음 휴지 길이가 회귀 모형의 가장 많은 변화량을 설명하며 세 평가 집단의 유창성 점수를 가장 잘 예측하는 것으로 드러났다. 하지만 원어민과 비원어민 교사 집단의 모델은 입력 변수와 변수의 상대적인 순위 측면에서 완전히 동일한 반면 동료 학생 집단의 회귀 모형은 교사 집단의 모형과 차이가 있음이 발견되었다. 그리고 무음 휴지의 평균 길이와 같은 휴지와 관련된 특질은 하 수준의 유창성을 가진 학생을 구분하는 데 사용되고, 발화 속도와 같은 속도 관련 특질은 상 수준의 유창성을 가진 학생을 구분하는 데 사용되고 있음이 밝혀졌다.

이러한 연구 결과들은 원어민 교사와 비교하여 유창성 평가자로서 비원어민 교사들의 역할에 대한 논의에 바탕이 되었고, 동료 학생 평가의 타당성과 신뢰성을 논의하는 토대를 만들었다. 연구 1과 연구 2에서 드러난 여러 실증적인 증거들을 바탕으로 원어민과 비원어민 교사는 모두 제2언어로서 영어 말하기 유창성을 비슷한 방식으로 인식하고 평가하는 것으로 결론 지을 수 있고, 비원어민 교사 평가자도 원어민 교사 평가자와 같이 유창성 평가에 동일한 평가자로 기능할 수 있음을 보여주었다. 하지만 동료 학생 집단은 교사 집단에 비해 다른 평가 양상을 보

여주었고, 이는 이들이 한국의 EFL 환경에서 능숙한 유창성 평가자로서의 역할을 수행하기 위해서는 많은 교육적인 노력이 필요함을 시사한다.

현 논문은 다양한 평가 집단이 학생들의 영어 말하기 유창성을 어떻게 인식하고 평가하는지를 체계적으로 분석하며, 한국 EFL 환경에서 타당하고 신뢰성이 있는 유창성 평가 방식을 수립하는데 기여한다. 게다가 본 연구는 동료 학생들의 평가를 교사 집단의 평가와 비교하여 연구함으로서 동료 평가의 가능성과 한계에 대한 직접적인 증거를 제시하였다. 또한 연구 방법론과 관련하여 본 연구는 유창성의 두 가지 측면인 인식 유창성과 발화 유창성을 결합함으로서 유창성의 다차원적 구성을 이해하는데 도움을 주고, 양적인 방법론과 질적인 방법론을 결합하여 다양한 평가 집단의 평가 패턴을 종합적으로 이해하는데 기여한다.

핵심어 : 영어 말하기 유창성, 유창성 평가, 인식 유창성, 발화 유창성, 채점자 변동성, 동료 평가

학번 : 2015-30489