교육학석사학위논문

Korean English Teachers' Decision Making Processes in Scoring Middle School Students' English Essays

한국인 영어교사들의 중학생 영어 작문 채점 시 보이는 의사결정 과정 연구

2020년 2월

서울대학교 대학원

외국어교육과 영어전공

강 민 희

Korean English Teachers' Decision Making Processes in Scoring Middle School Students' English Essays

by

Min Hee Kang

A Thesis Submitted to
the Department of Foreign Language Education
in Partial Fulfillment of the Requirements
for the Degree of Master of Arts in Education

At the
Graduate School of Seoul National University

February 2020

Korean English Teachers' Decision Making Processes in Scoring Middle School Students' English Essays

한국인 영어교사들의 중학생 영어 작문 채점 시 보이는 의사결정 과정 연구

지도교수 소 영 순

이 논문을 교육학 석사학위 논문으로 제출함

2019년 12월

서울대학교 대학원

외국어교육과 영어전공

강 민 희

강민희의 석사학위논문을 인준함

2020년 1월

위 원 장 이 경 민

부위원장 최 정 은

위 원 소 영 순

Korean English Teachers' Decision Making Processes in Scoring Middle School Students' English Essays

Approved by Thesis Committee:

Byungmin Lee, Committee Chair

Jungeun Choi

Youngsoon So

Abstract

Korean English Teachers' Decision Making Processes in Scoring Middle School Students' English Essays

Min Hee Kang

English Major, Dept. of Foreign Language Education

The Graduate School of Seoul National University

The primary goal of the present study is to examine Korean English teachers' decision-making processes when scoring middle school students' English essays. A total of nine Korean English teachers participated in this study and was asked to rate ten essays written on a scale of 10. During the scoring processes, they produced concurrent think-aloud protocols and the collected data was transcribed in full and then analyzed qualitatively.

The analysis of the verbal reports revealed the following results. First, most of the raters were more oriented to judging grammatical accuracy than content when making scoring decisions. Second, the raters with longer teaching experiences were more likely to exhibit the characteristics of expert raters' scoring behaviors, which were "creating a scoring rubric," "showing

personal engagement with the writer," and "displaying initial scanning of the whole essay." In this line of reasoning, it was assumed that the length of teaching experience could be one of the major factors which contribute to the individual differences in scoring processes. Also, based on the analysis, Korean English teachers' 12 key scoring behaviors were formulated and these were categorized by three strategies, judgement, macro-judgment and interpretation strategies. Also, judgement and interpretation strategies were further divided into three focuses, language, organization, and content focuses.

In the end, this study will shed light on variability in EFL raters' cognitive processes in scoring essays as well as common strategies found in the variability. Moreover, this is expected to help rater trainers to develop fair scoring protocols and settle down more reliable writing assessment systems in Korean school environments.

Key Words: writing assessment, decision-making processes, cognitive processes, individual differences in writing assessment, scoring behavior, Korean English teachers

Student Number: 2017-23606

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

The current research investigates Korean English teachers' decision-making processes while scoring students' English essays. This introductory chapter provides the background, purpose, and outline of the study. Section 1.1 discusses the factors that motivated this study. Section 1.2 identifies the research questions and Section 1.3 ends the chapter by outlining the overall organization of the thesis.

1.1 The Motivation of the Study

Ever since enhancing communicative competence has become the foci of English education, second language assessment has undergone rapid changes (J. Kim, 2017). Traditional language assessments, which involved pencil-and-paper tests and multiple-choice questions, are replaced by performance-based assessments (McNamara, 1996).

Language performance assessments, especially measuring oral and writing proficiency, have gained much attention from the scholars as it necessarily involves human raters (Beck, Llosa, Black, & Giese, 2015; Sakyi, 2000). Schaefer (2008) claimed that "essay rating is a complex and error-prone cognitive process which introduces systematic variances in

performance ratings" (p. 466). This rater variability questions what is really involved in rater's decision-making processes.

Some scholars argued that there was only insufficient knowledge about the decision making or criteria which raters actually used to perform such evaluations (Cumming, 1990). In the similar vein, Barkaoui (2010) argued that despite the influences of a variety of factors on the rater's decision-making processes, the previous studies have mainly focused on such aspects as task requirements or essay-related features.

Still to date, the exact nature of the constructs that the raters assess remains uncertain and the rating scales used to measure the participants' proficiency are yet too imprecise and broadly defined (Cumming, Kantor, & Powers, 2002). To address the issues about scoring processes and how rating scales are interpreted by the human raters, many researchers (Crisp, 2008, 2012; Cumming et al., 2002; McNamara, 1996; Milanovic, Saville, & Shuhong, 1996; Vaughan, 1991) have sought to explore rater cognition.

Bejar (2012, p. 4) defined rater cognition as "the systematic studies of different factors that could affect the scoring process." Largely, those different factors could be categorized in two sides: rater-internal and rater-external factors. More specifically, Eckes (2008) reported that rater-internal

factors include the following features: the degree to which rater comply with the scoring rubric, the way raters interpret criteria, the degree of severity or leniency, the understanding and use of rating scale categories, and the degree of consistent application of rating criteria. On the other hand, rater-external factors encompass such features as rating scales or test prompts (McNarama, 1996). Schaefer (2008) argued that rater-external factors could be reduced with the test developers' efforts, but rater-internal factors were more likely to persist even after extensive rater training.

Ultimately, for the valid use of performance test results, more thorough examinations into rater cognition are required. Especially it is on high demand in the Korean EFL context. S. Kim (1999) argued that studies of assessing essays written in the first language have been conducted well and deep enough, but that of assessing essays written in the second/foreign language has a short history despite its complicatedness and necessity. Similarly, Cumming et al. (2002) figured out that the groups of raters differed in their evaluation of writing according to their cultural backgrounds and the genres of writings being assessed.

In this regard, this study aims to investigate Korean English teachers' rating processes by looking into their scoring behaviors manifested during

scoring processes. The qualitative analysis of the verbal reports will specify the writing constructs the Korean English teachers more attended to and unveil the detailed scoring processes they went through in arriving at the final decisions. This qualitative study is expected to benefit the researchers who seek to have a better understanding of Korean English teachers' writing scoring procedures and help rater trainers to provide more effective guidance for teachers.

1.2 Research Question

The specific research question that guides the current study is as follows.

1. What strategies do the Korean English teachers often use when making scoring decisions?

It will be investigated by qualitatively analyzing nine EFL teachers' verbal protocols that were collected during the scoring processes. This research question will provide opportunities to understand rating processes and how the scoring decisions are made.

1.3 Organization of the Thesis

The current thesis consisted of five chapters. The present chapter, Chapter 1, explains the factors that motivated the researcher to explore this subject and presents the research question it aims to address. Chapter 2 introduces the theoretical frameworks that shaped this study, that is, the precedent studies regarding the rater cognition and their scoring behaviors. Chapter 3 describes the research methodology implemented in this study, presenting the participant profiles, and data collection and data analysis procedures. Chapter 4 displays the results and discussion points of the study. Chapter 5 concludes the thesis by summarizing the major findings and discussing pedagogical implications.

Chapter 2. Review of Literature

In this chapter, literature related to the present study is reviewed. Section 2.1 introduces and reviews the previous research on rater cognition and the related issues. Then, Section 2.2 explains factors influencing rater variability. Section 2.3 is going to delve into EFL raters' scoring behaviors, and lastly the gaps in the literature will be discussed in Section 2.4.

2.1 Rater Cognition and the Related Issues

The entire cognitive process that raters go through in the course of assessment is termed as "rater cognition" and Eckes, T., Muller-Karabil, A., and Zimmerman, S. (2016) defined it as "the attentional, perceptual, and judgemental processes involved when raters award scores to examinees" (p. 156). The study of rater cognition is necessitated as it can explain the possible sources of rater variability and be the grounds for validation arguments for writing assessments.

Some researchers pointed out the significance of studying "validity" in the use of performance-based assessment. Connor-Linton (1995) reported that "writing assessment research focused on improving the reliability of rating scales and procedures, potentially at the cost of the ratings' validity" (p.

762). Lumley (2005) made similar points. He argued like the following.

> "while the pursuit of reliability remains an essential consideration researchers have also pointed out, over a period of time, how the validity of performance assessment has been insufficiently addressed […]. Questioning the role of reliability was an insufficient condition for validity in the context of performance assessment" (p. 247).

As pointed out by the above scholars, obtaining high levels of reliability does not guarantee the validity of the assessments. In order to comprehend what the ratings really mean and make use of the results in a reasonable way, it is required to understand what the raters are actually doing and why they are doing it during writing assessments. In the ensuing sections, the studies of rater cognition will be explored for the aim of addressing the issues related to validity.

2.1.1. Rater Cognition Model

A number of research have tried to make a prototypical model for how raters come to make scoring decisions. This study introduces three major rater cognition models (Freedman & Calfee, 1983; Sakyi, 2000; Wolfe, 1997).

First, Freedman and Calfee's (1983) "information-processing model of holistic raters" explains how raters articulate evaluative decisions of written compositions.

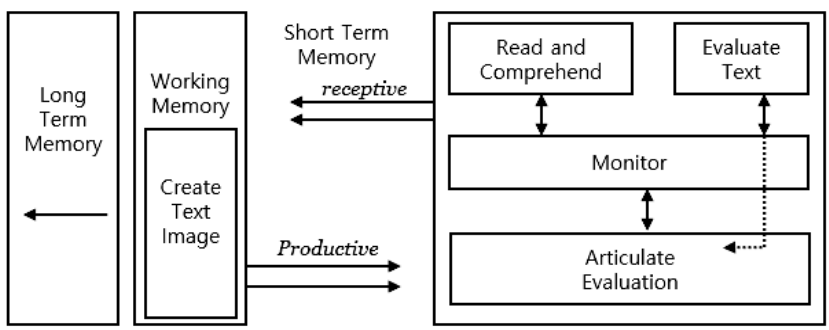


Figure 2.1. An Information-Processing Model of Rating a Composition (Freedman & Calfee, 1983, cited in Sakyi, 2000, p. 131)

In information-processing model, raters come to create text images in working memory by receiving information made by the interactions of reading, evaluating, and monitoring processes. The constructed text images are later stored in long-term memory and used as the basis for articulating evaluative decisions. According to Knoch (2009), the most important aspect of this model is that the raters make scoring decisions based on constructed text images, not actual texts. In other words, each rater can arrive at different evaluative decisions about the same texts based on how they perceive and comprehend them.

Yet, Barkaoui (2010) criticized this model for its linear sequences of scoring processes where the raters read the essays, form a mental representation of it, and then articulate a rating decision. He refuted that the scoring sequences exhibit rather recursive patterns where the raters go back and forth each stage.

Second, Wolfe (1997) sophisticated Freedman and Calfee's (1983) model by suggesting two features: a framework of scoring and a framework of writing. Figure 2.2 illustrates this idea.

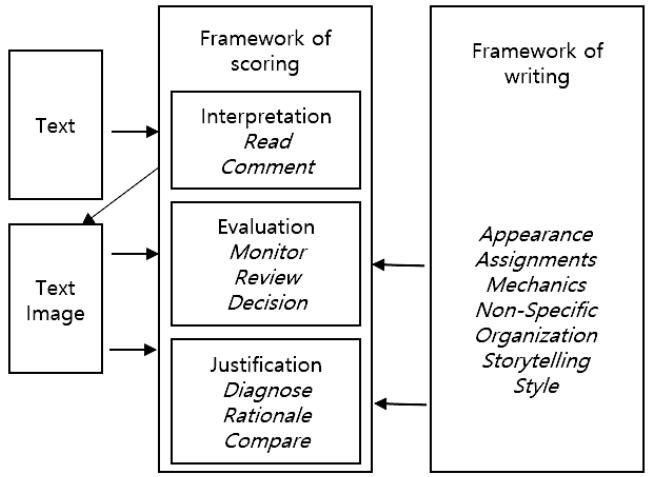


Figure 2.2. Wolfe's Model of Scorer Cognition (1997, p. 89)

The framework of writing is regarded with how raters perceive essay-related features such as appearance, organization, and storytelling of the essay. This, together with texts, affects the framework of scoring, which refers to raters' processing actions such as interpretation, evaluation, and justification processes. By going through the framework of scoring, raters get to create text images independently, which are later used as the basis for generating scoring decisions. The principal idea of making evaluative decisions about the essays not far from what was proposed in Freedman and Calfee's (1983) information-processing model.

What should be noted here is that framework of scoring occurs based on raters' perception of framework of writing. In other words, how the raters interpret and evaluate given essays are influenced by how they perceive the levels of the essays. Also, the created text images affect back to framework of scoring. This is in line with Barkaoui's (2010) claim, which is scoring processes exhibit recursive patterns where raters go back and forth each stage.

Lastly, Sakyi (2000) contended that still there is only limited information and evidence about how raters decide the levels of learners' writing ability and suggested an alternative rater cognition model. Based on the collected verbal protocols from six experienced raters, he suggested "a tentative model of holistic scoring process," as illustrated in Figure 2.3.

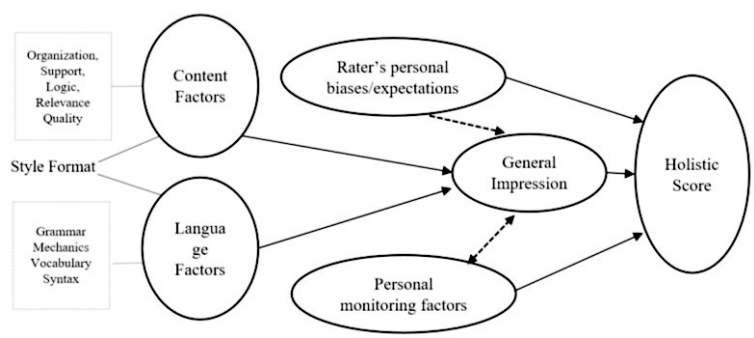


Figure 2.3 A Tentative Model Showing Factors Affecting Holistic Scores of Written Compositions (Sakyi, 2000, p. 146)

To illustrate, raters mostly attend to content-related and language-related factors during writing assessment. These factors interact each other, creating a general impression of a particular essay. The general impression is also influenced by raters' personal biases, expectations, and monitoring factors. All the features in circles play parts in arriving at holistic scores.

The three studies commonly support the idea that individual raters comprehend given texts differently and articulate scoring decisions based on their perception. Among these three models, this study mainly adopts Wolfe (1997)'s model of scorer cognition as it can explain the aspects of writing raters more attended to and how it affects scoring decisions.

2.2 Factors Influencing Rater Variability

The above section discussed the general cognitive patterns raters exhibit during scoring processes. This section is going to examine the probable factors that deviate raters from the desirable paths. Eckes (2008) defined rater variability as "the variability that is associated with characteristics of the raters, but not with the performance of examinees" (p. 155) and rater variability tended to persist even after extensive rater training. Cho (2014) claimed that figuring out the sources of rater variability and

reducing them are required for better writing assessment procedures.

Largely, there are two sources of rater variability: rater-internal and rater-external factors. As this study aims to investigate the effects of rater-internal factors, this will be the main focus of the discussion in Section 2.2.1.

2.2.1 Rater-Internal Factors

The precedent studies have reported that reading styles (Crisp, 2012; Milanovic et al., 1996; Vaughan, 1991), rater bias (Eckes, 2008; Engelhard, 1994), and the amount of writing assessment experiences (Cumming, 1990; Huot, 1993; Wolfe, Kao, & Ranny, 1998) are the three main sources of rater variability. In the following sections, each factor is going to be described in depth.

2.2.1.1 Rater's Reading Styles

Reading is the most fundamental processes that occur during writing assessment and thus, reading behaviors are directly associated with how raters evaluate the essays (Barkaoui, 2010; Crisp, 2012; Vaughan, 1991). This current research is going to suggest three perspectives toward raters' reading styles.

First, Vaughan (1991) explored rater cognition by analyzing think-aloud protocols of six experienced native English and EFL raters and proposed five reading styles as in Table 2.1.

Table 2.1

Five Reading Styles (Vaughan, 1991, p. 118)

1. The single-focus approach: To approach writing assessment with the rater's own existing methods
2. The first impression dominating approach: The first positive or negative impression that the marker received influences the whole marking session
3. The two-category strategies: To focus on two constructs such as content and grammar or content and organization
4. The laughing rater: To consider the writer is immature or non-native and have a negative impression on the essay
5. The grammar-oriented rater: To focus on mostly on grammatical features on the essay

The configuration of five reading styles is expected help categorize rater types and figure out their scoring tendencies.

Similarly, Milanovic et al. (1996) suggested four reading-marking styles: the principled two-scan/read, the pragmatic two-scan/read, read through, and the provisional mark approach. First, "the principled two-scan/read" refers to an initial scanning of composition for length and appearance followed by the second reading to confirm if first judgement is correct. Second, "the pragmatic two-scan/read" is to read the scripts twice

before assigning a mark to the script. Shaw and Weir (2007) noted that this reading style occurs only when the marker encountered difficulties in the script or in the marking environment and had to re-read to determine a mark. Third, “read through” is reading through a script to pick up its good and bad points. Lastly, “the provisional mark approach” is characterized by a single reading of a script, but with a break in the marking flow. This is done to discover whether the rest of the script confirms or denies the initial assessment of merits. The raters opt to utilize particular reading styles based on their prior experiences.

On the other hand, Wiseman (2012) put forward somewhat different perspectives about raters’ reading styles. Unlike the above two researchers who categorized reading styles and presented its relations with the marking strategies, he argued the impact of “rater as reader” on judgments of essays. This is to say, raters of essays written in the second language are readers and consistently making expectations about the essay. He added that considering the nature of reading process, raters’ expectations are as important as the quality of text itself. This reading behavior is also referred to as “the personalization of the examinee,” “sympathetic reading of the essay,” or “rater involvement or engagement with the essay.” According to Huot (1993),

"rater as reader" is typically exhibited by expert raters.

2.2.1.2 Rater Bias

"Rater bias" is also considered to attribute to rater variability (Cho, 2014; Eckes, 2008; Engelhard, 1994; McNamara & Lumley, 1995). Engelhard (1994) defined rater bias as "the tendency on the part of raters to consistently provide ratings that are lower or higher than warranted by student performances" (p. 98). Also, according to McNamara and Lumley (1995), rater bias is to be harsh on certain constructs perceived more important and lenient for the constructs perceived less important. Wigglesworth (1993) figured out that some raters were especially harsher when rating such features as grammar, fluency, or vocabulary.

Schaefer (2008) pointed out that the studies of bias analysis can increase inter-rater reliability. That is, the understanding of each rater's unique scoring patterns enhances rater-consistency by helping the raters to use the rating scales in a more consistent fashion. Ultimately, this is expected to increase fairness, accuracy, and the validity of assessments.

2.2.1.3. The Amount of Writing Assessment Experiences

Many researchers found that the amount of experiences in writing

assessment influences the raters' scoring performances. Huot (1993) defined expert raters as those who had extensive experiences in holistic scoring and novice raters as those who were not given rater training or did not have experiences in writing assessments. His study revealed the following results. First, expert raters exhibited plentiful instances of expressing personal engagement with the writer than novice raters. Second, expert raters were directed in terms of the elements they were assessing, whereas the focuses of novice raters' comments were varied. Lastly, expert raters exhibited fluent reading processes, but novice raters' reading processes were interrupted a lot.

In Cumming's (1990) study, experienced raters were collected from teachers with more than seven years of experiences in ESL composition instructions, meanwhile, novice raters were selected from a student-teacher group with no prior teaching and scoring experiences. Through verbal report analysis, the researcher figured out several distinctive scoring behaviors found across experienced raters. First, they envisioned the situation of the writer, and second, they directed their own reading processes toward the key criteria in the texts. Third, they reflected and summarized their rating judgements collectively. On the contrary, novice raters were more involved with editing phrases and classifying errors than did experienced raters.

Furthermore, Cumming et al. (2002) figured out some differences in scoring sequences between the two rater groups. That is, experienced raters started the scoring processes by scanning the overall layout such as length, neatness of handwriting, or paragraphing. Then, they judged grammaticality, comprehensibility, and rhetorical organizations while reading essays in detail. For the last stage, they monitored their initial judgements and articulated the final decisions. Table 2.2 displays the experienced raters' scoring processes.

Table 2.2

A Prototypical Sequence of Decision Making while Writing Assessment, Cumming et al. (2002, p. 74).

1	Scan the composition for surface-level identification, such as length, format, paragraphing, script (typed or handwritten)
2	Engage in interpretation strategies and read essays while exerting certain judgement strategies a. classifying error types b. identifying comprehensibility c. interpreting rhetorical strategies d. envisioning the situation and personal viewpoint of the writer
3	Articulate a scoring decision, while summarizing and reinterpreting judgements.

Lastly, in Wolfe et al.'s (1998) study, experienced raters turned out to show a higher degree of intra-rater agreement. This was possible because the extensive experiences in scoring assessment shaped the raters' internalized

scoring rubric, which in most cases corresponded to predetermined scoring rubric. On the contrary, less proficient raters manifested self-generated perspectives toward the scoring rubric and exhibited a lower level of intra-rater reliability.

In short, the scoring behaviors of the raters with extensive experiences in writing assessment can be summarized as the following: personal engagement with the writer, envisioning the writer's situations, directedness to the key criteria of writing, and higher degree of intra-rater agreement. On the other hand, the novice raters were characterized by interrupted reading flow, varied foci of assessment, and the orientation to grammatical features.

So far, rater-internal factors were explored in terms of reading style, rater bias, and the amount of writing assessment experiences. In the following section, the experienced EFL raters' scoring behaviors are going to be explored to set the grounds for the analysis of Korean English teachers' scoring behaviors.

2.3 EFL Raters' Scoring Behaviors

The precedent studies related to second language studies have

investigated EFL raters' scoring behaviors through the comparison with the native English raters. For instance, Kobayashi and Rinnert's (1996) research on Japanese teachers of English reported the following results. First, Japanese raters were more lenient in assessing linguistic features than did native English raters as they lacked in native intuition and could not detect as much unnatural expressions as native raters did. Second, in general, Japanese raters gave higher holistic scores than native raters did. Third, rhetorical patterns, which are largely influenced by culture, affected the writing assessment the most. Even though Japanese raters were assessing English compositions, they preferred Japanese rhetorical patterns and gave higher grades to them.

As a similar line of research, Lee and Choe (2012) compared the scoring behaviors between native English teachers and Korean English teachers. They revealed the following three distinctive scoring tendencies between the two groups. First, the Korean teachers gave higher scores than did the native speakers in both holistic and analytic scoring, and the differences were statistically significant. Second, the native raters utilized their native intuitions to judge linguistic features of the writing, while the Korean raters did not. Third, both groups treated the logical flow of ideas as the most important criterion, but the Korean raters were more sensitive to

judging linguistic features of essays. This study supports the idea that the raters from different background and prior experiences displayed distinguishable scoring patterns.

Cho (2003) compared native English and Korean English teachers' scoring processes and their use of rating criteria shown during writing assessment. The participants were asked to grade English compositions in four categories such as content, organization, language, and fluency and the different scoring behaviors of the two groups were revealed as follows.

First, Korean raters exhibited repeated use of reading and interpretation strategies, which was not often observed in native raters. Second, native raters mostly devoted to ideational and rhetorical features of the essays, while Korean raters were more involved in interpreting scoring rubric and detecting linguistic errors. Third, rater variables were found even within the Korean rater group. This is to say that, the Korean raters with different backgrounds and prior experiences, performed differently to a great extent. Thus, the scoring processes of non-native raters with different personal backgrounds should be interpreted with careful caution.

Lastly, Shin (2015) contrasted the scoring tendencies of experienced native English and experienced Korean English teachers. The Rasch analysis

found that the two groups were not statistically distinguished in terms of rater bias, severity, the amount of attention devoted to writing constructs, and the grades. This is to say, EFL raters with plentiful knowledge and experiences in writing assessment are expected to perform like native raters. This also implies the necessity of providing less experienced English teachers with considerable opportunities to accumulate experiences and knowledge in writing assessment.

So far, the previous literature on scoring behaviors of native English raters and EFL English teachers were compared. In short, non-native raters mostly employed reading and interpretation strategies and were more oriented to detecting grammatical errors while grading essays. Nevertheless, the EFL raters with broad knowledge and experiences in writing assessment are expected to demonstrate native-like writing assessment performances.

2.4 Gaps in the Literature

The review of previous literature has identified the two areas that need more exhaustive investigations. First of all, the Korean English teachers' decision-making processes and the aspects of writing they attend to require more detailed discussions. Even though Korean English raters' scoring

behaviors were studied through the comparison with native raters, their latent cognitive processes, i.e., how they construct internalized scoring rubric and apply it to the essays, are under-researched. Secondly, Korean English teachers' scoring behaviors shown on middle school students' essays are not much investigated. Most precedent literature has dealt with scoring processes manifested on relatively high-level essays, which were written by under-graduate students or test-takers for TOEFL test. As the quality of written compositions affects the raters' scoring behaviors, middle school students' essays can bring about different results.

Therefore, this study attempts to fill in these two gaps through qualitative analysis of think-aloud protocols. Detailed examinations of the Korean EFL teachers' scoring processes will enrich the understanding of the complicated decision-making procedures of writing assessment.

Chapter 3. Methodology

This chapter contains the discussion of the methodological approach and research design employed to explore the research question set forth in Chapter 1. This chapter includes four sections that describe participant recruitments, employed instruments, the general research procedures, and data collection and analysis.

3.1 Participants

A total of nine raters was recruited using an online noticeboard. A recruitment notice was posted on an online community site of which English teachers are members nationwide and all participants voluntarily applied for this study. Their background information was collected through a preliminary survey.

3.1.1. Preliminary Survey

The preliminary survey included questions about a variable that might influence Korean EFL teachers' writing assessment performances. The variable considered for this study was the different experiential backgrounds about the writing assessment. Many articles (Eckes, 2008; Huot, 1993; Sakyi,

2000; Wiseman, 2012; Wolfe et al., 1998) claimed that raters' different backgrounds contribute to the individual rating processes and utilization of the scoring rubrics. More specifically, it was argued that the raters who had broad experiences in scoring procedures and received extensive training performed differently from those who had not in terms of the qualitative analysis of scoring behaviors. In this study, the length of teaching career, experiences of writing assessment, and the amount of any training the participants had taken were treated as the possible sources of the raters' experiential backgrounds.

3.2 Instruments

3.2.1 Writing Samples

Ten writing samples were collected from ten students attending a middle school in Seoul during the second week of June 2019. They were all members of an English club and took a writing test in an after-school program. Their English proficiency levels ranged from intermediate to advanced based on their school records. Also, most of them mentioned that they have received extra private English education to improve their proficiency.

As a related topic with the textbook, students were asked to write an autobiography, assuming that they are at their 80s. While writing the essay, students were free to ask questions regarding grammatical features and expressions to be used to teacher and/or peers. They were also allowed to use teacher's laptop computer to look for unknown words but not allowed to use personal electronic devices or to refer to any online translators.

When writing the essay, the instructor asked the students to fulfill two requirements. First, the essay should be more than 150 words, and second, it should include three grammatical features covered in the textbook, which were "non-restrictive relative pronoun," "relative pronoun *what*," and "present perfect progressive." Some students highlighted on the required grammatical features to make them noticeable (refer to Appendix 1 for the students' English essays). However, the presence of these two requirements were not said to the raters.

3.3 Scoring Procedure

The raters, recruited through an online community site at the end of June 2019, completed a preliminary survey. They were then requested to grade students' writings and produce concurrent think-aloud protocols by the

end of July. In what follows, more detailed explanations on scoring procedures and producing verbal reports are going to be provided.

3.3.1 Scoring Processes and Think-Aloud Protocols

The nine raters who agreed to participate in the study were asked to employ holistic scoring procedures in assessing students' essays. Park (2006, p. 65) summarizes holistic scoring method as the following.

> A rater assigns a single score to a text based on the overall impression of the text. This scoring method reflects the idea that writing is a unidimensional entity and can be captured by a single scale which integrates the inherent qualities of the writing.

Barkaoui (2010) pointed out that during holistic scoring raters have to weight certain traits and combine the assessments of the different traits to arrive at a scoring decision. Considering the research question of this study, which is to figure out the strategies often used and the constructs more attended to when

deciding the levels of the written composition, the holistic scoring method was the proper scoring method to be utilized.

During the course of holistic scoring, the participants were asked to generate concurrent think-aloud protocols. Ericsson and Simon (1980) were the first who claimed the usefulness of think-aloud protocols in qualitative analysis of research. They argued as in the below.

> One of the most direct and widely used methods to gain information about subject's internal state is to instruct them to think aloud or talk aloud. With this procedure, the heeded information may be verbalized either through direct articulation or by verbal encoding of information that was originally stored in a nonverbal code. (p. 219)

Sakyi (2000) highlighted the effectiveness of applying think-aloud protocols in the studies of writing assessment. He put that think-aloud protocols had the potential of providing rich information on the mental processes of holistic raters as they read and score essays. Following this

viewpoint, a considerable amount of literature had utilized think-aloud protocols in their studies of writing assessments (Cumming, 1990, Cumming et al., 2002; Huot, 1993; Lumley, 2005; Milanovic et al., 1996; Sakyi, 2000; Vaughan, 1991; Wolfe et al., 1998).

In the current study, the following measures were taken to help participants better carry out think-aloud protocols. First, the researcher provided all the raters with a voice recorded file of think-aloud demonstration to help them get a picture of how it goes. Plus, not to bias the raters' actual scoring procedures, the demonstration was conducted about a subject unrelated to the target task, which was to explain how to use an e-book. Second, the raters were directed not to look over the papers before starting scoring. Third, they were told not to make pauses for more than three seconds, just to make sure to speak out every detail of their thoughts. Fourth, the raters were requested to score the essays in the given order that had been provided to them in order to prevent any ordering effect on their ratings. Milanovic et al. (1996) mentioned that a random order of scripts could function as another variable during the assessment. Lastly, all participants completed the think-aloud protocols individually at the participant's home. This was done to allow

them to have enough time to verbalize and to minimize researcher effects on the participants' performance.

3.4 Data Collection and Analysis

Table 3.1 illustrates the general outline of data collection and analysis procedures.

Table 3.1

The General Outline of Data Collection and Analysis

Stage	Data Collection	Analysis
1	Preliminary Survey	Qualitative Analysis
2	Scoring and producing verbal reports	
3	Data transcription and translation	
4	Data coding and analysis	

For the first phase of the research, the background information was gathered through the preliminary survey, then, concurrent verbal reports were collected in the second stage. The verbal reports were produced in Korean to facilitate the raters' thought processes. At the third stage, the researcher transcribed the verbal reports and then translated them into English. At the

last stage, the translated data were segmented by idea units and coded by scoring behaviors. The detailed procedures of segmenting the data and constructing coding schemes are described in Section 3.4.1.

3.4.1. Data Coding and Analysis

The raters' voice-recorded verbal reports ranged from the minimum of 20 minutes to the maximum of 36 minutes in length. The transcribed files were parsed into idea units, which were defined as "a message segment consisting of a topic and comment that is separated from contiguous units syntactically and/or intonationally" (Ellis & Barkhuizen, 2005, p. 154, cited from Lintunen & Makila, 2014). Also, Panos (2015) explained some distinguishing features of idea units. First, idea units are segmented by either a falling or rising intonation. Second, idea units are delimited by pauses, and lastly, they are expressed in a single clause formed by a verbal predicate and the phrases associated with it.

When applying these criteria to the collected data, the researcher made some modifications as the data was produced in Korean. First, contouring intonations were not considered during parsing procedures since Korean language sentences do not often include sentential intonations. Also, pauses more than three seconds were considered as a dividing point between

clauses. Third, repetitive lines were not counted as a single idea unit. The segmented idea units were grouped by similar focus, which was later used to form individual scoring behaviors. This led to the construction of a total of 12 key scoring behaviors in the end.

Chapter 4. Results and Discussion

This chapter presents the results of the qualitative analysis implemented in order to address the research question: What strategies do the Korean English teachers often use when scoring essays? Section 4.1 describes the results of the preliminary survey and Section 4.2 presents each rater's scoring characteristics. Then, Section 4.3 summarizes the key 12 scoring behaviors derived from the analysis and this chapter concludes with discussion points drawn from the research findings in Section 4.4.

4.1. The Analysis of the Preliminary Survey

All the nine participants' background information was gathered through the preliminary survey. The four questions provided to the participants are described in Table 4.1.

Table 4.1

Survey Questions

1) How long have you been teaching English?
2) How much proportion does the writing assessment account for the whole performance-based assessment? Please answer in percentage (%).
3) Have you taken any teacher training courses or lectures related to writing assessment? If you have, please describe them specifically.
4) What do you think is the most important thing to consider during writing assessment?

The questions 1 and 2 were implemented to see the participants' prior experiences in teaching and assessing writing and the question 3 to check the amount of trainings that they had taken so far. The last question was employed to figure out the participants' perceived significance of writing constructs. Table 4.2 summarizes their responses.

Table 4.2

Raters' Responses to the Preliminary Survey

Rater	(1) Experiences of Writing Assessment		(2) Experiences of receiving assessment-related courses	(3) Perceived Importance of Writing Construct
	The length of teaching career	Writing assessment portion		
A	3 years	10-20%	Taken 1-2 classes of assessment in general in college	Content
B	0.6 years	50%	None	Intelligibility
C	3.9 years	75%	None	Coherence, grammar, vocabulary
D	0.8 years	10%	None	Intelligibility
E	2 years	25%	Taken one writing assessment class in graduate school	Grammar and content
F	6 years	50%	Taken one language assessment class in college	Grammatical correctness
G	11 years	40%	Taken 2-3 teacher training sessions about process-oriented assessment	Setting a rubric, objective scoring, notifying test-takers of scoring criteria
H	7 years	60%	Taken 10 classes of writing assessment throughout college, graduate school, and teacher training	The variety use of vocabularies, expression of ideas, and grammar is not that important

			sessions	
I	5 years	60%	Taken one process-oriented writing class in graduate school	Content and creativity

Note. All the participants' names were replaced with alphabetical letters to conceal the information of the raters extraneous to the study.

The characteristics of the participants can be summarized in two ways. First, the raters' experiential backgrounds varied to a great degree in terms of the length of their teaching career and experiences in writing assessment. That is, the length of the raters' teaching experiences ranged from 0.6 to 11 years and writing assessment accounted for 10-75 percentages of the whole performance-based assessment. Secondly, the raters showed different thoughts about the perceived significance of writing constructs. Specifically, five participants considered "content related features" as the most important construct, one participant "grammar," two participants "content and grammar," and one participant "reliable scoring processes." In what follows, all participants' think-aloud protocols are going to be thoroughly examined in association with their responses to the preliminary survey.

4.2. Think-Aloud Protocol Analysis

In this section, descriptive accounts of a total of nine participants' think-aloud protocols are going to be presented. Each of the following subsections consists of two parts: raters' background information and scoring strategies utilized during scoring processes. From Section 4.2.1 to Section 4.2.9 deal with the qualitative analysis of individual raters' verbal reports and a summary of the analysis is suggested in Section 4.2.10.

4.2.1. Characteristics of Rater A and Her Scoring Processes

Rater A, a female English teacher who had three years of teaching experience at middle school, reported that "content" is the most important trait to be considered when scoring essays. The preliminary survey revealed this rater's insufficient amount of experiences in writing assessment compared to other participants. That is, currently, writing assessment accounted for only 10-20% of the whole performance-based assessment and it was usually done through filling in the blanks. Also, she put that she had received one class about language assessment in college.

The inspection of Rater A's think-aloud protocols displays her unique scoring behaviors in the following aspects. First, reading and interpreting

essays were conducted at sentential levels, which is usually followed by linguistic judgments. This is to say, she lacked in holistic comprehension and judgements of the content. The excerpt 1 illustrates her reading style.

Excerpt 1. Rater A, 10th essay

> First of all, the title is “meaningful life” um..... So, the first letters of each word should be capitalized. In the third line, this student wrote “I went to English kindergarten when elementary school.” A transition word “then” should be used for the natural connection of the two sentences. Also, in the fifth line, the definitive article “the” should be used in front of “high school.”

As shown in the above excerpt, she tended to read essays line by line, while making linguistic judgements on such features as article use or the appropriateness of expressions. Second, when evaluating the quality of the content, Rater A mostly considered “cohesion” and “elaborated description” as important traits than others. The excerpt 2 displays her concerns.

Excerpt 2. Rater A, 3rd Essay

> Overall, even though life events were vividly described, I think I can give only six points to this one because of many grammatical mistakes and low levels of cohesion.

The excerpt 2 manifests Rater A's concerns about grammatical accuracy, cohesion, and elaborated description when deciding scores. Third, she rated the levels of the essays by mentioning good and deduction points by three focuses: language, content, and organization. More specifically, she started scoring processes by mentioning the overall organization of the essay and then judged grammaticality and idea development while reading through the compositions. At the end of the scoring processes, she decided how many points to take off from the full marks. Rater A's entire scoring processes were summarized in Table 4.3.

Table 4.3

Rater A's Scoring Processes

		Scoring Process			Point deduction
		Language focus	Content focus	Organization focus	
1	Good Points		·Cohesion		
	Deduction Points		·The lack of elaborated description	·No title	-1
2	Good Points	·The use of transition words ·Not many grammatical errors	·Detailed description of life events		
	Deduction Points			·Short length	-1
3	Good Points				
	Deduction Points	·Wrong verb tense ·Long sentences	·Repetition		-4
4	Good Points		·Cohesion	·Text format	
	Deduction Points	·Grammatical errors			-1
5	Good Points				
	Deduction Points	·Grammatical errors	·Bland storyline		-4
6	Good Points		·Elaborated description ·Cohesion		
	Deduction Points	·Grammatical errors		·Short length	-2
7	Good Points	·Rich expression	·Elaborated description ·Coherence	·Text format	

	Deduction Points	·Spelling mistake		·No title	-1
8	Good Points		·Elaborated description		
	Deduction Points	·Low level of coherence in the first paragraph			-1
9	Good Points		·Elaborated description		
	Deduction Points	·Spelling mistake ·Unnatural expressions			-1
10	Good Points	·Rich expressions	·Elaborated description ·Cohesion		
	Deduction Points	·Cohesion			-0.5

Rater A's scoring processes showed that linguistic elements influenced more on her scoring decisions than did content and organization. Even if she reported on the preliminary survey that "content" is the most important construct to be considered when deciding the quality of the essay, linguistic elements were revealed to have greater effects than content. When asked about the reason for this during retrospective interview, she mentioned that she believed the delivery of message bears the primary importance in writing assessment from the aspects of communicative competence, however, when actually doing the scoring, it was hard to decide the comparative levels

of each essays because it could be subjective by rater. Thus, she got to resort more to linguistic features for scoring.

Another point to make here is that she did not apply the same perspectives to all essays. This is to say, instead of setting a standardized scoring rubric, she evaluated the quality of the essay by deducting points when she noticed erroneous or unnatural features. This, consequently, would increase rater variability during scoring processes.

4.2.2. Characteristics of Rater B and Her Scoring Processes

Rater B, a female English teacher, has worked as English teacher for less than a year. She mentioned that 50% of performance-based assessment was conducted through writing assessments, which usually required students to write a short sentence. She reported that “intelligibility” is the most important construct to be measured during writing assessment. Table 4.4 illustrates her underlying scoring processes.

Table 4.4

Rater B's Scoring Processes

		Scoring Process			Point deduction
		Language focus	Content focus	Organization focus	
1	Good points	·Collocation use			
	How to improve				
	Deduction Points		·The lack of description of the life in old age		-1
2	Good Points	·Expression use ·Present perfect progressive use			
	How to improve	·The addition of preposition "in" after the verb "succeed"	·More elaborated description		
	Deduction Points				
3	Good Points	·Collocation use			
	How to improve	·Neat handwriting ·Deletion of "be verb" in "when I was grown up"	·More elaborated description		
	Deduction Points	·The lack of required grammatical elements	·The lack of description of the life in old age		-2
4	Good Points	·Expression use		·The use of signal words	
	How to improve	·The past tense use ·The addition of transition word/preposition			
	Deduction Points				
5	Good Points				
	How to improve	·Preposition/expre-ssion use ·The past tense use	·Insertion of transition words for logical connections		

			between sentences
	Deduction Points		
6	Good Points		
	How to improve	·Verb tense ·Spelling	
	Deduction Points		
7	Good Points	·The use of present participle ·High level vocabulary use	
	How to improve	·Vocabulary choice ·Spelling ·The past tense use	
	Deduction Points		
8	Good Points	·High level vocabulary use ·The use of gerundive after the verb "continue"	
	How to improve		
	Deduction Points		
9	Good Points	·Vocabulary choice ·The use of dummy "do"	·Elaborated description
	How to improve		
	Deduction Points		
10	Good Points	·The use of relative pronoun ·Vocabulary choice	
	How to improve	·The verb tense	
	Deduction Points		

Similar to Rater A, Rater B attended to language, content, and organization of the texts while scoring essays. Also, she arrived at scoring decisions by deducting some points from the full mark, which is 10 points. Rater B's distinctive scoring patterns could be summarized as in the following aspects. First, she was mostly concerned with grammaticality judgments during scoring processes. However, instead of taking off points for every erroneous or unnatural expressions, she provided revisions of those expressions. She mentioned that she deducted points only when at least two out of three required grammatical features (relative pronoun what, to-infinitival used as adverb, present perfect progressive) were not used. Second, elaborated descriptions of life events were of great importance in measuring the quality of content. The following excerpt 3 illustrates how she mentioned about the detailed description of content.

Excerpt 3. Rater B, 3rd Essay

> In the third line, this student said that "the preschool teacher took care of me." I am curious about *how* she took care of her. A little more detailed elaboration on this would explain why this student wanted to be a preschool teacher.

However, the degree of elaborated description was not reflected into scoring decisions. She remarked that she only judged if the days of childhood, adolescence, adulthood, and old age were all covered in the essay for scoring decisions. For instance, as for the essay 1 and 3, she deducted points for the lack of description of the life in old age, not for the lack of elaborated description of life events in general. Consequently, her overall scoring processes and tendencies made her as a lenient rater, resulting in eight essays receiving full marks.

4.2.3. Characteristics of Rater C and His Scoring Processes

Rater C is a male English teacher who has 3.9 years of teaching experience at high school. He reported that "grammaticality," "appropriate expression use," and "coherence of the essay" are the significant traits to be considered during writing assessment. Despite his beliefs in the importance of the three aspects, he turned out to be "grammar-oriented rater" (Vaughan, 1991) who gives "language" greater consideration in assessing second language writing ability. The excerpts 4 and 5 display his scoring tendencies.

Excerpt 4. Rater C, 5th Essay

> "I was born on April 20 in Seoul I moved to Sangamdong ... I entered to Sangji" I think *to* should not be used here. […] "I came back to Korea and I graduate Sangji elementary school," *from* is missing here. This student made a lot of mistakes related to preposition. […] Besides the preposition use, this student made a lot of grammatical errors. I think I should deduct some points.

Excerpt 5. Rater C, 4th Essay

> Overall, this essay contains high level of expressions and I cannot notice any big grammatical errors. I think I can give a high score for this one.

As evidenced by the above excerpts, Rater C's verbal reports were mostly concerned with grammaticality judgements, and consequently this scoring tendency led him to less consider "content." Also, his scoring processes were mostly comprised of giving Korean interpretations to each line. That is,

instead of making efforts to comprehend the essay holistically, he was mostly engaged in line by line translations of the essays. The detailed scoring processes are described in Table 4.5.

Table 4.5

Rater C's Scoring Processes

		Scoring Process			Point deduction
		Language focus	Content focus	Organization focus	
1	Good points				
	Deduction Points	·Several grammatical errors	·Bland storyline		-3
2	Good Points	·Not many grammatical errors	·Comprehensibility		
	Deduction Points				-1
3	Good Points				
	Deduction Points	·Unnatural expression use ·Verb tense ·Grammatical errors	·Not concrete description of the writer's life		-4
4	Good Points	·Relative pronoun use ·Good expression use		·Long length	
	Deduction Points				
5	Good Points				
	Deduction Points	·A lot of grammatical errors			-4
6	Good Points				
	Deduction Points	·A lot of grammatical errors			-3
7	Good	·Use of relative	·Comprehensibility		

	Points	pronoun “what”			
	Deduction Points				
8	Good Points				
	Deduction Points		·Illogical connection between sentences		-2
9	Good Points	·Good expression use		·Long length	
	Deduction Points	·Several grammatical errors			-1
10	Good Points			·Long length	
	Deduction Points	·Several grammatical errors			-2

As can be seen in the above table, grammaticality judgement of essays had the major influence on scoring decisions. Another point to make is that Rater C exhibited a certain degree of inconsistent scoring tendencies. For instance, he deducted one point for having several grammatical errors from the ninth essay, but he deducted two points from the tenth essay. Even though he had plentiful experiences in writing assessment, he exhibited a rather unbalanced approach to scoring processes.

4.2.4. Characteristics of Rater D and Her Scoring Processes

Rater D is a female English teacher who has 0.8 years of teaching experience at high school. She reported that she had implemented writing assessment only once, which was to introduce Korean tourist sites. She

remarked that “comprehensibility” is the most important construct in writing assessment.

She started scoring processes by reading and interpreting the essays, pointing out grammaticality, comprehensibility, and expression use without displaying summarization strategy just like the other raters. However, what made her distinguished from other raters was that she created her own scoring rubric by undergoing three revisions. The first scoring rubric came out after reading the first essay. Table 4.6 summarizes Rater D’s first scoring rubric.

Table 4.6

The First Version

	Scoring Focus	Scoring Guideline	Assigned Points
(1)	Language	The past tense use	3
(2)	Content	Coherence of the content	7

Her first scoring rubric consisted of two focuses: language and content. For the language focus, Rater D considered only if the past tense was used and assigned three points for this construct. Also, as for the content, she mostly considered the coherent flow of the content and seven points were assigned for this construct. Instead of specifying when to deduct points, full

marks were given if a particular essay satisfied the criteria.

Then, after reading the fourth essay, Rater D decided to divide "content" into "organization" and "idea." The excerpt 6 explains her thought processes for this decision.

Excerpt 6. Rater D, 4th Essay

> This essay is perfect in terms of organization, being divided into three paragraphs by using signal words like "first," "second," and "lastly." Seeing this, I thought it would be better to "content" into "organization" and "idea." I am going to change the previous scoring rubric by assigning two points to "grammar," three points to "organization," and five points to "idea."

After allotting full points for each scoring category, she came up with more sophisticated scoring guidelines as reading through the essays. Specifically, as for "organization," she considered if the essay included introduction, body, and conclusion with indentations. Also, the quality of content was decided by

whether the writer's whole life was covered. The following excerpt 7 describes how she came to judge the levels of the content in detail.

Excerpt 7. Rater D, 4th Essay

> This essay is mostly talking about school days like how she studied math or how she entered a good university. I think the writers' days of old age should be more described to make it as a good autobiography. So, I will deduct one point in content.

In the end, Rater D's revised version of scoring rubric was constructed as in Table 4.7.

Table 4.7

The Revised Version

	Scoring Focus		Scoring Guideline	Assigned Points	
(1)	Language		The use of the past tense	2	
(2)	Content	Organization	Introduction, body, and conclusion	3	8
		Idea	Covering the whole life (up to the 80s)	5	

When she has read up to the 7th essay, she thought that four

grammatical features (to-infinitival as adverb, *what* as relative pronoun, present perfect progressive, and non-restrictive relative pronoun) were required to be used by noticing the highlights. This made her establish the final version as illustrated in Table 4.8.

Table 4.8

The Final Version

Scoring Focus			Scoring Guideline		Assigned Points
(1)	Language	The use of required grammatical features	·To-infinitive as adverb	0.5	2
			·Present perfect progressive	0.5	
			·*What* as relative pronoun	0.5	
			·Non-restrictive relative pronoun	0.5	
(2)	Content	Organization	·Including three paragraphs: introduction, body and conclusion	3	3
		Idea	·Covering the whole life (up to the 80s)	-2	5
			·Low comprehensibility	-1	

In the final version, she looked for whether the four required grammatical features were used and assigned 0.5 points to each when making scoring decisions for language focus. However, the way evaluated the levels of idea was different from that of language and organization. Instead of assigning points for each scoring criteria, she took off one or two points when the essay seemed inadequate to be considered as a completed autobiography.

After coming up with the final version, she went back to the first essay to change her prior decisions. Only Rater D went through the whole set of essays again to apply the same scoring rubric.

In short, her scoring behaviors could be summarized in two ways. First, she thought of the specific writing constructs to be assessed and weighted them by considering relative significance. Second, she kept reviewing the previous essays to confirm her judgements. Even though this rater did not have extensive experiences in writing assessment compared to other raters, she demonstrated refined scoring processes by creating her scoring rubric and applying it to all essays in the same fashion.

4.2.5. Characteristics of Rater E and Her Scoring Processes

Rater E, a female English teacher who has two years of teaching experience at high school, reported that she has conducted writing assessments usually through guided writing tasks. Also, she mentioned that adopting a balanced view toward "grammar" and "content" is important in assessing writing ability.

The investigation of think-aloud protocols revealed that Rater E's scoring processes were not different from Raters A, B, and C with respect to

their orientation to judging grammaticality. Even if she mentioned that coherence and grammatical correctness should be reflected to scores, she mostly based her scoring decisions on the grammaticality of the essays. Also, even though the writers described their lives vividly and in a well-organized format, full points were not given if grammatical or expressionistic errors were observed. Table 4.9 shows her scoring processes.

Table 4.9

Rater E's Scoring Processes

		Scoring Process			Point deduction
		Language focus	Content focus	Organization focus	
1	Good points				
	Deduction Points	·Grammatical errors ·Spelling mistakes ·Misuse of prepositions	·Low coherence		-2
2	Good Points		·Elaborated description of the writer's job		
	Deduction Points	·Omission of prepositions			-1
3	Good Points		.		
	Deduction Points	·Verb tense ·Spelling mistakes	·Low coherence		-2
4	Good Points		·Coherent flow	·Good organization	
	Deduction Points				

5	Good Points				
	Deduction Points	·Grammatical errors	·The lack of meaningful life events		-2
6	Good Points		·Organized in a time-sequence		
	Deduction Points	·Grammatical errors			-1
7	Good Points		·Elaborated description		
	Deduction Points	·Grammatical errors			-1
8	Good Points		·Description of the writer's life values ·Organized in a time-sequence		
	Deduction Points				
9	Good Points		·Elaborated description of the writer's life		
	Deduction Points	·Minor grammatical errors			-1
10	Good Points		·Elaborated description of the writer's life		
	Deduction Points	·Minor grammatical errors			-1

From the above table, it could be noticed that she made scoring decisions without weighting writing traits. To be more specific, as for the first essay, Rater E mentioned four deduction points and took off two points in the end. It is elusive which aspects affected more than others for the final scoring decisions. The excerpt 8 this scoring tendency.

Excerpt 8. Rater E, 1st Essay

> This essay contains many linguistic errors such as spelling mistakes, preposition misuse, article omission, and omission of transition words. Also, I cannot understand some expressions as well. So, I will give eight points to this one.

As can be seen in the above excerpt, she arrived at scoring decision mainly based on the general impression of the essay, instead of coming up with specific scoring standards.

4.2.6. Characteristics of Rater F and Her Scoring Processes

Rater F, a female English teacher who has six years of teaching experience at middle school, reported that "linguistic accuracy" is the most important criteria to be reflected during writing assessment. When asked for the reason, she responded that only linguistic accuracy can be judged without making any controversy.

When grading the ten essays, she started the scoring processes by scanning the whole composition, mentioning features such as the quality of handwriting and the overall organization of the essay. According to Milanovic

et el. (1996), this reading tendency is termed as “the principled two-scan/read,” which is to scan the essay for length and appearance, followed by the second reading to confirm the initial decision. This initial scanning process was observed throughout the whole verbal reports. The following excerpts 9 and 10 describe this behavior.

Excerpt 9. Rater F, 6th Essay

> The title of this essay is “my happy life” and the neat handwriting gives a good impression of this essay.

Excerpt 10. Rater F, 3rd Essay

> As soon as I saw the 3rd essay, I thought I should give a low score to it. Considering that the writer is a third grader, the levels of vocabularies and sentence structures are relatively simple.

Rater F got to make the first impressions of the essays through the initial

scanning, which influenced the rest of scoring processes. Especially, when reading the third essay, Rater F received the negative first impression by the short length, clumsy handwriting, and the lack of paragraphing, which led her to give a lowest score out of the ten essays. This reading behavior is classified as "first impression dominating approach" according to Vaughan's (1991) five reading styles.

After overlooking the whole essay, this rater tried to grasp the content, mentioning significant life events covered in the essay. Then, at the last stage, she made evaluative scoring decisions by taking grammar and content into account. However, she did not construct any specific scoring guidelines for scoring decisions. Excerpt 11 illustrates how she made scoring decisions.

Excerpt 11. Rater F, 5th Essay

> I'd like to give seven points to this one. Even though there are not many grammatical errors, but things like making friends, having a good teacher, going to school, and servicing army are insufficient to make it as a good autobiography.

Like this, after reading the whole essay, she came to make evaluative decisions by taking off some points when the given essay does not satisfy her scoring standards. The detailed scoring processes are described in Table 4.10.

Table 4.10

Rater F's Scoring Processes

		Scoring Process			Point deduction
		Language focus	Content focus	Organization focus	
1	Good points	·Not many grammatical errors ·High level expressions			
	Deduction Points	·Omission of article(-1) ·Two awkward expressions(-2)			-3
2	Good Points				
	Deduction Points	·Grammaticality(-2)		·Short length(-1)	-3
3	Good Points				
	Deduction Points	·Simple sentence structure	·Not covering the writer's whole life ·Pointless development		-5
4	Good Points	·Rich expression use ·No grammatical errors		·Long length	
	Deduction Points				
5	Good Points				
	Deduction Points	·Two instances of ungrammatical expressions (-2)	·Listing of life events (-1)		-3
6	Good Points				

	Deduction Points	·Simple sentence structure(-1)		·Short length (-2)	-3
7	Good Points		·Elaborated description of life events	·Long length	
	Deduction Points				
8	Good Points				
	Deduction Points		·Simple contents		-2
9	Good Points	·High level expression use ·No grammatical errors	·Comprehensibility		
	Deduction Points				
10	Good Points	·No grammatical errors	·Detailed description of her life ·Expressed the values of her life	·Long length	
	Deduction Points				

Rater F's scoring processes revealed her five scoring behaviors as the following. First, she made more comments on language-related features than content and textual organization, and second, when it comes to content, elaborated descriptions of life events were appreciated. Third, inconsistent scoring methods were observed from the verbal reports. For instance, Rater F deducted one point for the second essay because of the short length, but two points for the sixth essay, even though the both essays filled up to 15 lines. Fourth, it is elusive which writing trait received more weight than others. As

she mostly made her scoring decisions without specific scoring guidelines, how she arrived at the final decision was not clearly accounted for. Lastly, she pointed out grammatical mistakes but did not provide any revisions for them. Except for the scoring sequence where Rater F exhibited the initial scanning of the scripts, the rest of scoring processes was similar to that of Raters A, B, C, and E. That is, they did not create specific scoring guidelines and based their evaluative decisions mostly on grammaticality and general impression of the essays.

4.2.7. Characteristics of Rater G and Her Scoring Processes

Rater G is a female rater who has 10 years of teaching experience at high school. She reported that "reliable scoring" is the most important trait to be considered during scoring processes and also, it is necessary to set a detailed scoring rubric in advance and notify students of it before writing sessions.

Rater G's scoring processes can be summarized in two ways. First, she mainly took "content" and "grammar" into consideration without weighting constructs just like the other raters. Second, she constantly made scoring decisions through the comparison to other essays. As mentioned

above, the scoring strategy of comparison is a part of macro-strategies, which are mostly observed in experienced raters (see 4.3.1), and lastly, she constructed her own scoring rubric after reading the all essays and revised her initial decisions according to it. The excerpt 12 displays this scoring behavior.

Excerpt 12. Rater G's comment

> I set three scoring criteria by reading through all the essays. First, did the essay deal with the writer's whole life? Second, did the essay contain the required grammatical features? Third, are there any other grammatical errors in the essay? Keeping these scoring criteria in mind, I am going to go back to the first one to see if my first decisions need changes.
>
> The first student barely fulfills the requirements, so I think I have to change it to 6 points. The second student satisfies the three criteria, so it stays at 10. About the third composition, it is hard to find if it follows all the requirements and also it has too many grammatical errors. As I cannot deduct points for every error, I'll deduct one point for one to

two errors, two points for three to four errors, three points for five to seven errors. For those with more than seven errors, four points are deducted. Besides the content of the third essay being too simple, I should deduct one more point. Thus, it gets five points.

The first paragraph of the excerpt 12 describes Rater G's own scoring guidelines and the second paragraph details how she regraded the essays. Moreover, when re-grading the third essay, she came up with more sophisticated scoring criteria about how many points to take off for the number of grammatical errors. Based on the whole verbal reports, Rater G's scoring rubric was created as in Table 4.11.

Table 4.11

Rater G's Scoring Rubric

Scoring Focus		Scoring Guideline	Scoring
Language	The use of required grammatical features	·To-infinitival is missing	-1
		·Non-restrictive use of relative pronoun is missing	-1
		·Present perfect progressive is missing	-1
	Linguistic correctness	·1-2 grammatical errors	-1
		·3-4 grammatical errors	-2
		·5-7 grammatical errors	-3
		·More than 7 errors	-4

Content	Organization	·Not showing the appropriate textual format (indentation or paragraphing)	-1
		·Length (less than 150 words)	-1
	Topic development	·Repetition	-1
		·Listing of events	-1
		·Not covering the whole life	-1
		·Non-coherent flow	-1

First, the scoring category of language focus was comprised of two parts: the use of required grammatical features and linguistic correctness. While reading the essays, Rater G thought that three grammatical features (to-infinitival, non-restrictive use of relative pronoun, and present perfect progressive) were required to be used. So, this rater deducted points for the number of the required grammatical features that were missing. Next, as for "linguistic correctness," she deducted points for the number of grammatical errors such as the omission of articles or misuse of singular/plural forms.

With regards to the assessment of "content," Rater G considered "organization" and the "topic development." For the organization, the overall structure of the essay and the length were taken into account. Lastly, for topic development, Rater G deducted one point when the essay did not show concise and logical development of ideas. The following excerpts display how she came to formulate scoring criteria when judging the quality of content.

Excerpt 13. Rater G, 3rd Essay

> In general, the writer of this essay mostly listed life events without elaborating her own identity or the values of her life. Also, I could see some redundancies in the repetitive descriptions of her changing dreams. So, I think I should deduct one or two points for topic development.

Excerpt 14. Rater G, 6th Essay

> This student described her imaged future very specifically, but I think she concentrated too much on how she got a job. I am going to deduct one point for this as autobiography is ought to describe the whole life.

As could be observed, she established her own scoring rubric while reading through essays one by one, which influenced her scoring decisions.

4.2.8. Characteristics of Rater H and Her Scoring Processes

Rater H, a female English teacher who has seven years of teaching experience at high school, reported the variety use of vocabularies and

expression of ideas are more important than grammatical accuracy in writing assessment. She responded that she took a total of 10 process-oriented writing sessions throughout college, graduate school, and teacher training courses.

Before starting to read the essay, she looked over the essay for organization, the number of paragraphs, and the length just like Rater G. However, her unique reading styles were found in the following aspects. First, she tried to understand the main points of the essay through summarization. Second, she continuously displays personal engagement with the writer. The excerpt 15 illustrates this reading strategy.

Excerpt 15. Rater H, 9th Essay

> I've only read up to two paragraphs, but I could see that this student thought a lot about what he wants to do in the future. I can vividly picture his life in my head by the way he described his life. If I could get a chance, I want to have a talk with this student.

As explained in Section 2.2.1.1, personal engagement is defined as

"psychological link with the writer" (Vaughan, 1991) and "taking time to interact with students' writing" (Huot, 1993). The excerpt 15 clearly demonstrates Rater H's engagement and sympathizing with the writer. Huot (1993) asserted that personal engagement is typically observed in experienced raters and this is what differentiates Rater H from others. Third, she tried to judge if the grammatical errors were local or global errors. If certain errors do not cause significant misunderstandings, she did not deduct any points for those errors. The following excerpt manifests this scoring behavior.

Excerpt 16. Rater H, 6th Essay

> "Company sales was increased" is a wrong expression and passive expression is one of the common mistakes committed by students. However, as this does cause big misunderstanding, I won't deduct a point for this.

Fourth, she provided the ways of improving the quality of the essays. For instance, as for the fifth essay, she commented that deleting redundant parts or adding more stories after retirement would make the essay better. Table

4.12 describes Rater H's detailed scoring processes and deduction points.

Table 4.12

Rater H's Scoring Processes

		Scoring Process			Point deduction
		Language focus	Content focus	Organization focus	
1	Good points	·Rich expression ·Grammatical accuracy	·Coherent development		
	How to improve			·Paragraphing	
	Deduction Points	·Local grammatical errors		·No title	-4
2	Good Points	·Verb tense	·Elaborated description	·Title ·Paragraphing	
	How to improve	·Use a variety of vocabularies			
	Deduction Points				-2
3	Good Points				
	How to improve			·Paragraphing	
	Deduction Points		·Focused only on how dreams changed		-4
4	Good Points	·Rich expression ·Use of grammatical features	·Logical development	·Easy to read ·Indentation ·Signal words	
	How to improve				
	Deduction Points				
5	Good Points				

	How to improve		·Delete redundant parts ·More stories after retirement		
	Deduction Points	·Repetitive vocabulary use ·Simple sentence structure			-5
6	Good Points	·Verb tense	·Topic development		
	How to improve				
	Deduction Points		·Limited use of vocabulary		-4
7	Good Points	·Use of grammatical features ·Rich expression use	·Good introduction ·Coherent development	·Long length ·Title	
	How to improve				
	Deduction Points				
8	Good Points		·Unnatural expressions		
	How to improve			·Paragraphing	
	Deduction Points	·Simple sentence structure ·Limited use of vocabulary			-4
9	Good Points	·No grammatical errors ·Rich expression use	·Elaborated description	·Paragraphing	
	How to improve		·Verb tense		
	Deduction Points				
10	Good Points	·No grammatical errors	·Reflected the bygone days ·Good conclusion		
	How to improve				

Deduction Points	

What draws attention in the above table is that she deducted some points without even mentioning deduction points as in the second essay. It is thought that this could be possible as she made scoring decisions through the comparison with other essays. In other words, she adopted "norm-referenced assessment" when evaluating the quality of the essays. Thus, the aspects of writing she more attended to can be only limitedly explained through the investigation her scoring processes.

4.2.9. Characteristics of Rater I and Her Scoring Processes

Rater I, a female English teacher who has five years of teaching experience at middle school, reported that she put emphasis on "content" and "creativity" when scoring essays. Reflecting this view, she graded students' essays with higher significance given to "content" and "creativity" than to "grammar" and "expression use."

Her scoring behaviors could be summarized in four aspects. First, she started scoring processes by reading and interpreting the essay without scanning for the overall layout. Second, while reading the essay, she behaved

more like a reader, not a rater. Instead of making judgements about the essay, she tried to understand what the writer wanted to say and the writers' personality. The excerpt 17 displays her reading style.

Excerpt 17. Rater I, 8th Essay

> This writer expressed the important values of his life, which is being active and enthusiastic about everything. Even though now he is old and not healthy as he was in the past, he is saying that he will keep pursuing his life.

Third, she made self-cautionary comments to guard against making prejudgments. The following excerpt 18 displays this tendency.

Excerpt 18. Rater I, 10th Essay

> I think this student's personality was not expressed enough in this essay. To me, the life events described here seemed rather superficial. I think I should deduct some points for this.

However, subjective this perspective could be, I will deduct only one point for this.

Considering that measuring the quality of the content could be more subjective than that of language use, she chose not to deduct many points for content. Fourth, she constructed her own scoring rubric throughout the reading processes. Interestingly, she did not put much emphasis on linguistic features compared to other raters. She did neither look for the required grammatical features nor count the number of grammatical errors. Instead, this rater deducted one point for simple sentence structure and grammatical errors regardless of its frequency. Also, when assessing the content, she mostly considered "topic development," deducting one point when the essay seemed not logical or lacked in creativity. Table 4.13 describes Rater I's detailed scoring criteria.

Table 4.13

Rater I's Scoring Rubric

Scoring Focus			Guideline	Scoring
(1)	Language	Syntactic features	·Simple sentence structure	-1
		Linguistic correctness	·Grammatical errors (regardless of the number of errors)	-1
(2)	Content	Topic development	·Not showing the logical topic development	-1
			·Listing of events	-1
			·Lack of details of life events	-1
			·Lack of creativity	-1

Up to now, all the participants' verbal reports were analyzed from qualitative perspectives. In the following section, a summary of a total of nine participants' scoring processes is going to be presented to figure out the similarities and differences across them.

4.2.10. Summary

In short, the nine participants' scoring behaviors could be characterized in terms of the ensuing aspects. First, most of the raters were more involved in grammaticality judgement than they were in measuring the quality of the content. Some raters factored the number of errors into the scoring decisions, while others considered whether those errors hindered the

readers' comprehension of the essays. Second, when it comes to content, the majority number of the raters considered "elaborated topic development." Third, some raters exhibited expert raters' scoring behaviors. In the Chapter 2, the characteristics of expert raters' scoring behaviors were discussed in three aspects: making a scoring rubric, showing personal engagement with the writer, and displaying the initial scanning of the essay. In this study, three raters (Raters D, G, and I) came up with their own scoring rubric while or at the end of reading the essays. Also, two raters (Raters G and I) showed "personal engagement with the writer." They perceived the writer of the essay as the author, showing emphasizing with the writer. Lastly, two raters (Raters G and H) displayed the initial scanning of the overall essay. In the next section, Korean English teachers' key scoring behaviors are going to be formulated based on the findings so far.

4.3 Key Scoring Behaviors

The analysis of verbal reports rendered the key 12 scoring behaviors[1] and these were categorized by three strategies, judgement, macro-judgement

[1] This table was established with reference to Cumming et al.'s (2002, p. 88) "Descriptive framework of the decision-making behaviors while rating TEOFL writing tasks."

and interpretation strategies. Then, judgement and interpretation strategies were further divided into three scoring focuses, which were language, organization, and content. Table 4.14 describes them in detail.

Table 4.14

Korean English Teachers' Scoring Behaviors

Strategies			Scoring behavior	Description
Judgement Strategies	Language focus	1	Assessing linguistic correctness	·Assessing grammatical correctness or the use of expressions in context ·Discerning between local and global error
	Organization focus	2	Assessing textual structures	·Assessing the text organization, length, and paragraphing
	Content focus	3	Assessing topic development	·Assessing the way how the ideas are developed. ·Usually measured by degree of the elaborated description of instances, creativity of the content, or whether it shows autobiographical development.
Macro-judgement Strategies		4	Creating scoring rubric	·Devising the rater's own scoring rubric
		5	Comparing with other essays	·Referring to the previous essays when making scoring decisions ·Changing the first scoring decision
Inter-pretation Strategy	Language focus	6	Commenting on linguistic features	·Commenting on grammatical correctness and expression use ·Commenting on orthographic features
	Organization focus	7	Commenting on textual organization	·Commenting on text organization, length, the use of title, or paragraphing
	Content focus	8	Reading and interpreting the essay	·Reading aloud the lines and giving Korean interpretations
		9	Summarization	·Giving a brief summary about the

			essay ·Confined to "content" related comments
	10	Commenting on topic development	·Concerned with coherent development of the content ·Including comments on creativity or elaborated description of life
	11	Commenting on Cohesion	·Considering the sentence level connection (e.g., the use of linker or the connectedness of the two sentences)
	12	Making an overall impression	·Commenting on the first/overall impression of the essay ·Making personal judgements/engagements with the writer

In what follows, the way each scoring strategy and behavior was constructed will be explained with the participants' verbal reports.

4.3.1. Judgment and Macro-Judgement Strategies and Scoring Behaviors

Judgement strategies indicate the raters' evaluation strategies for formulating a score and these are further categorized into three focuses: language, organization, and content.

For starter, "language focused judgement strategy" includes the first scoring behavior "assessing linguistic correctness." This is to judge grammaticality and the appropriateness of expression use in context. Also, it

encompasses judging if the errors are local or global errors. Next, "organization focused judgement strategy" contains the scoring behavior of "assessing textual structures," which is to rate the quality of the essay in terms of the number of paragraphs, use of indentation, or length of the essay. Third, "content focused judgement strategy" contains "assessing topic development," which is usually measured through the degrees of concrete descriptions of instances, creativity of the content, or whether it shows autobiographical development.

Macro-judgement strategies are defined as the behaviors that the raters exhibited when deciding how to deal with the overall set of compositions" (Cumming et al., p. 72). More specifically, these refer to the raters' establishing specific criteria on which they base their judgements and the researchers claimed that "the raters with extensive experiences in assessing EFL compositions adopted macro-judgement strategies to guide their judgements and confirm their scoring criteria" (Cumming et al., p. 75). In this current study, the fourth and fifth scoring behaviors were thought to be included in macro-judgement strategies. The fourth scoring behavior, "creating scoring rubric," refers to coming up with the rater's own specific scoring guidelines and the fifth scoring behavior "comparing with other

essays" is to change the initial judgements after reading other essays. Excerpt 19 illustrates the fifth scoring behavior.

Excerpt 19. Rater D, 4th Essay

> The previous essay vividly illustrated what is going to happen in the future. The writer mentioned instances such as running a restaurant, her business getting popularity, and opening restaurant branches at the age of 64. I can give five points for the previous one for including these specific examples. However, in this case, this essay focused too much on "going to university," which is not likely to be memorable at the age of 80. So, I will deduct one point for content part.

In what follows, the specific scoring behaviors shown when raters interpret the quality of essays are going to be explained.

4.3.2. Interpretation Strategies and Scoring Behaviors

Interpretation strategies are used for the aim of comprehending the

compositions and these are further divided by three focuses: language, organization, and content focus. First, “language focused interpretation strategy” includes the sixth scoring behavior, “commenting on linguistic features.” To illustrate, it refers to the comments on grammatical correctness, expression use, and orthographical features. Next, “organization focused interpretation strategy” includes the seventh scoring behavior, “commenting on text format,” which is concerned with mentioning text organization, length, use of title, or paragraphing.

Lastly, “content focused interpretation strategies” contain five scoring behaviors. The eighth scoring behavior “reading and interpreting the essays” indicates the raters’ reading aloud the essays and giving Korean interpretations to it. This strategy is shown as the most fundamental behavior to get the scoring processes started. The ninth scoring behavior is to give a brief summary of the essay and only content-related comments were counted as this behavior. In other words, a summary of the overall grammaticality of the essay was excluded. The tenth scoring behavior, “commenting on topic development” deals with the overall development of the essay. This is rather a broad concept, since it includes comments on creativity, coherence, specific description of life events, and judging if the essay covers the writer’s whole

life. Excerpt 20 shows rater I's thoughts on topic development. This rater gave credits for time sequenced topic development.

Excerpt 20. Rater I, 2nd Essay

> The life events in this essay are organized chronologically. This student wrote that she went to a specialized high school, opened a bakery at the age of 20, and did other things two years later. Putting her ideas in a time order makes this easy to read.

Also, the concerns about creativity and elaborated description of life events can be observed in excerpt 21.

Excerpt 21. Rater I, 10th Essay

> I can see the logical flow in this essay, but the life events are not described concretely enough. Also, it is limited in creativity. This one could have written better with some

detailed examples, I guess.

The eleventh scoring behavior “commenting on cohesion” is to consider the sentence level connections such as the use of transition words or linkers. Lastly, the twelfth scoring behavior “making an overall impression” contains several types of comments: commenting on the first impression by scanning the whole composition, making personal engagement, and giving an overall judgement at the end of the assessment.

So far, the key 12 scoring behaviors of Korean English teachers were described by three strategies with the participants’ excerpts. To be more specific, the configuration of scoring behaviors supports Wolfe’s (1997) model of scorer cognition as it explains how raters perceive the quality of the essay influences their processing actions. That is, interpretation strategies influence judgement/macro-judgement strategies and also, judgement/macro-judgment strategies affect back to interpretation strategies as well.

In the end, these findings are expected to help explain the aspects of writing the raters more attended to and discover their internalized scoring processes when arriving at the scoring decisions.

4.4. Discussion

The current study explored a total of nine Korean English teachers scoring processes through think-aloud protocol analysis. In the previous sections, descriptive accounts of the participants' verbal reports were presented along with the key 12 scoring behaviors found commonly across them. In this section, the participants' scoring tendencies are going to be discussed in two aspects: the effects of the length of teaching career on individual differences in scoring behaviors and the degree of consistency between the perceived significance of writing constructs and the actually weighted constructs.

For starter, to see if the effects of the length of teaching career on individual differences in scoring behaviors, Table 4.15 shows all raters' scoring characteristics by the length of their teaching career. Also, each rater's individual scoring differences were thought in terms of the scoring focuses and the demonstration of expert raters' scoring behaviors. More specifically, in this study, expert raters' scoring behaviors were summarized by three scoring characteristics: making a scoring rubric, showing personal engagement with the writer, and displaying the initial scanning.

Table 4.15

All Participants' Background Information and Scoring Behaviors

Rater		D	B	E	A	C	I	F	H	G
Length of teaching experience		0.8 years	Less than 1 year	2 years	3 years	3.9 years	5 years	6 years	7 years	11 years
Perceived significance of writing constructs		Compre-hensibility	Compre-hensibility	Grammar, content	Content	Coherence, grammar, vocabulary	Content, creativity	Grammatical correctness	The variety use of vocabularies, idea expression	Setting a rubric, objective scoring
Focuses of scoring		Language, content, organization	Language, content	Language	Language, content, organization	Language	Language, content	Language, content	Language, content, organization	Language content, organization
Characteristics of expert raters	Making a scoring rubric	Yes	No	No	No	No	Yes	No	No	Yes
	Personal Engagement with the writer	No	No	No	No	No	Yes	No	Yes	No
	Initial scanning	No	No	No	No	No	No	Yes	Yes	No

Note. The raters were ordered by the length of teaching experience.

The above table shows that the raters were not much differentiated in terms of the focuses of scoring. That is, raters except for C and E took language and content into consideration when judging the quality of the essays and four raters reflected textual organization to scoring decisions. Here, the participants' experiential backgrounds do not seem to have played a part.

However, when it comes to the demonstrations of expert raters' scoring behaviors, the raters' differing length of teaching career seemed to have contributed to individual differences. Except for Rater D, who showed the scoring behavior of "making a scoring rubric," those raters with more than five years of teaching experiences were more likely to display those behaviors.

For the second discussion point, only Rater G exhibited high degree of consistency between the perceived significance of the writing constructs and actually weighted writing construct. To illustrate, Rater G reported that "setting a scoring rubric and adopting objective scoring procedures" were important on the preliminary survey and she came up with a scoring rubric during scoring processes.

However, except for her, the other raters manifested discrepancies between their beliefs and the focuses of scoring. In case of Raters A, B, D, F, H, and I, the focuses of scoring outnumbered the writing constructs they

perceived important. That is, they mostly reported one construct, either content or grammar, is important before starting the scoring processes, but during the actual scoring processes, more than two constructs were reflected into scoring decisions. Meanwhile, Raters C and E mentioned that both content and grammar were important traits, but "language" was the sole scoring focus of their scoring processes.

For now, the reasons for these discrepancies could not be explicitly explained due to insufficient information about the participants. Future studies involving larger participants and thorough retrospective interviews with them are expected to explain this matter.

Chapter 5. Conclusion

The present study has sought to identify the strategies that are used by Korean English teachers while scoring written compositions. Section 5.1 summarizes the major findings of the study based on the results of qualitative analysis reported in Chapter 4. Then, Section 5.2 will discuss pedagogical implications this study brings in and lastly, Section 5.3 concludes the thesis by reporting the limitations and making suggestions for future research.

5.1 Major Findings

The guiding question for this study was "What strategies do the Korean English teachers often use when making scoring decisions?" To address this research question, a qualitative analysis of think-aloud protocols was utilized. The analysis of think-aloud protocols revealed the following scoring behaviors across the raters.

First, Korean English teachers took grammar, content, and organization into account when measuring the quality of the essays. In more detail, grammar played the greatest role in deciding the score than the other two constructs. When it comes to content, the task completion and elaborated description were considered important. As for organization, the length of the

essay or paragraphing were reflected to scoring decisions.

Second, the raters with longer experiences in writing assessment exhibited the characteristics of expert raters' scoring behaviors. They were more likely to demonstrate scanning for the whole text, making a scoring rubric, and showing personal engagement with the writer than less experienced raters. From this, it was assumed that the amount of experiences in writing assessment could be a meaningful variable in accounting for the individual differences in scoring processes.

Overall, this study complements the existing studies about EFL English teachers' scoring processes. It explored the aspects of writing the participants more attended to and explored the grounds on which they make scoring decisions. Ultimately, this study is expected to shed some light on the development of scoring rubrics for writing assessment and help rater trainers to develop reliable and fair scoring protocols.

5.2 Implications of the Findings

The discussion of the Korean English teachers' scoring behaviors provides meaningful implications to writing assessments in classroom environment. First, norming and training sessions are necessitated for the

settlement of reliable and valid implementation of writing assessments. The raters were varied in the aspects of writing constructs they attended to and the weight they assigned. Also, they showed a certain degree of discrepancy between the perceived significance of writing constructs and the actual scoring focuses. Through the rater discussion, individual raters' differing expectations and personal criteria can reach a consensus, and this would eventually ensure a higher degree of inter-and intra-rater reliability.

Second, teachers should be more like an instructor or a reader of essays, instead of a rater. Wiseman (2012) put that the EFL teachers frequently feel pressure to prepare for the test and meet the criteria outlined in the rubric. In this current study, the teachers' tendency to treat the writing assessment as a test was obviously witnessed. However, for the pursuit of the learners' improvement in second language writing ability, the EFL teachers should shift their focus from detecting errors toward seeing the writer's self-expression and communication with the audience. In the end, the shift of the rater's perspective could consolidate the formative role of assessment, which supports the development of second language writing skills in the classroom.

5.3 Limitations and Suggestions for the Future Research

The current study revealed Korean English teachers' key scoring behaviors and provided insights into EFL teachers' conception of the writing assessment. However, this study is not free from limitations. First, the size of the participant sample of this study was relatively small, largely due to the qualitative nature of the research methodology. Consequently, the results of the study should be interpreted with caution. Future research involving a larger sample size would help verify the validity of the current research.

Next, the task of a think-aloud protocol may have caused additional cognitive burden on the cognitive demands of completing the scoring task. Also, the participants could have had differing levels of familiarity with or dedication to the think-aloud task. Therefore, the results of the current study should be interpreted with such factors into account. Future research is recommended with more training provided to participants so that they become highly familiar and competent with the think-aloud tasks.

Despite these limitations, this study, together with future research efforts in exploring scoring processes would aid in set up a more reliable and valid writing assessment systems in classroom environment.

References

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly, 7*(1), 54-74.

Beck, W., Llosa, L., Black, K., & Trzeszkowski-Giese, A. (2015). Think-aloud as a diagnostic assessment tool for high school writing teachers. *Journal of Adolescent & Adult Literacy, 58*(8), 670-681.

Bejar, I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice, 31*(3), 2-9.

Cho, D. (2014). Investigating EFL writing assessment in a classroom setting: Features of composition and rater behaviors. *The Journal of Asia TEFL, 5*(4), 49-84.

Cho, J. (2003). *Native and non-native English raters' use of assessment criteria in the evaluation process of English compositions by Korean high school students* (Master's thesis). Retrieved from https://primoapac01.hosted.exlibrisgroup.com/primo-explore/search?query=any,contains,%EC%A1%B0%EC%9E%90%EB%A3%A1%20%EC%98%81%EC%96%B4%EA%B5%90%EC%9C%A1&tab=all&search_scope=ALL&vid=82SNU&lang=ko_KR&of

fset=0

Conner-Linton, J. (1995). Looking behind the curtain: What do L2 composition ratings really mean? *Teachers of English to Speakers of Other Languages, 29*(4), 762-765.

Crisp, V. (2008). Exploring the nature of examiner thinking during the process of examination marking. *Cambridge Journal of Education, 38*(2), 247-264.

Crisp, V. (2012). An investigation of rater cognition in the assessment of projects. *Educational Measurement: Issues and Practice, 31*(3), 10-20.

Cumming, A. (1990). Expertise in evaluating second language composition. *Language Testing, 7*(1), 31-51.

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

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

Eckes, T., Muller-Karabil, A., & Zimmerman, S. (2016). Assessing Writing. In Editors-Banerjee, J., & Tsagari, D (Eds.), *Handbook of second language assessment* (147-164). Retrieved from https://www.researchgate.net/publication/304743555_Handbook_of_Second_Language_Assessment/link/5778fb2308aeb9427e2bfaa1/download

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*(2), 93-112.

Ericsson, K., & Simon, H. (1980). Verbal reports as data. *Psychological Review, 87*(3), 215-251.

Freedman, W., & Calfee, C. (1984). Understanding and comprehending. *Written Communication, 1*(4), 459-490.

Huot, B. (1993). *The validity of holistic scoring: A comparison of the talk-aloud protocols of experienced and novice holistic raters.* (Doctoral dissertation). Available from ProQuest Dissertations and Thesis Global (Order No. 8817872)

Kim, J. (2017). Improving the validity of L2 performance assessments: Use

of many-facet Rasch measurement. *Studies in Foreign Language Education, 31*(3), 277-297.

Kim, S. (1999). A study on marking behaviors in performance assessment of English writing. *Journal of Educational Evaluation, 12*(2), 25-54.

Knoch, U. (2009). *Diagnostic writing assessment: The development and validation of a rating scale.* Frankfurt am Main: Peter Lang.

Kobayashi, H., & Rinnert, C. (1996). Factors affecting composition evaluation in an EFL context: Cultural rhetorical pattern and readers' background. *Language Learning, 46*(3), 396-437.

Lee, S., & Choe, H. (2012). Rating of Korean students' L2 writing: Similarities and differences between native and non-native raters. *Journal of Research in Curriculum and Instruction, 16*(3), 629-655.

Lintunen, P., & Makila, M. (2014). Measuring syntactic complexity in spoken and written learner language: Comparing the incomparable? *Research in Language, 12*(4), 377-399.

Lumley, T. (2005). *Assessing second language writing: The raters' perspective.* Frankfurt: Peter Lang.

McNamara, T. (1996). *Measuring second language performance.* Harlow,

Essex, UK: Addison Wesley Longman Ltd.

McNamara, F., & Lumley, T. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*(1), 54-71.

Milanovic, M., Saville, N., & Shuhong, S. (1996). *Performance testing, cognition, and assessment: Selected papers from the 15th language testing research colloquium, Cambridge and Arnhem*. New York: Cambridge University Press.

Panos, L. P. (2015). *Defining and operationalizing propositional complexity into idea units: Effects of mode, discourse type, task type and task complexity.* (Unpublished maters' thesis). Universitat de Barcelona, Spain.

Park, J. (2006). The study on the rater reliability of three scoring methods in assessing argumentative essays: Holistic, analytic, and multiple-trait scoring methods. *Foreign Language Education Research, 9*(1), 63-84.

Sakyi, A. (2000). Validation of holistic scoring for ESL writing assessment: How raters evaluate ESL compositions. In A. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers*

from the 19th language testing colloquium (pp. 129–152). Cambridge: Cambridge University Press.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*(4), 465-493.

Shaw, S. & Weir, C. (2007). *Examining writing: Research and practice in assessing second language writing*. Retrieved from https://www.researchgate.net/ publication/2395-88606_Examining_Writing_Research_and_Practice_in_Assessing_Second_Language_Writing_Studies_in_Language_Testing_26

Shin, K. (2015). Professional native- and non-native English speaking raters' evaluation of native- and near-native students' English writing. *Journal of the Research Institute of Korean Education, 33*(4), 401-421.

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp.111-125). Norwood, NJ: Ablex.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing, 10*(3), 305-319.

Wiseman, C. S. (2012). Rater effects: Ego engagement in rater decision-making. *Assessing Writing, 17*(1), 150-173.

Wolfe, W. (1997) The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing, 4*(1), 83-106.

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

Appendices

Appendix 1: Students' Essays

The first essay

Writing My Autobiography

Imagine that you are 80 years old and write an bioautography looking back your past.

Title:

I could not concentrate on my studies untill high school.
Because I liked the game. So I entered an ordinary college.
But I could have many good friends.
Time went by and I graduated the college. I felt happy
to be adult one hand but was worried about my future
I tried to do many things but I could not find what I
wanted. So I regretted my school days.
One day I tried to broadcast about games and my life.
At first, My broadcast was not popular. I was in despair
but never gave up and kept my broadcast style
So, more and more people got interested in my broadcast. I could
be popular soon.
I traveled the tourist sites with my broadcast friends.
Meanwhile, I tried to keep my broadcast style.
Now that I am 80 years old, I published a book of my
life. I satisfied with my life.

The second essay

Writing My Autobiography

Imagine that you are 80 years old and write an bioautography looking back your past.

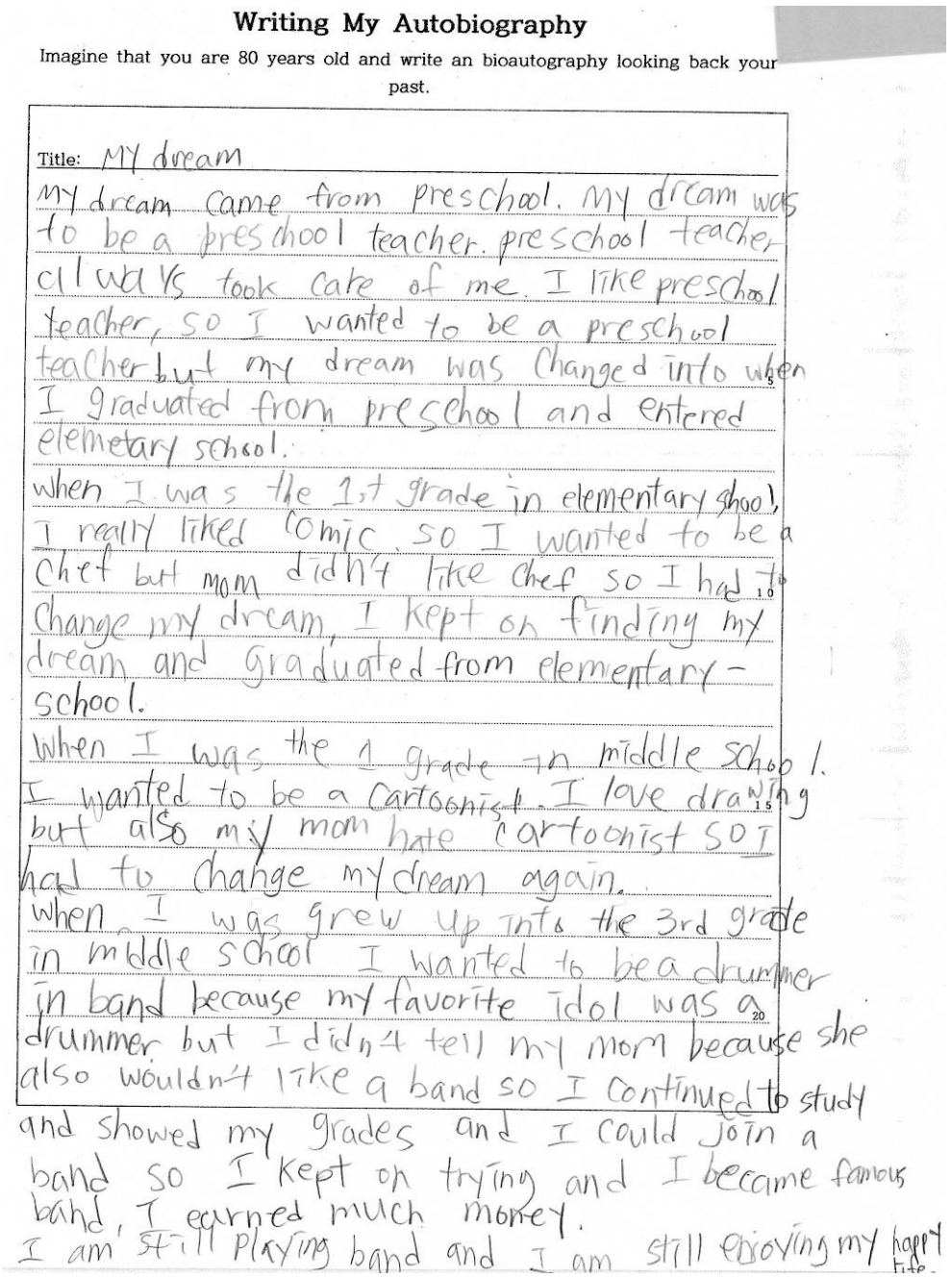
Title: My dream.

I was born on December 6th in 2004. I was the second of three children in my family. When I was an Elementary school student, my dream was being a fashion designer. But I was not good at drawing. So I changed my dream into a chef. I went to the specialized high school for cooking. When I was 20 years old, I worked at a bakery.

Two years later, I went to France to study cooking. When I was 25 years old, I got a job at the restaurant. One day, I participated in the cooking contest, and I won the first prize. Ten years later, I ran my own restaurant. I succeeded my business, which was known to the World. I had many restaurant's branches in the World. I retired when I was 64 years old, And I have been living with my three cats in my three-story building.

I am happy and enjoy my life at 80.

The third essay



Writing My Autobiography

Imagine that you are 80 years old and write an bioautography looking back your past.

Title: My dream

My dream came from preschool. My dream was to be a preschool teacher. preschool teacher always took care of me. I like preschool teacher, so I wanted to be a preschool teacher but my dream was changed into when I graduated from preschool and entered elemetary school.

when I was the 1st grade in elementary school, I really liked comic. so I wanted to be a chef but mom didn't like chef so I had to change my dream, I kept on finding my dream and graduated from elementary school.

When I was the 1 grade in middle school. I wanted to be a cartoonist. I love drawing but also my mom hate cartoonist so I had to change my dream again.

when I was grew up into the 3rd grade in middle school I wanted to be a drummer in band because my favorite idol was a drummer but I didn't tell my mom because she also wouldn't like a band so I continued to study and showed my grades and I could join a band so I kept on trying and I became famous band, I earned much money.

I am still playing band and I am still enjoying my happy life.

The fourth essay

Writing My Autobiography

Imagine that you are 80 years old and write an bioautography looking back your past.

Title: My life

I will write down my life. Since I was an only child, I had many benefits I raised three puppies and I could read books as much as I wanted

There are many memorable things in my life The first thing is that I entered Seoul National University. When I was a middle school student, the only thing that made me frustrated was maths However, I never gave up. I struggled to understand basic rules and solve lots of difficult problems As a result, when I was in high school, I could get better grades

The second thing is that I traveled around the world after graduating university. I traveled just close countries such as Japan and Singapore during childhood I dreamed to visit further worlds. The best country I've ever been is U.S It was like the heaven to me.

Lastly, I tried to find out something to suit me although I thought I wasn't talented in any field. Finally, I could find what I really wanted I would like to say "All your dreams can come

true if you have the courage to pursue them."

When I look back all, it doesn't look special. But I'm so proud of what I've done because I made every efforts to achieve my goals. Actually, I'm still trying to do my best for the rest of my life.

The fifth essay

Writing My Autobiography

Imagine that you are 80 years old and write an bioautography looking back your past.

Title: My life

I was born on April 20th in Seoul. When I was young, I lived in Seodaemun-gu. I moved to Sangam-dong when I was 7years old, and I entered to Sanji elementary school.

I met good friends and teachers at school. I went to America when I was 4th grade and so I entered to American elementary school. I came back to Korea at the end of a 6th grade. I graduated Sanji elementary school and I entered to Sungam middle school. In there I tried to find what I wanted to do in the future and I decided to go to the Korea military academy, so I studied very hardly. Finally I entered military school and I graduated as a top student. I became a soldier.

I was called Second MacArtur in military. After I retire, I'm spending happy life with my wife and my kids in very big house, which is in Gangnam.

170

The sixth essay

Writing My Autobiography

Imagine that you are 80 years old and write an bioautography looking back your past.

Title: My happy life.

When I was young, I lived in a happy family, and grew up nicely. My life was ordinary.

I had a great grade at korean S.A.T, so I could choose among lots of good university. I thought long time and finally chose sedgang university, which is good school in seoul. I spent great time in my university. After I graduated from university, I took an interview. Luckily, I passed the interview and entered Hyundae. I tried hard to learn many skills, but what I have been doing was not right for me. So I quitted the Hyundae when I was 32 years old. I built a small car design company and my company got popularity. Soon company's sales was increased and I became a famous company's CEO.

Now my company has become one of the best car design companys, and I am also healthy. so I still manage my company, I manage my company as long as possible

The seventh essay

Writing My Autobiography

Imagine that you are 80 years old and write an bioautography looking back your past.

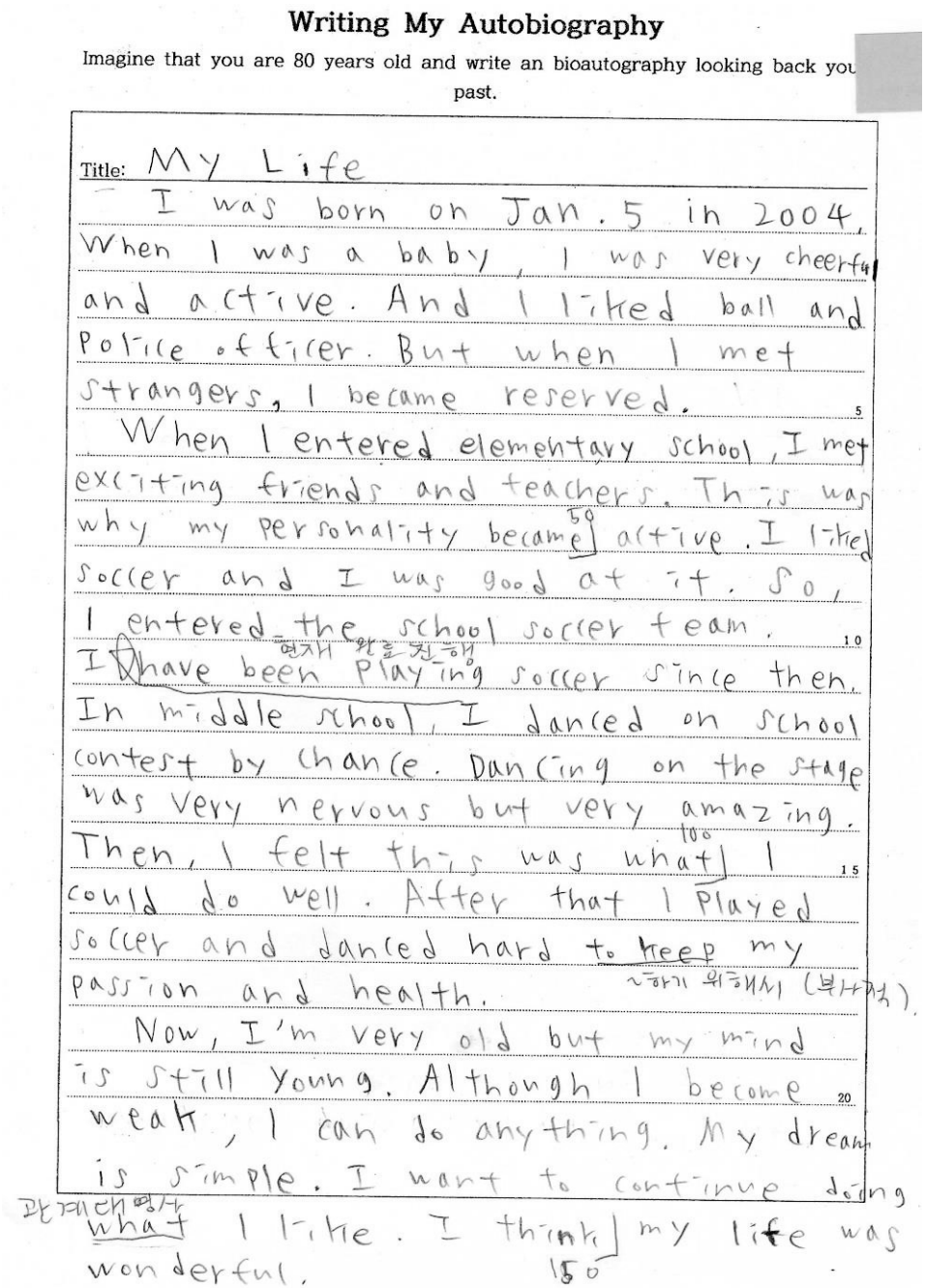
Title:

I was born on March 16, 2004, in Ilsan When I was young, I liked to help my friends I once helped my classmate who broke his ankle in P.E class I was also careful about other people's feelings. So when people around me got sad, I often tried to make the atmosphere brighter. I loved playing basketball, and I learned how to play as a team helping my team mates.

When I entered high school, I studied hard to do what I want in the future In 2023, I entered the medical school and studied for 6 years to become a surgeon While I was studying at university, I met a wonderful woman, and got married. I raised 2 wonderful boys. I loved them so much. They both became docters. When I was 40, I built my own hospital. I worked hard to make my patient's life better and healthier. I also read books about science and literature regularly to be smarter After my sons graduated high school, I traveled around the world with my wife, helping people who need medical help desperatly

I'm now 80 years old, and I have been writing books about my life I wrote 3 books, which all become best-sellers. Although my life may seem normal as a docter, I never gave up and enjoyed my life.

The eighth essay



Writing My Autobiography

Imagine that you are 80 years old and write an bioautography looking back you past.

Title: My Life

I was born on Jan. 5 in 2004. When I was a baby, I was very cheerful and active. And I liked ball and Police officer. But when I met strangers, I became reserved.

When I entered elementary school, I met exciting friends and teachers. This was why my personality became active. I liked soccer and I was good at it. So, I entered the school soccer team. I have been Playing soccer since then. In middle school, I danced on school contest by chance. Dancing on the stage was very nervous but very amazing. Then, I felt this was what I could do well. After that I Played soccer and danced hard to keep my passion and health.

Now, I'm very old but my mind is still young. Although I become weak, I can do anything. My dream is simple. I want to continue doing what I like. I think my life was wonderful.

The ninth essay

Writing My Autobiography

Imagine that you are 80 years old and write an bioautography looking back your past.

Title: If I am 80

I was born on December 9, 2004, in Korea. My family is my mom, dad brother and a cute dog. I was graduated from Myung-ji elementryschool, Saegam middle school, Sejong Academy of Science and Arts and finally KAIST university.

I have been living as an ordinary beat maker. 60 years ago, a famous rapper asked me to make a beat for him. I accepted his request and started to work for him. As I earned money, I could set up a club. One day, a foreigner rapper who visited my club suggested that I work for a club in United States. So, I went to the United States. When I was working as a D.J in the club, an NBA basketball team coach dropped by the club. He listened to the beat I made and asked me to become a DJ at the NBA basketball team home stadium.

After working in the NBA basketball team for 10 years, I went to Hawaii. I bought a three-story house and lived alone in that house. And by the many beats I made, I am in the Guinness Book of Records.

After living in Hawii for 20 years, I returned to Korea and met my family. Since the I have been making my own bucket list until I die.

I have never regretted living my life as a beat maker and a DJ. I haved loved myself and tried to overcome difficulties and hardships. Anybody doesn't knows when I'll die. It could be a little bit scary but I don't care. I haved loved myself, I do love myself, and I will love myself. I loved what I did!

The tenth essay

Writing My Autobiography

Imagine that you are 80 years old and write an bioautography looking back your past.

Title: Meaningful life

I was born on September 25, 2004, in Seoul, in a happy family. I can't clearly remember now, but I remember that I walked along the river in front of my house with my family everyday. I went to English kindergarten, went to elementary school. I met lots of friends in elementary school, and still maintaining friendship with who I met at that time. Later, I entered middle school and highschool. I had a talent in studying and finally graduated highschool with a perfect score. After graduating highschool, I went on a trip alone to California. With this trip, I got a dream of living in America. My childhood life was normal but filled with happiness.

My dream was living and working in America. However I did not have enough grade to go to an American university. I did not give up my dream and went to Yonsei university and went to UCLA as an exchange student. There, I got a chance to take part in Google's internship program. Finally, I became a real employee of Google. I was so happy to achieve what I have dreamed since I was young. I worked very hard with passion but one day, when I was 45, I got bored of working everyday. So I walked out of Google and started to travel around the world.

I am 80 years old, weak grandmother now. When I flash back through my life, I don't regret anything. I don't regret my choice to enter Google, and I don't regret my choice to travel around the world. I am old but I will do my best to live a meaningful life until I stop breathing.

국 문 초 록

한국인 영어교사들의 중학생 영어 작문 채점 시 보이는 의사결정 과정 연구

강민희
외국어교육과 영어전공
서울대학교 대학원

본 연구는 한국인 영어 교사들이 중학교 학생들의 영어 작문 채점 시 보이는 의사결정 과정을 밝히고자 실시되었다. 9명의 한국인 중고등학교 영어 교사들이 10명의 학생들의 글을 총체적 방식으로 채점하였으며 동시에 사고구술을 진행하였다.

사고구술 결과를 분석한 결과 교사들은 공통적으로 문법성 판단에 가장 큰 비중을 들였으며, 글의 내용을 평가할 때는 구체적인 묘사를 가장 중요시하는 것을 알 수 있었다. 또한, 경력이 높은 교사일수록 선행연구에서 밝혀진 전문가 평가자의 특성을 보일 가능성이 높다는 것이 드러났는데, 그것은 글쓴이와의 교감, 평가표를 만드는 것, 평가 시작 전 글 전체를 훑어보는 것이었다.

이 연구를 통해 한국인 영어 교사들이 글쓰기 채점 시 보이는 개별적 차이를 알 수 있었으며 앞으로 교사 연수자들이 신뢰도 있고 타당도 높은 채점 기준을 설정할 때 발판이 되는 시사점을 제공할 수 있을 것으로 기대된다.

주요어: 쓰기평가, 채점과정, 채점자 인지, 채점 행동, 채점 과정, 한국인 영어교사

학번: 2017-23606