교육학석사학위논문

Effects of Korean English Teachers' Perceived Criterion Importance on Scoring Behavior in L2 Writing Assessment

한국인 영어 교사의 채점 기준에 대한 중요성 인식이 영작문 채점 행동에 미치는 영향 분석

2023년 8월

서울대학교 대학원

외국어교육과 영어전공

이 영 주

Effects of Korean English Teachers' Perceived Criterion Importance on Scoring Behavior in L2 Writing Assessment

by

Young-Joo Lee

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

August 2023

Effects of Korean English Teachers' Perceived Criterion Importance on Scoring Behavior in L2 Writing Assessment

한국인 영어 교사의 채점 기준에 대한 중요성 인식이 영작문 채점 행동에 미치는 영향 분석

지도교수 소 영 순

이 논문을 교육학 석사 학위논문으로 제출함
2023년 6월

서울대학교 대학원
외국어교육과 영어전공
이 영 주

이영주의 석사학위논문을 인준함
2023년 7월

위 원 장 ______이 용 원______

부위원장 ______김 기 택______

위 원 ______소 영 순______

Effects of Korean English Teachers' Perceived Criterion Importance on Scoring Behavior in L2 Writing Assessment

APPROVED BY THESIS COMMITTEE:

____

YONG-WON LEE, COMMITTEE CHAIR

____

KITAEK KIM

____

YOUNGSOON SO

ABSTRACT

Effects of Korean English Teachers' Perceived Criterion Importance on Scoring Behavior in L2 Writing Assessment

Young-Joo Lee
English Major, Dept. of Foreign Language Education
The Graduate School of Seoul National University

The advent of Communicative Language Teaching has placed an emphasis on performance-based assessments to assess the ability to use a language. What distinguishes performance-based assessments from multiple-choice questions is the presence of the rating rubric. That is, how raters perceive and apply the rating scale plays a significant role in the evaluative process. Therefore, there has been a wide body of research investigating the interaction between raters and rating criteria, which aimed to enhance the validity of the performance tests by reducing the inconsistency of rating on the part of raters. Previous studies examining rater effects descriptively analyzed the rating criteria to which raters displayed more severity or leniency. However, few attempts have been made to understand the reason behind rater idiosyncrasy from a cognitive perspective. Hence, it is worthwhile to investigate how rater perception of the rating criteria can affect scoring behavior.

The purpose of the present study is to examine how perceived criterion importance can influence scoring behavior. Exploring the relation between rater perception of the rating criteria and scoring profiles will contribute to a better understanding of rater cognition, which in turn can help raters to be equipped with a more balanced view of rating criteria.

For this study, thirty Korean English teachers working at middle and high schools participated in the survey in which they were to indicate the importance of five rating criteria, Content, Organization, Vocabulary, Language use, and Mechanics. Participants also rated thirty writing compositions chosen from YELC (Yonsei English Learner's Corpus). Employing Many-facet Rasch measurement and Hierarchical Clustering, two types of raters were formed: Cognitive Rater Types, which were based on perceived criterion importance, and Operational Rater Types, which were derived from criterion-related biases. These two rater types were compared to analyze the relationship between rater perception and rating behavior.

The finding was that five Cognitive Rater Types (CRTs) and six Operational Rater Types (ORTs) were created. As all ORTs were composed of raters from different CRTs, it was not possible to investigate direct relationships between CRTs and ORTs. Therefore, based on the analysis of the mean bias measure in relation to the mean criterion importance rating from raters who belonged to the same ORT, but came from different CRTs, it was found that the effect of criterion importance on scoring behavior varied depending on the rating criteria involved. In Content and Mechanics, both severity and leniency bias were observed, but how biases were combined with criterion importance varied between these two rating criteria. In Content, severity bias was shown to be combined with the criterion importance higher than the average criterion importance whereas leniency bias was shown to be aligned with the criterion importance lower than the average criterion importance. However, in Mechanics, this pattern was reversed, revealing that severity bias was combined with less importance than the average criterion importance while leniency bias was aligned with higher importance than the average criterion importance. The disparity between Content and Mechanics was also identified in the analysis of the data from individual raters.

Although the study has a limitation in that participants were not trained raters for English writing assessments, the study′s endeavor to connect rater cognition to scoring behavior will help to expand the scope of the study researching rater effects.

Key Words: rater effects, perceived criterion importance, writing assessment, rater types, rater cognition, performance-based tests

Student Number: 2021-28726

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

This study aims to examine the relation between rater perceptions of the criterion importance and rater scoring behavior in assessments of writing performance. This chapter presents the purpose and organization of the thesis. Section 1.1 discusses the background and purpose of the study, followed by research questions in Section 1.2. Section 1.3 describes the organization of the thesis.

1.1 Background and Purpose of the Study

Throughout the history of language assessment, many endeavors have been made to accurately measure the language ability of test takers. The emergence and predominance of the Communicative Language Teaching (CLT) approach marked a dramatic shift in prevailing attitudes not only toward language teaching but also toward language testing. Previously popular traditional formats such as multiple-choice items were criticized as insufficient for providing accurate information about what test takers can and cannot do. As a result, moves toward performance-based measures in which test takers are asked to demonstrate proficient use of newly learned skills in an evaluative context emerged and remain increasingly embraced. At the core of these performance-based tests is the demand that test takers construct their responses by writing or speaking as a way to demonstrate their language ability (Bachman & Savignon, 1986; McNamara, 1996; Shohamy, 1995).

In addition to their differing formats, constructed-response performance tasks are differentiated from traditional testing methods (e.g., multiple-choice questions, T/F questions, and matching) in that scoring is mediated by human raters. As a result, the touted strength of performance-based testing is often overshadowed by claims of reduced objectivity in contrast to former modes of assessment. Indeed, it is often the case that rater-induced errors are involved in the process of judging and interpreting test takers' performances, thus undermining the validity of tests (Engelhard, 2002; Engelhard, Wang & Wind, 2018).

There are many sources of rater-related bias based on the types of interactions among the components that consist of test contexts: raters to test takers, raters to task types, raters to rating criteria, and raters to rating time. Raters, as humans, bring different experiences, beliefs, and cognition with them to the rating context, and, therefore, may be prone to assigning test scores, which may not be relevant to the test construct due to previously established, oftentimes unconscious biases. These biases come from a variety of sources including the rater's personal teaching and learning philosophies, previous training and rating experiences, beliefs about what constitutes good writing, and influences from the rater’s L1. Among these rater-related factors, the focus of the present research is to investigate the interaction between raters and rating criteria based on the assumption that differentially perceived importance attached to rating criteria will impact rating behavior in L2 writing assessments. Rating writing performances is essentially a complex cognitive process on the part of raters, which is mediated through the interactions with rating criteria. This, therefore, suggests that different perceptions and the resultant application of rating criteria may cause rater variation, thus affecting test scores and eventually threatening the validity of the assessment (Lumley, 2002).

How raters in L2 writing performance-based assessments differ in the application of rating criteria has been examined in a wide body of research (Engelhard, 2002; Paula & Huot, 1993; Wolfe, Kao, & Ranney, 1998). However, a majority of studies on rater bias to date have investigated the scoring patterns of raters, only revealing that raters vary in scoring profiles in relation to particular rating criteria. However, the question that still remains unanswered is what induces criterion-related bias from cognitive perspectives, thus leading the present study to advance the hypothesis that there is a link between perceived criterion importance and bias toward particular criteria.

Hence, responding to the gap in the previous literature on rater bias, the present study aims to bridge the gap between rater cognition and scoring behavior, which is based on the comparison of two different rater types; one is to be formed by perceived criterion importance and the other by criterion-related bias.

1.2 Research Questions

To address the inadequacy in the existing knowledge as to the nature of the link between raters' perceived criterion importance and their scoring behavior, the study poses the following research questions:

1. How can raters be classified into a group based on perceived criterion importance?
2. How can raters be classified into a group based on severity or leniency toward particular rating criteria displayed in live scoring sessions?
3. To what extent is perceived criterion importance related to scoring behavior?

1.3 Organization of the Thesis

This thesis consists of five chapters. Following this introduction chapter, Chapter 2 reviews the literature on theoretical and empirical research on L2 writing assessments pertaining to the present study. Chapter 3 delineates the research methods concerning participants, instruments, and procedures of the study, explaining data collection processes and quantitative measures adopted for the proposed study. Chapter 4 reports the outcome of the statistical analyses and provides a discussion concerning the research questions. The statistical analyses show how raters can be clustered into groups based on perceived criterion importance and criterion-related bias respectively along with the finding as to the relation between perceived criterion importance and rating behavior. Lastly, Chapter 5 summarizes the research findings, indicating the methodological and practical implications, and concludes with the limitations of the present study and future research suggestions.

CHAPTER 2

LITERATURE REVIEW

This chapter presents the theoretical and empirical research which has relevance to the present study. Section 2.1 explores how performance and language ability have been defined. Section 2.2 reviews research on rater effects focusing on the interaction between raters and rating criteria. Section 2.3 provides research on rater cognition in relation to rating criteria with a focus on Many-facet Rasch measurement analysis and Section 2.4 summarizes the chapter.

2.1 Defining and Assessing L2 Writing Abilities

2.1.1 Traditions of Language Performance Tests

Endeavors to involve performance in second language testing have predated the communicative language approach, which was proposed in the late 1960s and early 1970s. The purpose of language performance tests prior to the advent of theories of communicative competence was to select foreign students who wish to study in English-speaking countries or to teach those who needed to acquire communicative skills in a second language for a vocational purpose. Using the words of Messick (1994), the performance itself was the *target* of assessment. Hence, the early tradition of language performance tests did not pay attention to underlying language knowledge and abilities, the constructs of language tests. The introduction of Communicative Language approach, however, affected second language performance tests to align with theories of communicative competence, which thus resulted in more focus on what can be revealed through performances. In the words of Messick (1994), developments in language teaching influenced by communicative competence theories treated performance as the *vehicle* of assessment. That is, post-communicative traditions of language performance assessment centered on what performances can elicit with respect to underlying language ability and knowledge, thereafter the target of assessment. The two needs, which are selecting students and personnel based on foreign language ability and complying with communicative language theories, have constituted the tradition of second language assessment (McNamara, 1996).

A distinction between two traditions of second language performance tests was noted by McNamara (1996), who named a *strong* and a *weak* sense of the term *second language performance test*. Through the traditions of language performance tests, different weights have been given to task performance on the one hand and language performance on the other. McNamara (1996) offers the distinction here between a “strong sense” of language performance and a “weak sense”. The former values the completion of a task relevant to real-life as most important, with relatively less focus on the various language components elicited. Second language performance tests in the strong sense may or may not involve the assessment of language ability. In second language performance tests in the weak sense, however, the involvement of linguistic aspects is essential since the purpose of the tests is to assess language proficiency through a language sample. To those standing in favor of a weak form, the candidate’s ability to perform the task is not an actual focus since they believe that not performance but language performances can reflect underlying language ability. Language performance tests in a “strong sense” were prevailing even after Hymes’s model of communicative competence (1967).

The weak sense of the second language performance has its source in the work of Lado (1961) (McNamara, 1996). Viewing language through a structuralist framework, Lado argued that language ability consists of separate knowledge about phonology, structure, and lexicon, suggesting that knowledge of different language elements should be integrated into the skills of listening, reading, speaking, and writing. Although Lado’s discrete-point approach is credited with the attempt to understand underlying language ability, it did not take into account actual language use; Lado, thus, could not provide the reason why having grammatical knowledge cannot lead to the actual capacity to apply linguistic knowledge in a communicative setting.

Consistent with the work of Lado, Carroll (1961) argued for using sampled items from a large pool of possible items to assess language knowledge and skills. However, Carroll emphasized that performance components be involved in language testing, recommending that language tests should attend less to structural points or lexicon and more to the total communicative effect of an utterance. Carroll also stated that language tests should be able to predict the possible future use of a language, which seems to pertain to *predictive* validity. In line with Carroll, Davies (1968, 1977) claimed that the virtue of proficiency language tests lies in having the potential to provide evidence of examinees' control over language and predict future performance. Despite the importance given to the actual application of language knowledge in performance, Carroll and Davis did not explain explicitly what performance components in language performance tests should involve.

Clark (1972), another advocate of performance components in language performance tests, stood in favor of a strong sense of language performance tests and emphasized the importance of direct tests as replicating reality in test tasks. The area for reality replicated included "the setting and operation of the tests and the linguistic areas and content which they embody" (Clark, 1975, p.11). Clark stressed that examinees should not be evaluated on the basis of the mastery of the knowledge of vocabulary, morphology, syntax, pronunciation, and so forth, but rather on "the adequacy with which the student can communicate in specified language-use situations" (Clark, 1972, p.120). The notions of performance in language performance tests were established through Clark's assertions. However, as Clark stressed that direct tests should be situation-based, his sole focus on specificity is considered to raise the generalizability problem of language performance tests (McNamara, 1996). In line with Clark, Savignon (1972), who espoused a strong sense of language performance tests, argued for a distinction between linguistic competence and communicative competence and proposed that linguistic skills be assessed while communicating. Savignon stressed that the actual instance and use of language in communication as an indication of communicative competence should be involved in a language performance test. Distinguished from Clark, Savignon involved linguistic resources (e.g., mastery of the pronunciation and the lexical resources) as the elements of successful communication.

Morrow (1979), who also spoke for a strong sense of language performance tests, noted the importance of actual language use. Morrow paid attention to what candidates can achieve by using language, which he called a *performance criterion*. Notable is his attempt to link communicative testing and performance, which rested on seven features of communication. These seven features include interaction, unpredictability, appropriate language forms according to the situational and linguistic context, speaker's purpose, performance, authentic language, and achievement through language. Despite Morrow's explicit commitment to communicative competence, he seemed to stick to the performance-based tradition as reflected in his comment "Performance tests…are of most value in a communicative context." (Morrow, 1979, p. 152).

2.1.2 Theories to Define Constructs of Language Ability

As stated earlier, second language tests affected by pre-communicative traditions seemed to prioritize task completion, not regarding language use samples as a means to infer underlying language ability and knowledge. The focus of such early testers as Clark, Savignon, and Morrow was on how test tasks can be made similar to what candidates will encounter in their real life. However, the issue of the validity of language performance tests arose; though the incorporation of performance components could legitimize language performance tests as a direct method of testing, the question as to the validation of test scores as a true reflection of examinees' language ability was posed. This question was based on the suggestion that other non-linguistic factors may come into play in the evaluative process. Hence, theories about language ability were required to better understand what is involved on the part of test takers, which then could ultimately serve to justify the outcome of language tests as a reflection of examinees' language ability. Testers in favor of a weak sense in the 1980s and 1990s, focused on the specification of language components, thus relying on the elicitation of language performance.

Bachman and Palmer's model of language ability (1996) suggested that language ability should be extended beyond language-exclusive knowledge and ability to include non-linguistic components, such as cognitive competence and affective factors. Their theory of communicative language ability is, therefore, based on the view that language is an integration of various components. This more comprehensive view of language ability has provided a source by which second language performance tests and the inferences from test scores can be justified. Messick (1989) described validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores" (p.13).

Since Chomsky (1965) distinguished between *competence* and *performance*, many scholars have proposed models of language ability. Hymes (1967, 1972) elaborated on the work of Chomsky and advanced a theory of communicative competence. Notable in Hymes' work (1967, 1972) was a distinction between actual language uses in real time and abstract, underlying models of language ability. Hymes further suggested that these underlying parts consist of knowledge (i.e., grammatical and sociolinguistic rules) and capacities (i.e., models and rules of performance) under the term communicative competence. Referring specifically to underlying capacities, Hymes (1972) used the term "ability for use" (p. 283) as the potential to use knowledge in performance and to enable the performance of a particular task. Hymes's notion of performance is thus distinguished from the notion of Chomsky's performance since performance in Chomsky (1965) was understood to denote only actual instances of language use. Hymes, however, saw performance as comprising two parts: a potential realizing speech acts and the actual language use realization. Hymes's work has affected Canale (1983), Bachman (1990), and Bachman and Palmer (1996), all of whom paid attention to the underlying capacity that realizes actual communication. Hymes's notion of the underlying component in performance, the potential to enact the actual language use, was initially excluded by Canale and Swain (1980) but was later elaborated on and extended by Canale(1983), Bachman (1990), and Bachman and Palmer (1996).

A decade after Hymes's communicative competence model (1967), which was developed in reference to the L1 context, Canale and Swain (1980) proposed a communicative competence model for the L2 context. In their model, knowledge of language consists of four components: grammatical competence, sociolinguistic competence, strategic competence, and discourse competence. To illuminate, grammatical competence is composed of the knowledge of linguistic codes including vocabulary, morphonology, syntax, semantics, phonetics, and orthographies. Next, sociolinguistic competence refers to the ability to use knowledge of rules and conventions that determine the appropriateness of utterances in a sociolinguistic and sociocultural context. Strategic competence indicates knowledge of verbal and nonverbal communication strategies utilized for compensating for communication breakdown arising from inadequacy in any other competence. Finally, discourse competence, introduced later by Canale (1983), is concerned with mastery of combining forms and meaning to create a unity of spoken or written context, which is achieved through cohesion and coherence.

The model of Canale and Swain (1980) was criticized for the absence of the interaction among the components of communicative competence (Shohamy, 1988) and the exclusion of the underlying ability to apply language knowledge appropriately in communication (McNamara, 1996; Spolsky, 1985). Canale and Swain’s model of communicative competence (1980) has been influential for decades, dominating the areas of second language assessment and acquisition.

Elaborating on the work of Canale and Swain (1980), Bachman (1990) and Bachman and Palmer (1996) proposed a model of communicative language ability in which multiple components such as personal characteristics, language knowledge, topical knowledge, affective schemata, and strategic competence (metacognitive strategies) are interacting in a specific language use situation or task. Their model includes two main components, namely language knowledge and strategic competence. Language knowledge consists of organizational knowledge having grammatical and textual knowledge as its sub-components and pragmatic knowledge composed of functional and sociolinguistic knowledge.

Grammatical knowledge is involved when producing and comprehending grammatically correct utterances or sentences, comprising knowledge of vocabulary, syntax, phonology, and graphology. Textual knowledge is involved when producing and understanding spoken or written texts composed of two or more utterances or sentences, including knowledge of cohesion and of rhetorical organization. Thus, organizational knowledge is understood to be involved in controlling the formal structure of language.

Functional knowledge, which was called illocutionary competence by Bachman (1990), allows language users to relate the meanings of the utterances or sentences and the texts to the intentions of language users. Sociolinguistic knowledge allows language users to produce or interpret language that is relevant and appropriate to a particular language use situation.

Strategic competence, an area of language users' cognitive execution ability, is employed in goal setting, assessment, and planning. Although strategic competence is a non-linguistic cognitive component of language ability, the role it plays in communication needs to be noted. Bachman (1990) defined strategic competence as "a general ability, which enables an individual to make the most effective use of available abilities in carrying out a given task" (p.106). The significance of strategic competence lies in the provision of a place where other components of communicative competence interact with each other. Using language involves the language user's topical knowledge and affective schemata in addition to language knowledge since communication is not just the outcome of mere knowledge of language, but rather a combination of multiple areas of competence. Referring to strategic competence, Bachman and Palmer (1996) suggested that it leads language users to relate their topical knowledge and language knowledge to a language use situation, thus helping to select available language abilities relevant to a language use setting, and formulating a plan for successful communication. Thus, strategic competence would be properly characterized as an ability rather than knowledge since the exercise of strategic competence draws on language users' cognition to achieve a communicative goal. This, therefore, serves as justification for the idea that language performance tests should be developed to encompass all that can possibly affect communication as well as linguistic knowledge. The conceptualization and the establishment of strategic competence in Bachman (1990) were considered to address the stated issues in Canale and Swain model by McNamara (1996) since Bachman's strategic competence demonstrated how non-linguistic ability can be involved in communication and explained how different components constituting language abilities can be integrated.

As Bachman and Palmer (1996) suggested that language skills, which were traditionally divided into four areas of listening, reading, speaking, and writing, should be more properly defined as "contextualized realization of the ability to use language in the performance of a specific language use task" (pp.75-76). Thus, writing, which relates to the present research, can be better defined as a combination of task characteristics and the areas of language ability involved, rather than being characterized as an abstract skill. Moreover, multiple components of language knowledge and strategic competence in Bachman and Palmer (1996) were not advanced as being only relevant to the specific modes of language use. Hence, the proposed language construct in the model of Bachman and Palmer can be applied to assessing writing as long as writing ability is evaluated in the context of language users' performing a language use task.

2.1.3 Empirical Research Exploring the Nature of Writing Ability

Since the seminal work of Bereiter and Scardamalia (1987), writing with an orientation toward process rather than production was examined throughout a plethora of writing research based on the Cognitive Development Theory (Bereiter & Scardamalia, 1987; Deane, Odendahl, Quinlan, Fowles, Welsh & Bivens-Tatum, 2008; Flower & Hayes, 1981; Grabe & Kaplan, 1996; Kellogg, 2001; Levy & Ransdell, 1995; Odell, 1993; Raimes, 1987; Ruetten, 1991; Tedick & Mathison, 1995; Zamel, 1987). Empirical research examining the nature of writing ability in an L1 and L2 context has provided the possible language constructs that need to be noted in writing assessments.

Introducing the notion of a process, Flower and Hayes (1981) explored what writers actually do in the process of composing a piece of writing. While referring to a writing task as a *rhetorical problem*, Flower and Hayes suggested that the ability to formulate mentally stored representations concerning an audience, a task, and a goal of writing distinguishes between skilled and poor writers. While poor writers spend much time generating a text, tied to a given topic without consideration for the readers, good writers were concerned with the prospective readers, which they found to be helpful for producing new ideas. From the perspective of the model of Bachman and Palmer (1996), it appears that unskilled writers largely seem to rely on organizational knowledge, with a focus on following writing conventions and attending to surface-level textual features. However, skilled writers were shown to be more involved in applying pragmatic knowledge in that they identified the role of a writer as influencing readers and were shown to be attentive to rhetorical situations.

Contrastive to the work of Flower and Hayes, whose attention was on the differences between skilled and unskilled writers, Bereiter and Scardamalia (1987) characterized the nature of easy and difficult tasks according to two models of composing: *knowledge telling* and *knowledge transforming*. While in a *knowledge telling* model, a writer approaches writing with a naturally acquired ability, which almost everyone has, and utilizes the full potential of language competence and skills effortlessly, not spending time in planning and revising a composition. Most L1 writers are known to have little difficulty in applying a *knowledge tellin*g strategy in writing since they have already developed linguistic competence enough to form grammatical sentences. In contrast, in a *knowledge transforming* model, continuous and conscious effort to monitor what to write (i.e., content generation) and how to write (i.e., rhetorical planning) is required. This inevitably causes a stream of conflicts between the existing knowledge (i.e., topical knowledge) and the rhetorical goal. As a result of the resolution of those conflicts, new ideas concerning content and rhetoric are generated, which is considered a sign of *knowledge transforming*. Thus, differences between good and poor writers can be attributed to which type of composing model is primarily employed. According to Bereiter and Scardamalia, novice writers retrieve content from their memory and translate it into writing without deliberation, which is evidence of a *knowledge telling model*. Expert writers, on the other hand, use *knowledge transforming* strategies, which seem to coincide with the *strategic competence* proposed by Bachman and Palmer (1996). As more advanced models*, knowledge transforming models* entail utilizing meta-cognitive elements in the process of selecting what is relevant to the writing task at hand from a stock of stored topical knowledge. Flower and Hayes (1981) and Bereiter and Scardamalia (1987), noted for approaching writing within a cognitive view, have similar findings: expert writers perceive a writing task as a problem and are thus engaged in a problem-solving process in which they must determine what and how to write in order to achieve a communicative goal.

Zamel (1982) and Raimes (1987), based on the commonalities in composing strategies employed by L1 and L2 writers, suggested that L2 writing techniques should not be distinct from those of L1 writers. However, Zamel (1982) and Raimes (1987) did not seem to take into account the lack of linguistic knowledge of L2 writers who are still in the language development phase. Conversely, L1 writers may have little difficulty forming grammatically correct sentences, which thus can draw a sharp contrast to the limited linguistic aspects of L2 writers.

L2 writers' relatively insufficient knowledge of linguistic code has compelled many L2 writing researchers and practitioners to formulate their own research agenda. Ruetten (1991) suggested that in L2 learning rhetorical control (e.g., coherence, development, and organization) and grammatical control (e.g., structure of sentences and use of articles, verbs, and prepositions) do not develop in tandem. This lack of balance between the components of L2 writing abilities was reflected in Ruetten (1991). Some ESL placement essays were assessed with the rhetorical aspect evaluated higher than the grammatical one, yet the other essays displayed the reverse pattern. Moreover, different ratings were given to the same essay due to different weights attached to rhetorical and grammatical features among raters.

Tedick and Mathison (1995) examined the rhetorical features of *framing* and *elements of task compliance* displayed by L2 writers. Given that *Framing* concerns how easily a reader can identify the topic of an essay and predict the development of the essay, well-framed writings are likely to be easier to understand. *Elements of task compliance* refer to whether writers include task elements such as a definition, reasons, or examples as specified by a task prompt. An important observation was derived from the research; topic development affected test scores more significantly than compliance with the task, thus corroborating the argument that framing should be considered a legitimate construct of L2 writing ability. Interestingly L2 writers who lack linguistic proficiency appeared to write better on topics demanding field-specific knowledge than on general topics, thereby revealing the role of topical knowledge as complementary to the limited linguistic competence of L2 writers.

2.2 Factors that Affect L2 Writing Assessments

Bachman (1990) discussed factors affecting the test scores of performance assessments under three categories: *test method facets, personal attributes*, and *random factors*. Despite the detailed specifications of these elements, Bachman did not include rating situational variables such as rater and rating criteria factors as elements affecting examinees' language performance, which were later proposed by Kenyon (1992) and extended by McNamara (1995). Though Bachman (1990) and McNamara (1995) proposed different sub-components that constitute language performance assessment, the common premise is that observed language performance is not solely responsible for test scores as interactions among different factors in relation to examinees, tasks, raters, and rating criteria affect the test score each examinee receives.

2.2.1 Rater Effects on L2 Writing Assessments

Linacre (1989) stated that between one-third to two-thirds of variability present in test scores can be attributed to variability among raters, which in effect amounts to the differences in examinee ability. Thus, rater variability is considered a primary reason for different test scores assigned to the same writing performance (Ruetten, 1991). The importance of raters in test scores was noted by Stiggnins (1987), who stated that it is not examinee performance but professional judgment by raters that defines performance assessments. In the same vein, Hill, O'Grady, and Price (1988) stated that “raters do not function as neutral recorders of some physical reality.” (p.346); differences among raters with concerns to rating criteria interpretation and application need to be considered to directly affect test scores.

Rater variability is largely understood to be due to psychometric forces such as severity or leniency, the interactions between raters and other components of the rating situation (i.e., rater bias), intra-rater consistency or inconsistency, and patterns in using rating scales such as *the halo effect*, *central tendency*, and *restriction of range* (Lumley, 2005). Rater variability regarded as contributing to *rater effects* represents possible effects that raters may have on test scores. The term *rater effects* are often used interchangeably with *rater errors* and *rater bias* (Myford & Wolf, 2003). For simplicity, this study adopts the term *rater effects* to indicate the comprehensive effects on systematic variance in test scores, attributable to raters, and not variance attributable to the actual performance of examinees. Depending on what raters are interacting with, the function of rater effects can vary; some of these interactions include raters to task types (Eckes, 2005; Wigglesworth, 1993, 1994), raters to task prompts (Weigle, 1999), raters to examinees (Brown, 1995; Caban, 2003; Du, Wright, & Brown, 1996; Kondo-Brown, 2002; Kuiken & Vedder, 2014), and raters to rating scales (Engelhard, 2002; Fahim & Bijani, 2011; Knoch, Read, & von Randow, 2007; Kondo-Brown, 2002; Kuiken & Vedder, 2014; McNamara, 1996; Schaefer, 2008; Skar & Jølle, 2017; Wigglesworth, 1993).

2.2.2 What are Rating Scales?

The present research concerns rater scoring behavior, which can be affected by the interaction between raters and rating criteria. To better understand the interaction between raters and rating criteria, which are rating scale components, it must be initially investigated as to how rating scales have been defined in previous research. Rating scales are developed to function as a reference point on which rater judgment is based. Valdes and Arriarza (1992) considered rating rubrics as representing an implicit theory about the constructs of writing ability as well as implicit assumptions about the development of L2 writing skills. Similarly, the importance of rating scales as manifesting constructs to be measured has been emphasized in other studies (North, 2003; Turner, 2000; Weigle, 2002).

Stiggins (1987) contended that the quality of performance assessment hinges on how the *performance criteria* are specified. *Performance criteria* refer to examinee behaviors or traits that can be observed and therefore include skills that are aimed for as an instructional goal. Clark and Watson (2019) stated that despite the perception of rating scales as more relevant to reliability than to validity, well-developed rating scales can contribute to enhancing the validity of language assessments. Considering rating scales as having a significant bearing on test validity, Huot (1993) also argued that based on the rating scales the evaluative process in performance tests can be controlled. In other words, raters' decision-making and cognitive exertions based on the rating scales are likely to draw relatively unbiased interaction between a rater and a rating scale. As Bejar (2012) discussed the significance of rater cognition as a basis for interpreting the outcome of a language test and consequently drawing validated inferences about examinees' language ability, impartial interactions between raters and rating criteria are of paramount importance.

In the same vein, McNamara (1996, p. 19) criticized the fact that rating scales have only been discussed in relation to reliability. He asserted that the object of measurement is reflected in rating scales. However, the presence of explicitly stated constructs of language ability in rating scales is not sufficient to ensure the validity of tests, due to rater inconsistency and idiosyncrasy in the use of rating criteria. It is inconceivable that all raters interpret and use rating criteria in the same way; some may place higher importance on content and development than linguistic accuracy and language usage whereas other raters can display reversed scoring foci. Concerning rater variability, McNamara (1996) noted that raters use rating criteria idiosyncratically but with systematic patterns within. He claimed that rater variability, though seemingly chaotic, can be predicted based on the overall rater severity and the overall criteria difficulty, which will eventually render the categorization of raters feasible.

2.2.3 Interactions between Raters and Rating Criteria

The significance of the relation of raters to rating criteria can be appreciated by considering how performance tests differ from selected-response questions; only performance tests need human raters who should base their evaluative decisions on rating criteria. In contrast, since the examinee's language ability in multiple-choice questions is evaluated based on the total number of correct answers, raters' subjectivity is not involved in the evaluation. Conversely, in performance-based tests, the interaction between raters and rating scales emerges as illustrated in Figure 2.1.

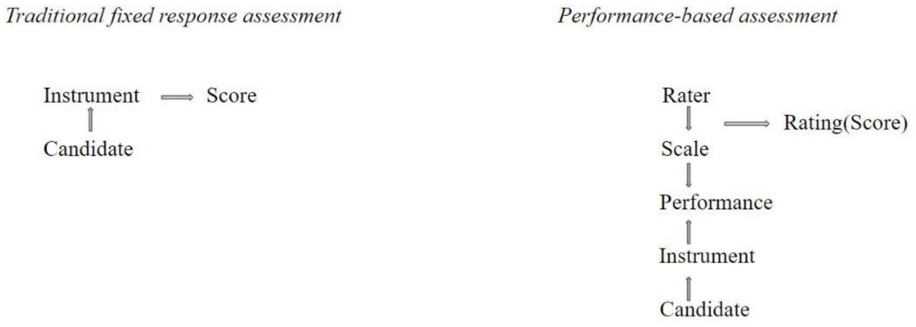


Figure 2.1. The Characteristics of Performance Assessment (McNamara, 1996, p.9)

Many studies have provided varying explanations as to the interaction between raters and rating criteria (Barkaoui, 2010; Cumming, 1990; Heidari, Ghanbari, & Abbasi, 2022; Huot, 1990; Li & He, 2015;

Plakans & Gebril, 2017; Vaughan, 1991; Winke & Lim, 2015). Cumming (1990) researched strategies that experts and novice raters employed in the evaluation of essays. All raters seemed to implicitly know that students' ESL proficiency and writing expertise are different, regardless of rating experiences, thus treating each as a separate, distinct factor to be considered in their evaluations. However, novice and expert raters exhibited qualitatively different approaches to assessing writing performance. Whereas the main concern of expert raters was to form a general impression of the text by distributing their attention evenly across all rating criteria, the efforts of the novice raters centered on editing grammatically erroneous textual features. Given that there was no special attention placed on any particular rating criteria among expert raters, it can be possibly claimed that expert raters assess writing performance from a more balanced perspective than novice raters. Cumming demonstrated that the decision of which rating criteria is considered important rests on rating experiences. However, Cumming only compared different scoring foci specific to each group of raters, not addressing the relation between raters' differentially perceived criterion importance and their distinctive scoring profiles.

Huot (1990, 1993) and Vaughan (1991) explored the decision-making processes of raters who used a holistic rating scale and obtained similar results. They both suggested that raters were more concerned with content and organization than language use. Vaughan further stated that some raters were shown to base their evaluative decision on criteria that were not stated in the rating rubric. Huot (1990) cited a list of previous studies showing that raters were likely to be most affected by content and organization among the components of rating scales (Breland & Jones, 1984; Freedman, 1979; Vaughan, 1991).

The dominance of attention to organization has also been illustrated in recent research (Barkaoui, 2010; Heidari et al., 2022; Li & He, 2015; Plakans & Gebril, 2017; Winke & Lim, 2015). Barkaoui (2010), Winke and Lim (2015), and Plakans and Gebril (2017) discovered that attention to organization can enhance rating consistency among raters using an analytic rating scale. Conversely, relatively less importance attached to the convention of writing (i.e., spelling and punctuation) among raters was observed in Milanovic, Saville, and Shen (1996) and Winke and Lim (2015). Milanovic et al. (1996) concluded that among different textual features, spelling and punctuation seemed to be considered the least important, and they also found that raters tended to put different weights on essay features.

Sakyi (2000), who also studied the ways in which raters place differential importance on criteria concerning various textual features, found that certain positive features were associated with high marks and other negative features with low marks. In Sakyi's study, other than the two essays which gained the highest and the lowest score, all other essays displayed a combination of good and bad marks. As there was no weight prescribed for rating criteria, different ratings were assigned to the same essay depending on which quality of the text each rater perceived as important. For example, whereas a certain essay was evaluated highly by those who saw language as important, the same essay was assigned a low mark by those who considered content crucial. Synthesizing rater comments on the essays, Sakyi suggested three main essay features, (1) content & organization, (2) grammatical & mechanical errors, and (3) sentence structure & vocabulary, as affecting rater evaluative decisions. Given that the raters in Sakyi used a holistic rating scale, the most significant finding of this research is that raters tend to place different weights on textual features according to the textual feature’s perceived importance to each rater. Sakyi’s work is distinguished from other studies; whereas many studies merely considered the focal textual feature without explicitly addressing the reason for the differences in attention to rating criteria, Sakyi ascribed differences in rater scoring to rater perception. This analysis shows the need for further studies interpreting the origin of rater effects. Sakyi's work, however, did not involve a statistical analysis demonstrating how raters differ in their scoring behavior, delineated in terms of severity or leniency toward particular rating criteria.

Cumming, Kantor, and Powers (2002) described what elements skilled raters focused on by using concurrent verbal reports from raters. They investigated decision-making processes in order to validate a new scoring scheme for the writing component of the Test of English as a Foreign Language (TOEFL). What was notable about their research was that the main focus was not on textual features that raters attended to in evaluating ESL and EFL written compositions, but rather on rater decision-making in the evaluation process. They found that two groups, each consisting of 10 ESL / EFL raters and 7 EMT (English-mother-tongue) raters respectively, seemed to be convergent in the use of decision-making behaviors. Their behaviors include a self-monitoring focus (e.g., rereading composition and considering personal biases), a rhetorical and ideational focus (e.g., accessing reasoning, topic development, task completion, and coherence), and a language focus (e.g., accessing error frequency, lexis, syntax, and spellings or punctuation). However, the scoring foci among the two groups were different. Whereas ESL and EFL raters tended to focus on language rather than rhetoric and ideas, EMT raters seemed to balance their attention between language, ideas, and argumentation. The different scoring foci between these two groups were analyzed to be derived from different conceptualizations of writing; ESL/EFL raters approached writing as a language learning process whereas EMT raters regarded it as a means to expressing ideas in a literary manner. Contrastive to Cumming (1990) in which the importance attached to superficial textual features, such as writing conventions and language use, was characterized as the trait of inexperienced raters, Cumming, Kantor, and Powers found that the perception of English, which should vary among ESL/EFL raters and English native raters, affects scoring foci. Added to the analysis of different conceptualizations of writing among raters was the discussion of how raters' varying scoring foci were influenced by examinee writing proficiency; while raters were shown to attend more to language in low proficiency essays, greater concern with ideas and rhetoric were noted in high proficiency compositions.

Lumley (2002), who also observed the rater-to-rating criteria interaction, placed particular interest in the actual rating process in which raters interpret and apply rating criteria. Challenging the prevailing assumption that well-constructed rating scales and proper rater training will suffice for raters to reach sound evaluative judgments on examinee language ability, Lumley demonstrated that these two conditions alone are insufficient for ensuring validated assessments. Despite rater training aiming to inform raters of how to read the rating rubrics before the live rating sessions, the raters were found to struggle with matching their evaluative decisions with the suitable performance descriptors on the rating scales. In Lumley's study, raters were observed to rely heavily on their overall impression of the text initially, and then to find the proper descriptors that they considered relevant to the construct they were dealing with. However, raters encountered difficulty when there were no corresponding descriptors that concerned the construct they deemed as important. For example, as the length of the text was not a construct being measured, any statements relating to essay quantity were not included in the rating rubric. While experiencing difficulty locating a proper descriptor to match the observed performance at hand, some raters were shown to be reluctant to assign a good mark to short essays and criticized the rating scale for being of no help. Referring to raters' use of the unstated content on the rating scale, Lumley (2002) and Vaughan (1991) found that rating scales were not employed as intended, attesting to the possibility that raters bring their own unique rating strategies to justify their decisions. Despite insight into rater variability, the study was not without limitations; the cognitive aspects of raters were only addressed as a coping strategy to resolve the mismatch between the raters' impressionistic initial judgment of the texts and the content of the rating rubric. Lumley's research did not provide accounts of how rater cognition can be involved when raters created an initial impression of writing performance, which can essentially be associated with perceived criterion importance. Lumley only focused on the observation that raters did not use the rating criteria as they were supposed to be referred to. For example, questions such as "Why do some raters resort to the unstated criteria such as the length of the text?" and "When two competing criteria, namely relevance and clarity of meaning, were considered, why do some raters choose relevance over clarity whereas others do the reverse?" can have been analyzed in reference to criterion importance. To illuminate, as for the first question, a possible answer could be that the quantity of writing might be an overriding criterion to those raters who did not want to assign a good score to short essays. Similarly, the criterion importance can answer the second question; when raters should determine which criteria to be employed, they are likely to prioritize what they consider as important. Taken together, beyond the observation that raters based their evaluation on the unstated content of the rating rubric, it would be worthwhile to reveal the reason why these unstated rating criteria matter to the raters. This reason will inevitably be associated with raters' perceived criterion importance.

2.3 Rater Cognition

Rater cognition can be a broad term since it can include anything that might transpire in the minds of raters who are conducting evaluations. Bejar (2012) identifies rater cognition as "the attributes of the raters that assign scores to student performances, and their mental processes in doing so" (p.2). This definition denotes both the personal aspects of raters (e.g., a rater’s age, gender, education, and language background) and the architecture of human information processing along with the metacognitive strategies employed while rating (Han, 2015). Various topics relating to rater cognition have been addressed to date, including which rating strategies are deployed (Cumming et al., 2002; Lumley, 2002), how native English-speaking raters and nonnative English teachers differ in the perception of rating criteria (Hijikata-Someya, Ono, & Yamanishi, 2015), how raters differ in the interpretation of rating criteria (Ono, Yamanishi, & Hijikata, 2019; Wang, Engelhard, Raczynski, Song, & Wolfe, 2017), how different scoring behavior emerges depending on the type of rating scales (Heidari et al., 2022; Jeong, 2019), and how differently perceived criterion importance affects rater scoring behavior (Eckes, 2012). Among these, the present research aims to examine the relation of perceived criterion importance to rater scoring behavior empirically, which has been relatively unexplored so far.

Although rater cognition can be broadly defined, as seen by the wide scope of research topics covered, rater cognition in the present research is limited to exploring how raters’ perception of rating criteria is related to criterion-related bias. Eckes (2008, 2012) attempted to bridge the gap between rater cognition and rater scoring behavior. Based on self-reported results that inquired about the importance attached to nine rating criteria (i.e., fluency, train of thought, structure, completeness, description, argumentation, syntax, vocabulary, and correctness), 64 experienced raters, who were working as evaluators for the writing section of the Test of German as a Foreign Language (TestDaf), were classified into six cognitive rater types (CRTs) (Eckes, 2008). Eckes (2012) had the raters evaluate the essays to verify raters' effects of criterion importance ratings on scoring behavior; Eckes found the existence of rater-criterion bias, severity being displayed towards criteria that raters reported as extremely important in rating writing performances. Conversely, for those criteria regarded as less important, raters showed leniency. Based on these criterion-related biases, operational rater types (ORTs) were identified, which were to be compared to CRTs. By combining the respective findings of two studies (Eckes, 2008, 2012), Eckes concluded that rating criteria perceived as highly important are more likely to be evaluated more severely and rating criteria deemed less important tend to be rated more leniently. Therefore, Eckes succeeded in illuminating how rater perception of rating criteria influences scoring behavior.

Despite contributing to connecting rater cognition with scoring behavior, two questions about the reliability of the research findings can be posed. First, in Eckes (2012) only the data of 18 out of the 64 raters, who originally participated in the rater cognition research (2008), were analyzed. As six CRTs in Eckes (2008) were formed, it would have been more appropriate to include all six CRTs to examine the comprehensive effects of differentially perceived criterion importance on severity or leniency of raters. Hence, having a limited number of only 18 raters belonging to four out of the six cognition types can raise the question of replicability: Would the same result emerge if the raters from the two excluded rater cognition types are included in the comparison of CRTs and ORTs?

Secondly, similar to the limitations of the study noted by Eckes, the four-month lapse of time between the data collection, the survey of criterion importance ratings, and the live scoring sessions could possibly weaken the evidence for the link between rater cognition and rating behavior. Considering that rater perception might change within the period of four months, the link between the perception of criteria and rating behavior might be undermined, which potentially harms the credibility of the research. While it is undeniable that Eckes (2012) dealt with an unexplored but critical topic that associates rater cognition with rating behavior in the field of language assessment, selectively choosing an underrepresented sample from the previous research on which the following research (Eckes, 2012) was extended can limit generalizability as well as the reliability of the study. Moreover, the lengthy period between the data collection phase and the actual rating sessions in which criterion-related bias measures (i.e., the interaction between raters and rating criteria) were obtained can also threaten the validity of the research findings. Taken together, further research with the same raters who participate in both the criterion importance survey and the live rating sessions without any passage of time in between needs to be conducted.

2.3.1 Research on Scoring Behavior Using Many-facet Rasch Measurement (MFRM) Analysis

Studies researching rater cognition are premised on the idea that there are interactions between raters and rating criteria. Despite differences in the details of the observations and implications of each study, the overarching finding is that raters are inconsistent and idiosyncratic in the use of rating criteria. Several attempts have been made to understand rater scoring patterns using Many-facet Rasch measurement (Linacre, 1989) focusing on the interaction between raters and rating criteria (Eckes, 2012; Kondo-Brown, 2002; McNamara, 1996; Park, 2012; Park & Shim, 2014; Schaefer, 2008; Shin, 2010; Wigglesworth, 1993).

Many-facet Rasch measurement (Linacre, 1989), or MFRM is a latent trait model of probabilities in which testing components called *facets* are calibrated independently of one another and placed within a common frame of reference called *logits* (log odds units). Included in *facets* are, for example, task types, raters, rating criteria, and rating occasions (Barkaoui, 2014). Expanding on the basic Rasch model (Rasch, 1960) in which two facets, *item difficulty* and *examinee ability* are calibrated, Many-facet Rasch measurement adds the facet of *rater severity* or other facets of interest, thereby allowing the construction of possible combinations of interactions of all facets constituting the assessment setting. Thus, a Many-facet Rasch measurement allows test developers and researchers to model various facets in a testing setting, estimating the impact of each facet on the evaluation process. In particular, bias analysis is the main function of MFRM. Wigglesworth (1993) referred to bias analysis as identifying any systematic sub-patterns of behavior occurring from an interaction of a particular rater with a particular aspect of the rating situation (p.309). McNamara (1996) compared the interactions between raters and rating scales to those between test-takers and test instruments; just as the interactions between test-takers and test instruments are measured, so are the interactions between raters and rating scales (p.121). Hence, information on the raters and rating scale is required to locate the source of rater variability, which can possibly be discovered through MFRM (McNamara, 1996).

Several empirical studies identified biased rater-rating criteria interactions by using MFRM. For instance, Kondo-Brown (2002) explored the scoring patterns of three trained native Japanese-speaking raters who evaluated Japanese English learners' writing performances employing

MFRM and found rater scoring patterns unique to each rater. However, due to the small number of raters in the study, there was a limitation with generalizing the findings. Addressing the issue of the small sample of Kondo-Brown (2002), Schaefer (2008) conducted bias analysis using MFRM with 40 raters who evaluated 40 essays written in English by Japanese university students. Among six rating criteria, namely content, organization, style and quality of expression, language use, mechanics, and fluency, one subgroup of raters displayed severity toward content and organization, but leniency for language use and mechanics. However, another subgroup of raters showed reversed scoring patterns with severity toward language use and mechanics and leniency for content and organization. Noting that one subset of rating criteria is evaluated harshly whereas another subset is rated leniently, Schaefer suggested that raters take compensatory strategies in applying rater criteria, which was also confirmed in Ekces (2012). Despite the relatively large number of raters, however, Schaefer did not provide the possible reason behind the displayed opposite scoring patterns, for which Eckes (2012) noted perceived criterion importance as the mediating variable affecting raters' severity bias.

As for the investigations conducted in Korean educational settings, Park (2012), Park and Shim (2014), and Shin (2010) also employed MFRM, focusing on the interactions between raters and rating criteria. Whereas these three studies situated in a Korean context can be valuable for the attempt to analyze rater scoring patterns statistically, they all had a very limited number of raters and were centered on the descriptions of observed rater behavior without connecting the research findings to raters’ criterion perception.

2.4 Summary

This chapter dealt with how the concept of language performance assessments has been conceptualized, discussed theories of language ability, and reviewed empirical research on L2 writing ability. In particular, rater effects, in relation to the interactions between raters and rating criteria, were introduced as a significant factor affecting L2 writing assessments. The discussion covered empirical research addressing different scoring foci among raters and investigating criterion-related bias based on MFRM. The discussion of rater effects identified a paucity of studies that connect rater cognition and operational behavior, thus suggesting a need to explore how raters’ perceived criterion importance affects scoring behavior.

CHAPTER 3

METHODOLOGY

This chapter centers on how research data was collected and how collected data was statistically analyzed. Section 3.1 concerns the recruitment of participants and Section 3.2 presents the instruments used in the research. Section 3.3 explains the research procedures for collecting and analyzing data. Lastly, Section 3.4 summarizes the chapter.

3.1 Participants

A total of 30 in-service Korean English teachers working at middle or high schools participated in this research. Participants were recruited through a poster in which the purpose and procedure of the research were explained. All participants voluntarily notified the researcher of their intention to take part in the research by clicking on the ‘Agree’ button for the rating criterion importance survey and filling in the consent form for rating 30 essays.

3.2 Instruments

3.2.1 Questionnaire

The survey questionnaire was developed to investigate the different weights each participant generally assigns to five rating criteria (i.e., Content, Organization, Vocabulary, Language use, and Mechanics) when evaluating writing performances by using a four-point rating scale ranging from *extremely important*, *very important*, *important,* to *less important*. The five rating criteria were adopted from ESL Composition Profile developed by Jacobs, Zingraf, Wormuth, Hartfiel, and Hughey (1981). Content refers to the prepositional content and coherence in the flow of a text. Organization indicates connections and unity across paragraphs and cohesion of a text. Vocabulary specifies the word range and choice relevant to the context, lexical diversity, and word form mastery. Language use pertains to grammatical accuracy and syntactical variety and Mechanics denotes capitalization and generally superficial aspects of a text.

The survey was conducted online in a Google Survey Form in which the participants were instructed to indicate the importance of each rating criterion that they usually bring toward a general writing assessment context without assuming a specific writing situation. The scheme of the four-point scale of criterion importance was adopted from Eckes (2008). Included in the questionnaire was a brief definition of each rating criterion to ensure that all participants could have an equal understanding of the operational meaning of the rating criteria being employed. The questionnaire is provided in Appendix 1. Participants were requested to complete the survey before the evaluation of the 30 essays. The estimated time for completing the survey was one to three minutes. The survey results were later subject to a two-way facet (i.e., rater × rating criteria) analysis.

3.2.2 Essays to be Rated

The thirty essays to be assessed were selected from YELC (Yonsei English Learner's Corpus), which was accumulated in 2011 from an L2 English writing placement test for Younsei University's first-year students. The proficiency of 30 chosen essays ranged from A2 to C1 in CEFR (Common European Framework of Reference); five essays were from A2, seven from B1, five from B1+, eight from B2, three from B2+, and two from C1. All participants were provided with the same set of essays presented in the same order. All thirty writing compositions were argumentative essays, the topic of which was 'Should people use their real name on the Internet?' All 30 essays to be rated were presented both in a text file and an image so that raters could choose either form based on preference. The essay ratings had to be submitted via the shared online Google Form. The approximate time for evaluating one essay was estimated to be five to ten minutes.

3.2.3 Rating Rubric

Two analytic rating scales, which were specifically developed for the L2 writing evaluative context, were modified for the rating rubric being employed in the present research; the one commonly known as ESL Composition Profile developed by Jacobs et al. (1981), and the other by Connor-Linton and Polio (2014), which revised Jacobs et al's rating rubric.

Sasaki and Hirose (1999) stated that Jacob et al's rating rubric was noted for defining the construct of L2 writing ability clearly and showing the validity of the rating scale. Jacobs et al's ESL Composition Profile has been widely implemented as an analytic rating rubric for assessing writing proficiency levels in ESL/EFL programs (Bacha, 2001; Cho, 1999; Ghanbari, Barati, & Moinzadeh, 2012; Ishikawa, 2018; Kim, 2020; Setyowati, Sukmawan, & El-Sulukiyyah, 2020). In Jacobs et al's rating rubric differential weights were attached to each rating criterion (i.e., 30 % for content, 20% for organization, 20% for vocabulary, 25% for language use, and 5% for mechanics of the total mark, respectively). However, some researchers discussed that uneven weightings could result in a distorted perception of the rating criteria on the part of raters, thus arguing for employing an equal point scale rather than differential weightings (Connor-Linton & Polio, 2014; Kim, 2020). Distributing weights evenly across rating criteria was to prevent raters from assuming that a rating criterion with heavy weight is the most important.

Both Jacobs et al's (1981) and Connor-Linton and Polio's rating rubric (2014) tapped into the same five components of textual features: content, organization, vocabulary, language use, and mechanics. However, there are two notable differences between these two analytic rating scales. First, Connor-Linton and Polio's rating rubric adopted a 20-point scale across the rating criteria whereas differential weightings across criteria were assigned by Jacobs et al's. Second, the descriptors in Connor-Linton and Polio were different from those of Jacobs et al. Connor-Linton and Polio's descriptors were developed to reflect raters' evaluations of what actually improved in the students' writing compositions, which were written at different points in time over a semester, and thus were displaying changes in writing quality. Raters in Connor-Linton and Polio's study (2014) were asked to rank the given essays written by the same writer based on quality and to discuss what they perceived determined the quality changes. Connor-Linton and Polio (2014) said that their descriptors reflected a practical measure of writing quality and showed enhanced reliability and a higher correlation with a holistic rating than Jacobs et al.'s rating rubric.

The present study employed the descriptors of Connor-Linton and Polio’s rating scale (2014) with some descriptors concerning Language use adopted from the Jacobs et al. rubric. However, a 5-point scale instead of a 20-point across the rating criteria was adopted due to the estimated difficulty of discerning the multiple scoring bands. Furthermore, there were revisions to Mechanics rating criterion. Whereas appropriate paragraphing was deemed as an important factor in Connor-Linton and Polio (2014), paragraphing was deleted in the rubric of the present study since all essays in YELC were composed in one paragraph. Other elements such as proper capitalization, punctuation, and spelling under Mechanics remained relevant aspects of L2 writing ability. The rating rubric was provided both in Korean and English to prevent any possible misinterpretation. The rating scale is attached as Appendix 2.

3.3 Procedures

This section presents the procedures of data collection and analysis. Section 3.3.1 provides descriptions of the data collection process. In Section 3.3.2, methods employed for data analysis are discussed.

3.3.1 Data Collection

The data collection was initiated after gaining approval from the Ethics Committee of the Institutional Review Board (IRB) of Seoul National University. Thirty in-service Korean English teachers teaching in a middle or high school were recruited as participants through a recruitment poster advertised in an online community of in-service Korean English teachers.

Before responding to the criterion importance survey, participants were requested to attend an individual Zoom meeting in which each participant was informed of the research procedures, which consisted of the criterion importance survey and the evaluation of 30 essays. The respective Zoom meeting approximately took 10 to 20 minutes. This Zoom meeting aimed to notify participants of the rating rubric to be employed in the evaluation so that they could achieve a similar understanding of the operational definition of the rating criteria as well as the rating scale structure. The individual Zoom session, in effect, was assumed to serve as a kind of norming session to help participants to apply the rating rubric consistently across the rating criteria and the scoring structure. Some questions as to the rating rubric posed by participants were addressed during the Zoom meeting. Additionally, participants were notified of the fact that they should provide five ratings per essay according to each rating criterion, all of which ranged from 1 to 5. Participants were also provided with a notice stating if they still wish to participate, they should submit the consent form before rating the essays; following this, they were asked to complete the evaluations of all 30 essays within 30 days. Finally, participants were informed of their right to withdraw consent of participation at any point in the research without any penalty. The researcher had 30 individual Zoom meetings with the respective participants.

Directly after the individual Zoom meetings, all materials needed for research participation were provided to the participants online: a link to the criterion importance survey in Google Form, 30 essays presented in the format of both a text file and an image, the rating rubric in Korean and English, and a link for participants to record the assigned rating scores. The reason for presenting essays in two formats both as a text and an image file, was to allow participants to choose either form depending on personal preference in order to facilitate the evaluation. The text file was offered for those who may prefer the printed version of the essays, and the electronic one was given to those who might want to evaluate the essays on a computer.

Accompanied with the online survey was a description of the research and an online consent form for the survey. Before starting the survey, participants were asked to read the research descriptions carefully and click on the 'Agree' button provided, showing their intent to participate in the research. The time taken for completing the survey was estimated to be two to three minutes. The responses to the survey were collected for two months. The time constraint for completing the evaluation of 30 essays was 30 days at maximum. The outcome of the evaluation, the test scores of the essays, were gathered for two months.

3.3.2 Data Analysis

Two statistical methodologies, a Many-facet Rasch measurement analysis (MFRM) and hierarchical clustering, were applied for addressing the research questions of the present study. SPSS 29 for hierarchical clustering and FACETS (version 3.85, Linacre, 2011) for a MFRM analysis were used in the study. Below is a discussion of what each analysis is and how these statistical methods were employed to address the respective research questions.

Clustering raters draws on McNamara (1996) and Eckes (2008) who noted that raters are not homogeneous in terms of perceived criterion importance. They suggested that raters can fall into distinctive types, each characterized by distinct scoring foci. Eckes (2008, 2009) found that some raters displayed a strong focus on vocabulary and syntax, whereas others put significantly more weight on structure or fluency. Raters having different scoring foci, thus, do not form a single homogeneous group and can be differentiated into rater types, each defined by distinct scoring foci (Eckes, 2015). Eckes (2008) empirically demonstrated that raters could be grouped into different rater cognitive types, which were later linked to operational rater types identified in Eckes (2012). The present research motivated by two studies by Eckes (2008, 2012) conducted two hierarchical clustering analyses; one was to form cognitive rater types (CRTs) based on perceived criterion importance and the other was to build operational rater types (ORTs) according to criterion-related bias respectively. The terms CRTs and ORTs were adopted as they were used in Eckes (2008, 2012).

Hierarchical clustering, also known as hierarchical cluster analysis, is a widely implemented method for the classification of objects. This method generates clusters in which objects within a cluster display similarities to each other and differences from objects in other clusters. Johnson (1967) stated that hierarchical clustering enables clustering or the arrangements of the subjects under the study into homogeneous groups based on empirical measures of similarity. The similarity between clusters, which is measured by the distance between clusters, can be computed by several linkage criteria such as Single linkage, Complete linkage, Average linkage, Centroid linkage, and Ward’s method.

In Single linkage, two clusters are merged in a way that the two closest members of each cluster have the smallest distance. In Complete linkage, which is the opposite of Single linkage method, two clusters merge when the two farthest members of these two clusters have the largest distance. In Average linkage, the distance between two clusters is defined as the average of distances between all pairs of the elements, each pair from two clusters. In Centroid linkage, the distance between two clusters is set as the distance between the two mean vectors of two clusters. In Ward’s method, the distance between two clusters is computed as the increase in the combined error sum of squares, after merging two clusters into a single cluster. Thus, Ward’s method generates clusters that yield minimized within-group dispersion at each binary fusion (Murtagh & Legendre, 2014). What linkage criterion to choose in conducting hierarchical clustering varies according to the theoretical considerations of investigations.

Hierarchical clustering initially operates by treating each data point as a separate cluster, then identifying and merging two clusters that are closest together based on the chosen linkage criterion. These steps are iterated until all the clusters are merged and one cluster is generated in the end. The visual representation of the result of hierarchical clustering is provided through a tree diagram, which is called a dendrogram. A dendrogram displays the hierarchical relationship between all the data points.

Since all estimations resulting from FACETS are indicated as logits, which is interval-scale, Ward's method was adopted due to its advantage for analyzing interval-scale data. Each cluster obtained from hierarchical clustering comprised raters sharing a distinctive pattern of perceived criterion importance (i.e., CRTs) and raters sharing a scoring profile toward rating criteria (i.e., ORTs), which thus distinguished the given cluster from others.

A MFRM analysis is widely implemented for its ability to factor relevant variables (e.g., examinee proficiency, rater severity, rating criteria difficulty, and scale categories difficulty) that are assumed to affect the test scores into measuring the impact of the respective variables, or *facets* on the scores that raters assign to examinees. Obtained under MFRM analyses is a set of indices such as a fixed chi-square index, separation index, separation reliability index, infit index, and outfit index, all of which indicate variability across the elements of each facet.

There were two uses of MFRM in the study; at first, a two-way facet analysis (i.e., rater × rating criteria) as a preliminary analysis was employed to ascertain variability in the degree of raters' perceived criterion importance as well as the function of the rating scale of the survey as intended. The necessity of conducting a MFRM analysis at this stage is closely associated with addressing the first research question, 'How can raters be classified into a group based on perceived criterion importance?’ To illustrate, clustering raters into cognitive rater types (CRTs) can only be possible when there is rater variability of perceived criterion importance. For instance, if participants similarly consider all rating criteria as very important or less important, there is no way of proving the effect of differentially perceived criterion importance on scoring behavior. Indices from MFRM such as chi-square, separation index, and separation reliability can confirm that raters differentially perceived criterion importance and that all rating criteria were viewed differently. Furthermore, it also needs to be investigated whether the rating scale functioned as intended; that is, higher ratings denote higher importance attached to rating criteria. The functioning of the rating scale can be examined by the monotonic increase in the values of the category thresholds and the average rater importance measures as the importance of the rating category increases (i.e., from less important to important, from important to very important, from very important to extremely important). Only after verifying the rater variability of perceived criterion importance and the functioning of the rating scale through a MFRM analysis, can raters be then categorized into cognitive raters types (CRTs) according to perceived criterion importance, thus answering the first research question.

There was another employment of MFRM when conducting a bias analysis between participants and the rating criteria, whose result provided the input to the categorization of raters into operational rater types (ORTs). Hence, the second use of MFRM pertained to dealing with the second research question, ‘How can raters be classified into a group based on severity or leniency toward particular rating criteria displayed in live scoring sessions?' Two Rasch models, the rating scale model and the partial credit model, were used to model three facets: raters, examinees (essays), and rating criteria. Of particular interest in the study were the interactions between raters and rating criteria, thus for all possible combinations of raters and rating criteria, a MFRM bias analysis provided evidence that a rating given by a particular rater on a particular criterion was higher or lower based on the overall rater severity measure and the overall criteria difficulty measure. As a result of hierarchical clustering based on MFRM bias measures, raters with a similar scoring profile were grouped in the same ORT, which concerns the second research question.

To address the third research question, 'To what extent is perceived criterion importance related to scoring behavior?', two approaches were employed. The first was to combine the mean value of the criterion-related bias measures of raters in each ORT with the mean value of the criterion importance ratings of the same raters, thereby enabling the investigation to be based on the data of groups of raters. The first approach applying the mean values of the bias measures and the criterion importance ratings, however, may not be sufficient for the investigation into the relationship between criterion perception and rating patterns since the mean is sensitive to outlier values. To deal with the potential problem of the qualification of mean values as a representative, the second approach adopted was to directly compare the bias measure of an individual rater with the criterion importance rating of the same rater, thus not involving the mean values.

As for the group-based approach, since all ORTs were composed of raters coming from different CRTs, it needed to determine what can be a representative for the criterion importance rating. The mean of criterion importance ratings was chosen over the common importance marker. In Eckes (2012) only the common importance marker across the constituting CRTs in each ORT was adopted. In Eckes (2012), moderate importance, which was indicated as 3 or 2 in the criterion importance rating scale, was not considered in the analysis of the relation of CRTs and ORTs if moderate importance was identified as the common importance marker. That is, only the common importance marker, which was either extreme importance (i.e., 4 on the rating scale) or less importance (i.e., 1 on the rating scale), was the subject of the analysis. In the present study, however, rather than the common importance marker, the mean of the criterion importance ratings of the raters in each ORT was employed in matching the perception and bias. The reason was that the common importance marker alone did not seem to represent the overall perception tendency of ORTs considering that there was not a single ORT that was composed of the raters belonging to the same CRT.

The advantage of opting for the average criterion importance rating can be evident by setting a hypothetical situation. For example, it can be possible that raters of an ORT came from two CRTs whose respective importance ratings for a particular criterion were moderate importance (e.g., 3 on the rating scale) and extreme importance (i.e., 4 on the rating scale). The criterion bias for the ORT was leniency. In this case, focusing on the fact that there was no common importance marker could not provide any basis for analyzing the relation between criteria perception and scoring behavior. However, if the mean of the importance ratings was chosen over the common importance marker, the mean value of the importance rating (i.e., 3.5 in this example case) could allow the investigation of the relation of two variables, perception and bias; the finding, in this case, would be that even criteria perceived as somewhat important can be rated leniently.

As for the mean criterion bias measure of each ORT, five bias values per rating criterion were calculated. Positive bias values denote more leniency of raters on the criterion involved than other criteria based on the overall rater severity measure and the overall criterion difficulty measure. Conversely, negative bias measures refer to more severity on the part of the rater toward the criterion involved than any other criteria based on the overall rater severity measure and the overall criterion difficulty measure. Applying a so-called half-logit rule (Draba, 1977), mean bias measures more than (absolute) 0.50 logits were chosen as signaling severity or leniency bias, which were later compared to the mean of the criterion importance ratings.

Concerning the second approach, which was based on the bias measure and criterion importance rating of an individual rater, the same half-logit rule (Draba, 1977) was applied to detect the criterion bias.

The investigation into the relationship between the bias measures displayed as severity or leniency and the criterion importance ratings thus addressed the third research question.

3.4 Summary

This chapter presented research methods for the current study with a discussion of the context of the study, participants, and instruments. Furthermore, it described how two quantitative analyses, hierarchical clustering and MFRM, were conducted in addressing the research questions.

CHAPTER 4

RESULTS AND DISCUSSION

This chapter presents the findings of the study and discussion related to the research questions. Section 4.1 provides descriptive statistics for criterion importance ratings and the ratings of 30 essays categorized into six levels of CEFR. Section 4.2 reports on the outcomes of a two-way facet Rasch analysis, which acted as a preliminary step to the formation of CRTs. Section 4.3 provides the results of CRTs, which described how raters were categorized into different CRTs. Section 4.4 presents FACETS results focusing on rater measurements along with the figures of inter-rater agreement and intra-rater reliability. Section 4.5 delineates the results of ORTs and Section 4.6 analyzes the relationship between perceived criterion importance and scoring behavior based on the data both from groups of raters and individual raters.

4.1 Descriptive Statistics

Table 4.1. shows the means of the criterion importance ratings across participants; the mean value for Content was highest (3.62), followed by Organization (3.12), Vocabulary (2.27), Language use (2.12), and Mechanics (1.54), thus interestingly coinciding with the order of the presentation of the rating criteria.

Table 4.1 Means of the Criterion Importance Ratings per Rating Criterion (N=30)

Rating criteria	The Mean of criterion importance ratings
Content	3.62
Organization	3.12
Vocabulary	2.27
Language use	2.12
Mechanics	1.54

Table 4.2 reports descriptive statistics for the means and standard deviations of the ratings of six levels of essays (i.e., A2, B1, B1+, B2, B2+, and C1), categorized according to CEFR (Common European Framework of Reference) across the rating criteria. 30 Essays which were assigned ratings comprised five from A2, seven from B1, five from B1+, eight from B2, three from B2+, and two from C1.

Table 4.2 Descriptive Statistics for Essay Ratings across the Rating Criteria (N=30)

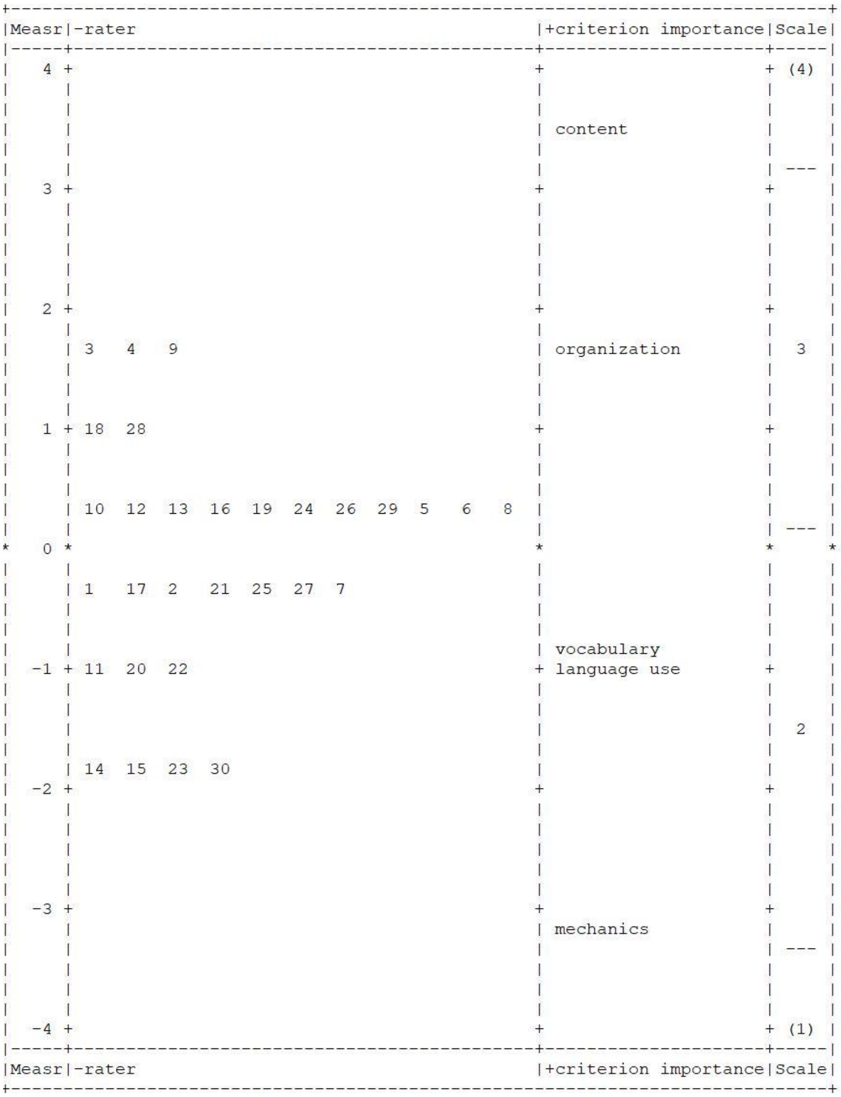
Criteria		A2	B1	B1+	B2	B2+	C1
Content	Mean	2.97	3.48	3.30	3.99	4.47	4.15
	SD	1.05	0.96	1.01	0.85	0.74	0.80
Organization	Mean	2.89	3.58	3.20	3.95	4.54	4.13
	SD	1.02	1.03	1.00	0.90	0.72	0.85
Vocabulary	Mean	3.01	3.26	3.23	3.73	4.38	4.22
	SD	0.99	0.94	0.89	0.87	0.68	0.74
Language use	Mean	2.77	3.24	2.99	3.74	4.42	4.32
	SD	1.13	1.00	0.98	0.84	0.65	0.70
Mechanics	Mean	3.23	3.69	3.73	4.29	4.62	4.62
	SD	1.29	1.19	1.07	0.81	0.61	0.56

As seen in Table 4.2, it was notable that the increase in writing proficiency of the essays did not coincide with higher ratings; as the level of CEFR increased from B1 to B1+ and from B2+ to C1, the mean values of the corresponding ratings for the four rating criteria, Content, Organization, Vocabulary, and Language use decreased. As for Mechanics, however, as the CEFR level rose from B1 to B1+ and from B2+ to C1, the mean of the assigned ratings increased slightly and remained the same, respectively. This negative relation between the proficiency level and the ratings observed in the sections of B1 to B1+ and B2+ to C1 towards all criteria except Mechanics could be attributed to the difference in the type of the rating scale used in YELC and the present study; whereas the CEFR levels were assigned to the essays in YELC based on the use of a holistic rating scale, the present study adopted an analytic scale, composed of five rating criteria. Concerning differences between a holistic and an analytic rating scale in the evaluation of writing performances, Carr (2000) found that test scores derived from these two rating scales were not comparable. Thus, it may well be that there was no coincidence between the CEFR levels in YELC based on a holistic rating scale and the essay ratings in the present study using an analytic rating scale.

4.2 Two-way facet Rasch Analysis

Participants were instructed to indicate the criterion importance in the survey, whose results became the input to the hierarchical clustering of raters into different rater cognitive types. However, before categorizing participants, an investigation was needed as to the rater variability of perceived criterion importance; if all raters perceive the rating criteria similarly very important or less important, an attempt to cluster participants based on perceived criterion importance cannot make much sense. Thus, a two-way facet Rasch analysis was to ascertain that participants differentially perceived criterion importance. Additionally, a probe into how the rating scale of the survey functioned was needed; that is, a higher importance rating should be used to indicate high importance attached to a criterion and less importance ratings should be used to denote less importance placed on a criterion. This was also examined through a two-way facet Rasch analysis.

The overall data-model fit was satisfactory since unexpected observations whose (absolute) standardized residual was equal to or greater than 2 or 3 were not observed at all, indicating that importance ratings assigned by participants were not divergent from the estimates calculated by a computer facet program. Figure 4.1 exhibits the Wright map, also known as the variable map, showing the calibrations of raters and criterion importance. Table 4.3 presents summary statistics from the facet analysis.



```
+------------------------------------------------------------------------------+
|Measr|-rater                                      |+criterion importance|Scale|
|-----+--------------------------------------------+---------------------+-----|
|   4 +                                            +                     + (4) |
|     |                                            |                     |     |
|     |                                            |                     |     |
|     |                                            | content             |     |
|     |                                            |                     |     |
|     |                                            |                     | --- |
|   3 +                                            +                     +     |
|     |                                            |                     |     |
|     |                                            |                     |     |
|     |                                            |                     |     |
|     |                                            |                     |     |
|     |                                            |                     |     |
|   2 +                                            +                     +     |
|     |                                            |                     |     |
|     | 3   4   9                                  | organization        |  3  |
|     |                                            |                     |     |
|     |                                            |                     |     |
|     |                                            |                     |     |
|   1 + 18  28                                     +                     +     |
|     |                                            |                     |     |
|     |                                            |                     |     |
|     |                                            |                     |     |
|     | 10  12  13  16  19  24  26  29  5   6   8  |                     |     |
|     |                                            |                     | --- |
*   0 *                                            *                     *     *
|     |                                            |                     |     |
|     | 1   17  2   21  25  27  7                  |                     |     |
|     |                                            |                     |     |
|     |                                            |                     |     |
|     |                                            | vocabulary          |     |
|  -1 + 11  20  22                                 + language use        +     |
|     |                                            |                     |     |
|     |                                            |                     |     |
|     |                                            |                     |  2  |
|     |                                            |                     |     |
|     | 14  15  23  30                             |                     |     |
|  -2 +                                            +                     +     |
|     |                                            |                     |     |
|     |                                            |                     |     |
|     |                                            |                     |     |
|     |                                            |                     |     |
|     |                                            |                     |     |
|  -3 +                                            +                     +     |
|     |                                            | mechanics           |     |
|     |                                            |                     | --- |
|     |                                            |                     |     |
|     |                                            |                     |     |
|     |                                            |                     |     |
|  -4 +                                            +                     + (1) |
|-----+--------------------------------------------+---------------------+-----|
|Measr|-rater                                      |+criterion importance|Scale|
+------------------------------------------------------------------------------+
```

Figure 4.1 Variable Map of Raters' Criterion Importance Ratings

The variability can be interpreted from two perspectives; how raters differed in the perception of the rating criteria and how rating criteria were differently perceived by raters. As can be seen in Table 4.3, whereas the variability across rating criteria measures was substantial, the variability across rater measures was not. This contrasting trend between rater measures and criteria measures can also be evinced by three separation indices (i.e., fixed chi-square, separation index, and separation reliability index).

Table 4.3 Summary Statistics for the Many-facet Rasch Analysis of Raters' Criterion Importance Ratings (N=30)

Statistics	Raters	Rating criteria
Mean measure	-0.08	0.00
Mean SE	0.83	0.34
Chi-square	38.1	206.4*
df	29	4
Separation index	1.15	10.19
Separation reliability	0.27	0.98

Concerning the summary statistics for raters, it seemed that there was not much variability across raters in terms of perceived criterion importance as demonstrated by low chi-square value, low separation index, and low separation reliability.

The congruence in rater perception of rating criteria, however, can be possibly attributed to the limited number of rating criteria, five, involved in the study in comparison to seven criteria used in a previous rater cognition study (Eckes, 2008). It is likely that the more criteria are employed in the criterion importance survey, the more variability among raters can be induced due to a greater number of possible combinations of differentially perceived criterion importance. The difference in the number of rating criteria involved between the present study and Eckes (2008, 2012) may be due to the difference in the type of the writing task at hand. For example, three criteria, namely *Completeness*, *Description*, and *Argumentation* in Eckes (2008) are aspects related to task realization; *Completeness* refers to the degree to which all of the points stated in the task description are addressed; *Description* indicates the degree to which content of a table or a diagram included in the task prompt is concisely summarized; *Argumentation* represents the degree to which pros or cons of a particular view is expressed. Examinees in the present study were merely required to compose a piece of argumentative essay without the need to comply with any further specifications of the task except that they had to express their ideas concerning the given topic. Thus, the rating criteria, which had relevance to task realization and thus were deemed as important in Eckes's study, were excluded from the present study, which led to the relatively smaller number of rating criteria involved in this study. A high mean standard error for rater measures (i.e., 0.83) may have been derived from a lack of information provided concerning raters. That is, as each rater merely assigned five importance ratings across the rating criteria, this may not be enough to model raters. However, the concern of the present research was not how accurately the rater facet can be modeled, but how differently raters attached the importance to the rating criteria. Thus, a high value of SE is not considered to undermine the possibility of clustering raters.

Several incidents falling outside the 0.5/1.5 fit range, however, demonstrated a considerable degree of rater variation. By squaring the standardized residuals and averaging over the elements of the facets involved, two summary statistics, infit and outfit mean-square fit statistics, are obtained (Eckes, 2011). Both infit and outfit statistics have an expected value of 1.0. As the mean square value is the ratio between observed and expected variance, the expected mean square value is 1, which indicates observed variance equals expected variance (Wright and Masters, 1982).

Rater misfit, which is over 1.5, is considered to show more variation than expected whereas rater overfit, whose value is less than 0.5, denotes predictability in a rater’s ratings. Thus, both rater misfit and overfit indicate unexpected responses beyond the model estimation. Contrasting with the lack of variability of rater measures represented by a low chi-square and a low separation index, rater fit statistics showed substantial deviations from model expectations; exactly half the infit values and more than half the outfit values were located outside the 0.5/1.5 fit range as summarized in Table 4.4. Hence, raters were shown to generate heterogeneous patterns of criterion importance ratings.

Table 4.4 Frequencies of Rater Fit Statistics (N=30)

Fit range	Infit		Outfit	
	n	%	n	%
fit < 0.50 (overfit)	10	33.3	12	40
0.50≦fit≦1.50	15	50	13	43.3
fit > 1.50 (misfit)	5	16.7	5	16.7

When it comes to variability across criterion measures, three separation indices sufficiently demonstrated that the rating criteria were perceived differently by raters; the fixed chi-square value (i.e., 206.4) was highly significant, thus rejecting the null hypothesis that all criteria were viewed as similarly important; the separation index (i.e., 10.19) suggested that there were at least 10 statistically distinct strata of criterion importance; lastly, the separation reliability (i.e., 0.98) index showed that criteria were extremely well distinguished in terms of perceived importance.

Taken together, rater variability in terms of perceived criterion importance was not secured, which was assumed to arise from a relatively small number of the rating criteria involved in the survey. However, variability across criteria measures was identified. That is, it was proved that the rating criteria were viewed significantly differently, thus the categorization of raters can possibly be justified.

Another use of a two-facet analysis was to examine whether the importance rating scales functioned as expected; that is, a higher rating should denote higher importance attached to a rating criterion. The successful function of the rating scale can be confirmed by monotonic increases in the category thresholds and average rater importance measures computed per category, which are demonstrated in Table 4.5. In addition, outfit mean square values for the rating scale categories approached the expected value of 1. As denoted by the relative frequencies, the distribution of importance ratings manifested the trend that centered around the middle point, *important* and *very important*.

Table 4.5 Functioning of the Criterion Importance Rating Scale

Category	Freq. %	Threshold	Average measure	Outfit
Less important	13%		-3.02	0.9
Important	39%	-3.18	-1.16	1.0
Very important	31%	0.16	1.13	1.1
Extremely important	17%	3.02	3.49	0.8

As set out in Table 4.6, both infit mean square values and outfit mean square values for all rating criteria fell within the 0.5/1.5 fit range, thus satisfying unidimensionality assumption within the set of rating criteria.

Table 4.6 Measures and Fit Statistics of the Rating Criteria

Rating criteria	Measure	SE	Infit	Outfit
Content	3.43	0.37	0.89	0.97
Organization	1.61	0.32	0.87	0.87
Vocabulary	-0.84	0.33	0.54	0.53
Language use	-1.05	0.33	1.10	1.11
Mechanics	-3.15	0.36	1.41	1.38

To summarize, the two-way facet analysis revealed that rating criteria were well differentiated by raters even though there was little variability across rater measures, which can be attributed to the relatively small number of the rating criteria involved. Thus, based on variability across criteria measures, raters can possibly be categorized into different cognitive rater types (CRTs), which will be presented and discussed in the following.

4.3 Cognitive Rater Types (CRTs)

CRTs, based on perceived criterion importance, were identified as seen in the tree diagram (Figure 4.2), the dendrogram, which was yielded by the hierarchical clustering analysis (SPSS 29). The clustering solution from Ward’s was adopted due to its tendency to maximize the significance of differences between clusters.

To the left of the tree diagram is rater identification denoted by their identification number ranging from 1 to 30. For example, 1 indicates Rater 01. Hierarchical clustering according to the congruence in the criterion importance ratings located raters into more inclusive classes in a way that minimizes the increase in sums of squares, which consequently yielded five CRT clusters, namely CRT A to CRT E. The larger the distance (i.e. from left to right at the bottom of the figure) between raters, the dissimilar they are in terms of their perceived criterion importance as displayed in the distance between CRT A (or CRT B or CRT C or CRT D) and CRT E. Conversely, the minimum distance shown between raters suggests that they have identical or similar views on criterion importance as seen in many cases such as between raters 5 and 29; raters 21, 27, and 13; raters 14 and 20; raters 24, 26, 12, and 8; raters 2, 25 and 1; raters 7, 17, and 6; raters 22 and 23; raters 4, 9, and 3; raters 10, 16, and 28.

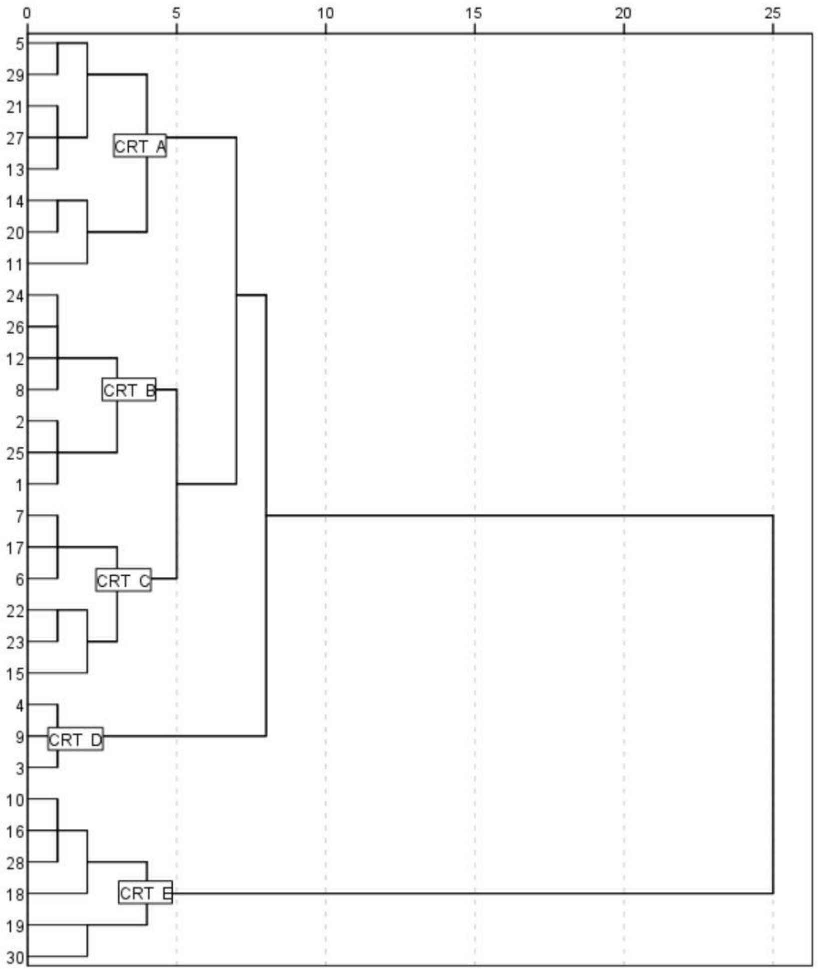


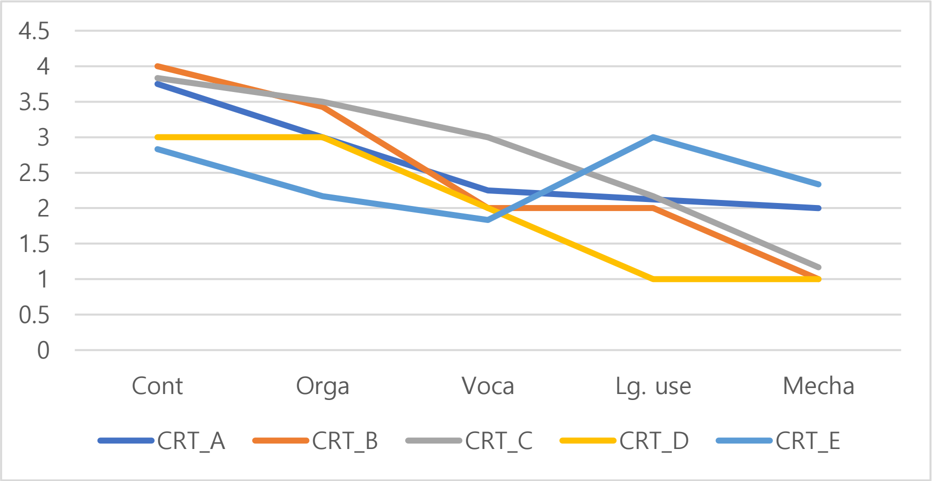
Figure 4.2 Hierarchical Clustering Solution for CRTs

Below is a description of how each CRT type was composed, each having a distinctive pattern in terms of criterion importance. Table 4.7 displays the means of criterion importance ratings among each CRT and Figure 4.3 graphically shows how the criterion importance profile of each CRT was different from one another.

Table 4.7 Means of the Criterion Importance Ratings among each CRT (N=30)

CRT	Rater No.	Cont	Orga	Voca	Lg. use	Mecha
A	5,29,21,27,13,14,20,11	3.75	3	2.25	2.13	2
B	24, 26, 12, 8, 2, 25, 1	4	3.43	2	2	1
C	7, 17, 6, 22, 23, 15	3.83	3.5	3	2.17	1.17
D	4, 9, 3	3	3	2	1	1
E	10, 16, 28, 18, 19, 30	2.83	2.17	1.83	3	2.33

Note: Cont is short for Content; Orga for Organization; Voca for Vocabulary; Lg. use for Language use; Mecha for Mechanics. Extreme importance was marked as 4, moderate importance as either 3 or 2, and less importance as 1 on the rating scale.



Note: Color figure available online.

Figure 4.3 Criterion Importance Profiles for CRTs

As can readily be seen in Table 4.7, CRT A comprised 8 raters (i.e., 5, 29, 21, 27, 13, 14, 20, and 11). Most raters in CRT A shared a tendency to perceive Content as extremely important whereas the remaining criteria were considered moderately important. CRT B was composed of 7 raters (i.e., 24, 26, 12, 8, 2, 25, and 1), all of whom shared a perception of Content as extremely important, but regarded Mechanics as less important and the remaining three criteria as moderately important. CRT C consisted of 6 raters (i.e., 7, 17, 6, 22, 23, and 15), whose perceived criterion importance was very similar to that of CRT B. The difference that made a distinction between these two clusters, however, was that all raters in CRT C marked 3 on Vocabulary while all in CRT B gave 2 for the same criterion. Most raters in CRT C showed extreme importance toward Content and Organization, displaying moderate importance towards Vocabulary and Language use and low importance towards Mechanics. CRT D comprised 3 raters (4, 9, and 3), all of whom perceived none of the criteria as extremely important and gave moderate importance ratings to Content, Organization, and Vocabulary and less importance to Language use and Mechanics. Considering that Language use was considered moderately important in other CRTs, CRT D was distinguishable in that it assigned less importance to Language use. CRT E consisted of 6 raters (i.e., 10, 16, 28, 18, 19, and 30), displaying the most distinctive pattern in the criterion importance among all CRT clusters in that no rating criteria received extreme importance or low importance rating. Most of the raters in CRT E gave moderate importance to all rating criteria. Overall, the findings of the criterion importance ratings were that Content was perceived as extremely important in three clusters (i.e., CRT A, CRT B, and CRT C) whereas Mechanics was assigned low importance in three clusters (i.e., CRT B, CRT C, and CRT D). Organization received moderate importance from four ORTs except for CRT C in which half of the raters perceived Organization as extremely important. Vocabulary was perceived as moderately important by all CRTs. Language use was also viewed as moderately important by all CRTs except for CRT D in which Language use was regarded as less important.

When comparing the present study with Eckes (2008), several patterns of commonalities and differences, in terms of the perceived criterion importance on Organization, Vocabulary, and Language use, emerged. Under the category of Linguistic realization in Eckes's cognition study, there were three criteria: *Syntax*, *Vocabulary,* and *Correctness.* Syntax can be compared to Organization and Language use in this study since Syntax in Eckes's study referred to examinees' ability to use cohesive elements as well as syntactically correct structures. Vocabulary under Eckes (2008) denoted the same operational definition as Vocabulary in this study. Correctness pertains to Vocabulary, Language use, and Mechanics of the present study as it stood for the degree to which the text does not exhibit morphosyntactic, lexical, or orthographical errors in Eckes's study. Thus Eckes' three criteria, *Syntax*, *Vocabulary*, and *Correctness* can be compared to the present study's four rating criteria, Organization, Vocabulary, Language use, and Mechanics.

The commonalities revealed were that extremely high importance was rarely placed on any other rating criteria than Content. To be specific, extremely high importance attached to *Syntax*, *Vocabulary*, and *Correctness* in Eckes (2008) was observed only three times among 18 cases (i.e., 6 CRTs in relation to 3 rating criteria), which was also a pattern that emerged in the outcome of the current study; extremely high importance was observed for the equivalent rating criteria only once (i.e., Organization in CRT C) out of 20 cases (i.e., 5 CRTs in relation to 4 rating criteria). Additionally, another shared observation between these two studies was that a dominant importance rating across the four rating criteria was moderate importance with a relatively small account of low importance identified.

However, rater variability of perceived criterion importance, which can be represented by the incidents of all possible importance ratings, was more clearly observed in Eckes than in the present study. Given that 64 raters participated in a rater cognition study in Eckes (2008), a relatively smaller pool of participants (i.e., 30 raters) in this study can probably be the reason for the lack of criteria perception variability.

Though raters largely seemed to share views on criterion importance, raters in the study were shown to be placed in a different cognitive rater type, exhibiting a distinctive pattern and thus distinguishing one cluster of raters from another. Consequently, concerning the first research question, ‘How can raters be classified into a group based on perceived criterion importance?’, it can be concluded that raters formed different cognitive types based on their perceived criterion importance.

4.4 Many-facet Rasch Analysis on Essay Ratings

4.4.1 Inter-Rater Agreement

Assumed under the use of the FACETS program is that raters are independent experts; raters are presumed to rate according to the expertise in rating based on the same understanding of the construct being measured. They are at the same time expected not to act as “scoring machines” (Linacre, 1998). This implies that raters are not overly dependent on one another (Eckes, 2015). To verify local independence among raters, the Rasch-Kappa index (Linacre, 2014) was consulted, whose value is expected to approach zero when the assumption is satisfied.

Rasch-kappa= (Obs%-Exp%) / (100- Exp%)

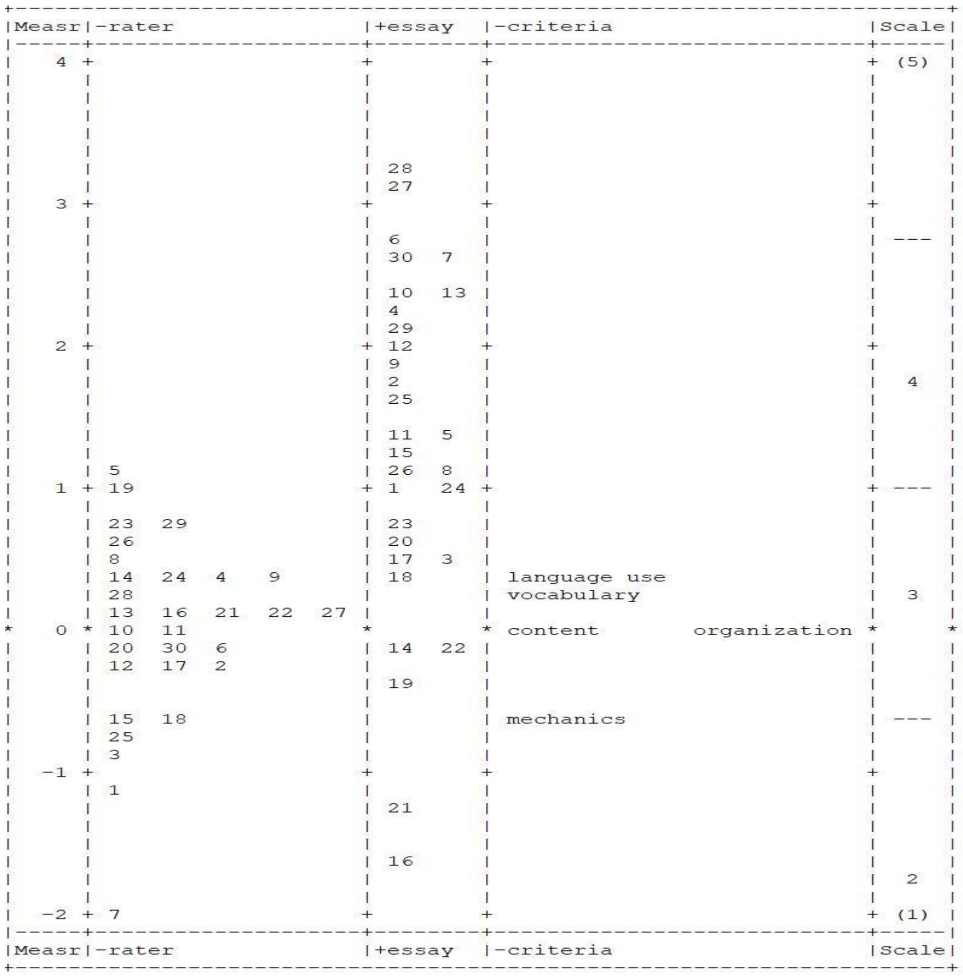
The observed proportions of exact agreements between raters under the rating scale model and the partial credit model were both 37.8%; the expected proportions of exact agreements between raters under the rating scale model and the partial credit model were 36% and 36.6%, respectively. Inserting these proportions into the equation stated above, the Rasch-kappa index, a value of 0.02 was calculated under the two Rasch models, which is close to zero. This indicates that the raters were independent of one another and that they had a common understanding of the rating criteria.

4.4.2 Rater Measurement Results

The overall data-model fit was satisfactory in that the (absolute) standardized residuals equal to or more than 2 accounted for only 0.38% (i.e., 17 out of 4500 cases) without the occurrence of the (absolute) standardized residuals equal to or more than 3 at all. The preset critical percentages for the (absolute) standardized residuals equal to or more than 2 and 3 are 5% and 1%, respectively (Eckes, 2008).

Before conducting a bias analysis between raters and the rating criteria, rater variability in terms of severity, a fixed chi-square value, separation statistics, and intra-rater reliability needs to be discussed first. For a graphical illustration of the complete set of calibrations, the Wright map is provided in Figure 4.4 and Figure 4.5. Figure 4.4 and Figure 4.5 were derived from the three-facet *rating scale model* (RSM) and the three-facet *partial credit model* (PCM), respectively. The RSM presumes that the set of threshold parameters, which define the structure of the rating scale, is the same across all elements of a facet. Thus, in the RSM, it is assumed that raters may use the rating scale in a highly similar manner and that the probability of an examinee receiving a rating between two adjacent categories is 50% provided that the ability of the examinee is in one of those two categories (Andrich, 1998). Alternatively, in the PCM, the threshold parameters are specified in a way that allows for variable rating scale structures across the elements of a facet (Eckes, 2015). For example, the PCM assumes that the difficulty of an examinee receiving a rating of 4 in Content is not equivalent to that of obtaining the same rating in Organization. The PCM also implies an assumption that raters may be inconsistent in interpreting the relative step difficulty of the available scoring categories.

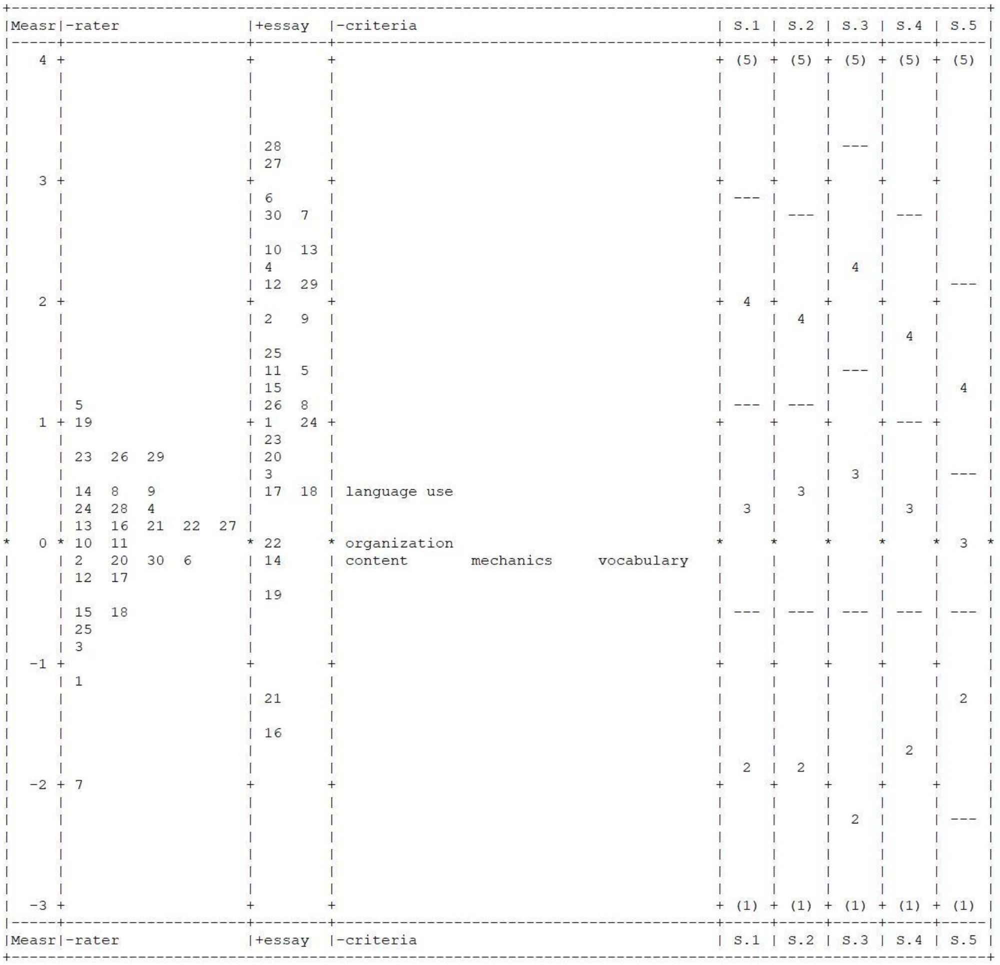
Hence, Figure 4.5 displays different calibrations from those in Figure 4.4., which may have been caused by the differences in the thresholds across the rating criteria. However, notable was that the bias analysis measures between raters and the rating criteria under the RSM and the PCM were similar, thus choosing either the RSM or the PCM does not affect the categorization of raters based on the criterion-related bias.



```
+---------------------------------------------------------------------------+
|Measr|-rater                |+essay  |-criteria                      |Scale|
|-----+----------------------+--------+-------------------------------+-----|
|   4 +                      +        +                               + (5) |
|     |                      |        |                               |     |
|     |                      |        |                               |     |
|     |                      |        |                               |     |
|     |                      |        |                               |     |
|     |                      |        |                               |     |
|     |                      | 28     |                               |     |
|     |                      | 27     |                               |     |
|   3 +                      +        +                               +     |
|     |                      |        |                               |     |
|     |                      | 6      |                               | --- |
|     |                      | 30  7  |                               |     |
|     |                      |        |                               |     |
|     |                      | 10  13 |                               |     |
|     |                      | 4      |                               |     |
|     |                      | 29     |                               |     |
|   2 +                      + 12     +                               +     |
|     |                      | 9      |                               |     |
|     |                      | 2      |                               |  4  |
|     |                      | 25     |                               |     |
|     |                      |        |                               |     |
|     |                      | 11  5  |                               |     |
|     |                      | 15     |                               |     |
|     | 5                    | 26  8  |                               |     |
|   1 + 19                   + 1   24 +                               + --- |
|     |                      |        |                               |     |
|     | 23  29               | 23     |                               |     |
|     | 26                   | 20     |                               |     |
|     | 8                    | 17  3  |                               |     |
|     | 14  24  4   9        | 18     | language use                  |     |
|     | 28                   |        | vocabulary                    |  3  |
|     | 13  16  21  22  27   |        |                               |     |
*   0 * 10  11               *        * content       organization    *     *
|     | 20  30  6            | 14  22 |                               |     |
|     | 12  17  2            |        |                               |     |
|     |                      | 19     |                               |     |
|     |                      |        |                               |     |
|     | 15  18               |        | mechanics                     | --- |
|     | 25                   |        |                               |     |
|     | 3                    |        |                               |     |
|  -1 +                      +        +                               +     |
|     | 1                    |        |                               |     |
|     |                      | 21     |                               |     |
|     |                      |        |                               |     |
|     |                      |        |                               |     |
|     |                      | 16     |                               |     |
|     |                      |        |                               |  2  |
|     |                      |        |                               |     |
|  -2 + 7                    +        +                               + (1) |
|-----+----------------------+--------+-------------------------------+-----|
|Measr|-rater                |+essay  |-criteria                      |Scale|
+---------------------------------------------------------------------------+
```

Figure 4.4 Variable Map from the Many-facet Rasch Analysis of 30 Essay Ratings under the Rating Scale Model

Note: In the second column, numbers represent 30 participants, who also participated in the criterion importance rating survey. In the third column, numbers indicate 30 essays, which were assigned ratings by participants. The horizontal dashed lines in the fifth column indicate the category threshold measures for the five-category rating scale.



```
+------------------------------------------------------------------------------------------------------------+
|Measr|-rater              |+essay  |-criteria                                 | S.1 | S.2 | S.3 | S.4 | S.5 |
|-----+--------------------+--------+------------------------------------------+-----+-----+-----+-----+-----|
|   4 +                    +        +                                          + (5) + (5) + (5) + (5) + (5) |
|     |                    |        |                                          |     |     |     |     |     |
|     |                    |        |                                          |     |     |     |     |     |
|     |                    |        |                                          |     |     |     |     |     |
|     |                    |        |                                          |     |     |     |     |     |
|     |                    | 28     |                                          |     |     | --- |     |     |
|     |                    | 27     |                                          |     |     |     |     |     |
|   3 +                    +        +                                          +     +     +     +     +     |
|     |                    | 6      |                                          | --- |     |     |     |     |
|     |                    | 30  7  |                                          |     | --- |     | --- |     |
|     |                    |        |                                          |     |     |     |     |     |
|     |                    | 10  13 |                                          |     |     |     |     |     |
|     |                    | 4      |                                          |     |     |  4  |     |     |
|     |                    | 12  29 |                                          |     |     |     |     | --- |
|   2 +                    +        +                                          +  4  +     +     +     +     |
|     |                    | 2   9  |                                          |     |  4  |     |     |     |
|     |                    |        |                                          |     |     |     |  4  |     |
|     |                    | 25     |                                          |     |     |     |     |     |
|     |                    | 11  5  |                                          |     |     | --- |     |     |
|     |                    | 15     |                                          |     |     |     |     |  4  |
|     | 5                  | 26  8  |                                          | --- | --- |     |     |     |
|   1 + 19                 + 1   24 +                                          +     +     +     + --- +     |
|     |                    | 23     |                                          |     |     |     |     |     |
|     | 23  26  29         | 20     |                                          |     |     |     |     |     |
|     |                    | 3      |                                          |     |     |  3  |     | --- |
|     | 14  8   9          | 17  18 | language use                             |     |  3  |     |     |     |
|     | 24  28  4          |        |                                          |  3  |     |     |  3  |     |
|     | 13  16  21  22  27 |        |                                          |     |     |     |     |     |
*   0 * 10  11             * 22     * organization                             *     *     *     *     *  3  *
|     | 2   20  30  6      | 14     | content       mechanics     vocabulary   |     |     |     |     |     |
|     | 12  17             |        |                                          |     |     |     |     |     |
|     |                    | 19     |                                          |     |     |     |     |     |
|     | 15  18             |        |                                          | --- | --- | --- | --- | --- |
|     | 25                 |        |                                          |     |     |     |     |     |
|     | 3                  |        |                                          |     |     |     |     |     |
|  -1 +                    +        +                                          +     +     +     +     +     |
|     | 1                  |        |                                          |     |     |     |     |     |
|     |                    | 21     |                                          |     |     |     |     |  2  |
|     |                    |        |                                          |     |     |     |     |     |
|     |                    | 16     |                                          |     |     |     |     |     |
|     |                    |        |                                          |     |     |     |  2  |     |
|     |                    |        |                                          |  2  |  2  |     |     |     |
|  -2 + 7                  +        +                                          +     +     +     +     +     |
|     |                    |        |                                          |     |     |     |     |     |
|     |                    |        |                                          |     |     |  2  |     | --- |
|     |                    |        |                                          |     |     |     |     |     |
|     |                    |        |                                          |     |     |     |     |     |
|     |                    |        |                                          |     |     |     |     |     |
|     |                    |        |                                          |     |     |     |     |     |
|  -3 +                    +        +                                          + (1) + (1) + (1) + (1) + (1) |
|-----+--------------------+--------+------------------------------------------+-----+-----+-----+-----+-----|
|Measr|-rater              |+essay  |-criteria                                 | S.1 | S.2 | S.3 | S.4 | S.5 |
+------------------------------------------------------------------------------------------------------------+
```

Figure 4.5 Variable Map from the Many-facet Rasch Analysis of 30 Essay Ratings under the Partial Credit Model

Note: S1 through S5 denotes Content, Organization, Vocabulary, Language use, and Mechanics, respectively.

Rater variability in terms of the level of severity is represented by a 3.08-logit spread in the RSM and a 3.11-logit spread in the PCM between the most severe rater (rater 5 from CRT A) and the least severe rater (rater 7 from CRT C). Except for rater 7, who displayed distinctive leniency, the approximate severity difference among raters was around 2.2 logits, which was much smaller than the logit spread of essays, approximately 4.95 logits.

Raters displayed variability in terms of the essay ratings, as supported by figures in Table 4.8. The separation statistics (Myford & Wolfe, 2003) provided evidence that rater severity measures were by no means homogeneous; in the RSM a fixed chi-square was highly significant, Q(29)= 812.1, $p < .001$, rejecting the null hypothesis that all raters were identical in terms of rating severity. Similarly, in the PCM a fixed chi-square was highly significant, Q(29)= 824.5, $p < .001$, rejecting the null hypothesis that all raters were identical in terms of rating severity. Hence, the significant Q index showed that at least two raters did not share the same severity level after accounting for measurement error. In addition, the separation index (i.e., H) suggested that there were between seven to eight statistically distinct strata of severity (H= 7.89 in RSM and H=7.94 in the PCM). Lastly, in terms of separation reliability, R was 0.97 under the two Rasch models, thus proving that raters can be well-differentiated in terms of severity.

Table 4.8 Summary Statistics for the Many-facet Rasch Analysis of Essay Ratings in Two Rasch Models (N=30)

Statistics	Raters in the RSM	Raters in the PCM
Chi-square (Q)	812.1*	824.5*
df	29	29
Separation index (H)	7.89	7.94
Separation reliability (R)	0.97	0.97

4.4.3 Intra-Rater Reliability

Another issue concerning rater effects is how consistently an individual rater employed the rating scale across examinees and rating criteria. Various factors were indicated as contributing to weakening rater consistency in the use of the rating scale; severity or leniency toward particular examinees or rating criteria can affect intra-rater consistency adversely; furthermore, unsystematic or unexpected factors such as changes in scoring conditions, rater fatigue, or transcription errors can also prevent raters from maintaining consistency in rating writing performances. Fit statistics are crucial in that they identified the degree to which each element is aligned with model expectations. Thus, the fit statistics examine the pattern of the *residuals*, the gap between the observed and the expected score through either a *mean square* value or *t* (McNamara, 1996). Infit and outfit mean square statistics, all of which have an expected value of 1.0, were consulted to examine intra-rater consistency.

The relevant fit range is between mean ± twice the standard deviation of the mean square statistics (McNamara, 1996). The infit statistics, due to their higher estimation precision, are favored over the outfit statistics when judging rater fit (Eckes, 2011; McNamara, 1996; Myford &Wolfe, 2003). Concerning rater misfits, those who have infit values greater than the upper limit (i.e., the mean plus twice the standard deviation) can be said to have a tendency to rate essays unexpectely, exhibiting more variation than expected in the ratings. Conversely, raters with less than the lower limit (i.e., the mean minus twice the standard deviation) are considered to show less variation than expected, thereby rendering their ratings too predictable and failing to provide any information that other raters do not give; this is called overfit.

The mean and the standard deviation for the infit value of the present study under both the RSM and the PCM were almost the same: 1.0 and 0.27 under the RSM and 1.01 and 0.27 under the PCM. Thus, the normal fit range based on the RSM was between 0.46 logits and 1.54 logits (i.e., 1±0.27×2) and the normal fit range based on the PCM was between 0.47 logits and 1.55 logits (i.e., 1.01±0.27×2). As three raters, rater 7 (1.58 logits), rater 18 (1.66 logits), and rater 29 (1.73 logits), were located over the upper limit, and thus were identified as misfits, these three raters were excluded from the following research procedure. Other than these three raters, no raters were identified as misfits or overfits. The fit statistics of all raters except for these three raters (i.e., rater 7, rater 18, and rater 29) fell between 0.46 and 1.54 and between 0.47 and 1.55, suggesting that a majority of raters in the present study scored essays in a consistent manner. Hence, the essay ratings from 27 participants, therefore, were subject to clustering raters based on criterion-related bias.

4.5 Operational Rater Types (ORTs)

This section is directly related to the second research question of the present study, which asks how groups of raters can be differentiated from one another in terms of severity toward particular rating criteria. Following the label Eckes (2012) used to designate the clusters formed based on criterion-related bias measures, the term Operational Rater Types (ORTs) was adopted in the present study. Implied under operational rater types is that raters belonging to the same operational cluster share a certain severity and/ or leniency scoring pattern toward particular rating criteria, which is distinctive from those of other operational rater clusters.

To identify raters who display a criterion-related bias, or show severity or leniency bias on particular rating criteria, two MFRM two-way interaction (i.e., rater × rating criteria) analyses were conducted, one based on the RSM and the other based on the PCM. The results of the MFRM analyses became the input to the categorization of raters into ORTs. Hence, raters sharing a similar scoring profile were to belong to the same ORT. The bias measures from these two Rasch models were almost the same, thus generating the same hierarchical clustering solution for raters based on criterion-related bias measures.

Figure 4.6 shows the tree diagram yielded by hierarchical clustering based on MFRM bias measures (in logits) in which six ORTs were identified. Ward’s method employed in the formation of CRTs was also used in generating ORTs. As seen in Figure 4.6, the number of raters for the set of ORTs through ORT 1 to ORT 6 was 4, 3, 3, 5, 3, and 9 respectively.

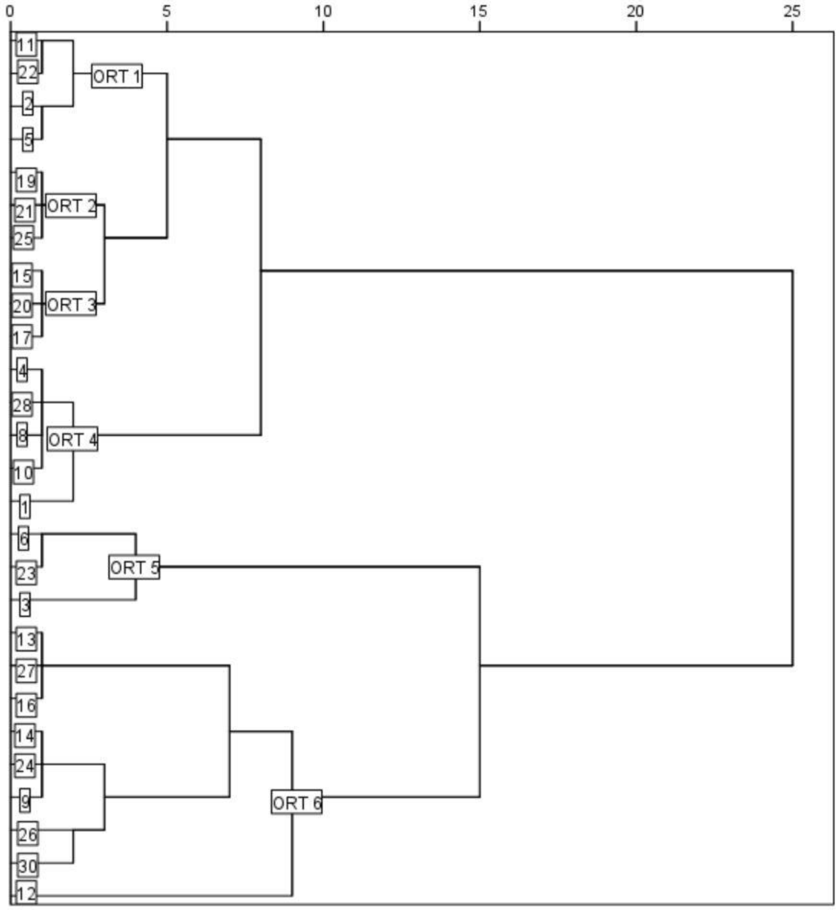
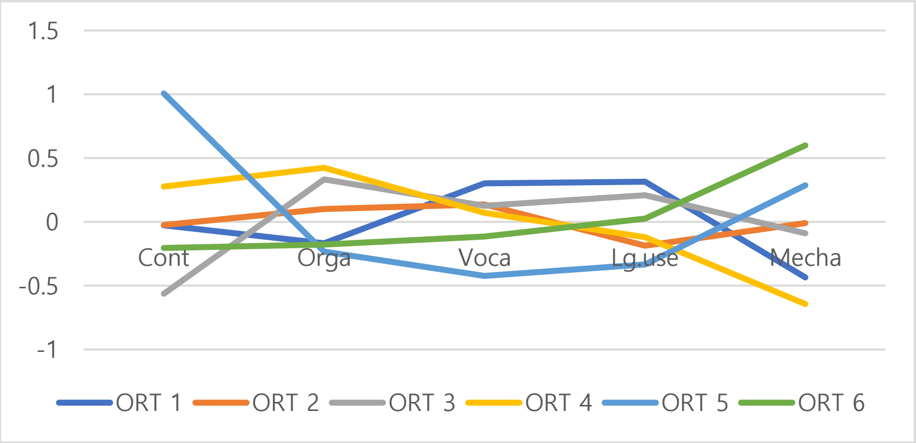


Figure 4.6 Hierarchical Clustering Solution for ORTs

A bias profile, a pattern of criterion-related bias, can be compared among the set of six ORTs as seen in Figure 4.7 (bias diagram) and Table 4.9. In Figure 4.7, each scoring profile represents the mean biases of five rating criteria across the raters in each ORT. Raters in the same ORT exhibited a shared criterion-related bias pattern, which was distinguishable from that of other ORTs. Looking at Figure 4.7, it was easy to see that the range of the bias measure was the largest (i.e., 1.57 logits) in Content. The second largest bias range was observed in Mechanics (1.24 logits), followed by the bias range of Vocabulary (0.73 logits), Organization (0.66 logits), and Language use (0.65 logits). Scoring profiles for Organization, Vocabulary, and Language use seemed to be similar across ORTs due to the relatively small bias size of less than (absolute) 0.5 logits.



Note: The mean bias measure of each ORT obtained per rating criterion is shown in logits (positive values representing leniency and negative values indicating severity). Cont is short for Content; Orga for Organization; Voca for Vocabulary; Lg. use for Language use; Mecha for Mechanics (color figure available online).

Figure 4.7 Bias Diagram for ORT 1 through ORT 6

Table 4.9 Mean Bias Measures among each ORT (N=27)

Criteria	ORT 1	ORT 2	ORT 3	ORT 4	ORT 5	ORT 6
Content	-0.03	-0.02	*-0.56(S)	0.28	*1.01(L)	-0.20
Organization	-0.17	0.10	0.33	0.42	-0.23	-0.18
Vocabulary	0.30	0.14	0.13	0.07	-0.42	-0.11
Lg. use	0.32	-0.19	0.21	-0.12	-0.33	0.03
Mechanics	-0.44	-0.01	-0.09	*-0.64(S)	0.29	*0.60(L)

Note: The average bias measures (in logits) across raters in each ORT were calculated. Four bias cases were asterisked with S and L denoting severity and leniency respectively.

Table 4.9 shows the average bias measures (in logits) of each ORT across the rating criteria. Positive bias values indicate that the observed score, the assigned score, was higher than the expected score, showing that raters rated leniently. Conversely, negative bias values suggest that the observed score was lower than the expected score, signifying a rater’s tendency to rate severely. ORT 1 showed leniency toward Vocabulary and Language use and severity toward Content, Organization, and Mechanics. ORT 2 displayed leniency for Organization and Vocabulary and severity for Content, Language use, and Mechanics. ORT 3 showed leniency for Organization, Vocabulary, and Language use and severity for Content and Mechanics. ORT 4 exhibited leniency for Content, Organization, and Vocabulary and severity for Language use and Mechanics. ORT 5 displayed leniency toward Content and Mechanics and severity for Organization, Vocabulary, and Language use. ORT 6 revealed leniency toward Language use and Mechanics and severity toward Content, Organization, and Vocabulary. What needs to be noted about the composition of the raters was that among all 27 raters, 9 raters were placed in ORT 6, accounting for 33% of the total participants. In addition, all nine raters in ORT 6 displayed leniency toward Mechanics and six of them showed leniency bias higher than positive 0.5 logits. Conversely, concerning the composition of ORT 4 having the second most number of raters, all five raters in ORT 4 exhibited severity toward Mechanics, and four of them showed severity bias smaller than negative 0.5 logits. Thus, it can be possibly said that Mechanics was the rating criterion toward which a large portion of raters in the study displayed biases.

A so-called half-logit rule (Draba, 1977) has been reliably applied to detecting rating criteria that were rated more severely or more leniently than expected (Eckes, 2012). Therefore, applying (absolute) 0.5 logits as the limit for detecting bias, bias measures over 0.5 logits were interpreted as signaling leniency toward the corresponding rating criteria whereas bias measures less than negative 0.5 logits were considered to signify severity bias. As seen in Table 4.9, four cases were identified as displaying a criterion-related bias: severity toward Content in ORT 3 and toward Mechanics in ORT 4 and leniency toward Content in ORT 5 and toward Mechanics in ORT 6. To illuminate, concerning Content and Mechanics respectively, a pair of ORTs displayed a bias measure with an opposite direction; whereas ORT 3 showed severity bias represented by negative 0.56 logits toward Content, ORT 5 displayed leniency bias with 1.01 logits for the same criterion; whereas ORT 4 exhibited severity bias with negative 0.64 logits toward Mechanics, ORT 6 showed leniency bias with 0.6 logits for the same criterion. Additionally, it was found that no biases were observed in Organization, Vocabulary, and Language use across all ORTs.

Taken together, although most bias cases had bias measures less than (absolute) 0.5 logits except for 4 bias cases, it was observed that all ORTs had a distinctive scoring profile from one another. Thus, concerning the second research question it can be said that a group of raters can be differentiated from one another based on the severity or leniency shown toward particular rating criteria.

When it comes to criterion-related biases, the outcome of previous literature can be consulted. There were mixed findings as to the interactions between raters and rating criteria. Raters were found to rate grammar more harshly than any other aspect of writing performance (Lumley & McNamara, 1995; Wigglesworth, 1993). Shin (2010), in which three Korean raters rated the essays written by Korean English learners, similarly uncovered that raters rated Grammar most harshly, but Organization the most leniently. Schaefer (2018) found that subgroups of raters exhibited the opposite patterns of scoring bias; one subgroup of raters showed severity toward Content and Organization and leniency toward Language use and Mechanics whereas another subgroup of raters displayed the reversed pattern. Shaefer's finding, however, did not address the reason behind the emergence of the opposite scoring patterns only to demonstrate that raters on a group basis could be distinguished in terms of criterion-related bias. Later, Eckes (2012) attributed the difference in scoring behavior among subgroups of raters in Schaefer's study to raters' different perceptions of criterion importance.

Observed was that previous studies appeared to dichotomize rating criteria into a content-related area (and an organization-related area) versus a convention-related area (e.g., content versus grammar or content and organization versus language use and mechanics). Concerning this, Schaefer (2008) suggested that raters seemed to undertake compensatory rating strategies in which a tendency to rate particular rating criteria severely is compensated for by a tendency to rate the other rating criteria leniently. The employment of this compensatory strategy, however, was not revealed in the present study, which was supposed to be due to the limited number of bias cases.

4.6 Relation between Criteria Perception and Scoring Behavior

Two investigations into the relationship between perceived criterion importance and criterion-related bias were conducted: one based on the measures from groups of raters and the other based on the measures from individual raters.

4.6.1 Group-Based Investigation

To examine the relationship between criteria perception and scoring behavior on a group of raters basis, the mean of the bias measures of raters from each ORT was compared with the mean of the criterion importance ratings of the raters belonging to the same ORT.

The raters in each ORT are presented in reference to CRTs in Table 4.10. All of the ORTs comprised raters coming from different CRTs; ORT 1 was from CRT A, B, and C; ORT 2 was from CRT A, B, and E; ORT 3 was from CRT A and C; ORT 4 was from CRT B, D, and E; ORT 5 was from CRT C and D; ORT 6 from CRT A, B, D, and E. None of the ORTs was composed of raters belonging to the same CRT.

Table 4.10 Rater Composition of ORTs in Relation to CRTs (N=27)

CRT ORT	A	B	C	D	E
1	5, 11	2	22		
2	21	25			19
3	20		15,17		
4		1, 8		4	10, 28
5			6, 23	3	
6	13, 14, 27	12, 24, 26		9	16, 30

Note: Numbers denote raters.

A casual glance tells that there seemed to be little association between CRTs and ORTs due to a lack of congruence of perceived criterion importance among raters in each ORT. If criterion perception affects scoring behavior, it may well be that ORTs comprise raters who belong to the same CRT or CRTs displaying a relatively short distance in the dendrogram of the CRT hierarchical clustering solution. However, all ORTs were composed of raters from different CRTs. In addition, considering that CRT B and CRT C were similar to each other, most of the ORTs except for ORT 1 comprising raters from CRT B and CRT C, seemed to suggest that raters sharing similar criteria perception in effect did not share a similar scoring pattern.

However, a more in-depth examination of the relationship between criteria perception and scoring behavior is needed, which entails two types of information (i.e., the mean of perceived criterion importance ratings and of criterion-related bias measures). Table 4.11 presents the mean of criterion importance ratings of raters from each of the four ORTs displaying bias measures. Since only four criterion bias cases out of 30 (i.e., 6 ORTs × 5 rating criteria) were identified, the analysis of the relationship between these four biases and the corresponding criterion importance ratings was conducted. Regarding the representative criterion importance rating of ORTs, as stated in Data Analysis, the mean of raters' criterion importance ratings was adopted.

Table 4.11 Four Criterion-Related Bias Cases and the Means of the Criterion Importance Ratings (N=27)

Criterion bias (logits)	The mean of the criterion importance rating	Rating criteria	ORT No.
Severity (-0.56)	4	Content	3
Leniency (1.01)	3.3	Content	5
Severity (-0.64)	1.4	Mechanics	4
Leniency (0.60)	1.7	Mechanics	6

Note: Extreme importance was marked as 4, moderate importance as either 3 or 2, and less importance as 1 on the rating scale of the criterion importance survey.

It needs to be noted that raters in the study were shown to place different weights across the rating criteria even though there were no differential weights prescribed on the rating scale employed in the present study. For ease of reference, the content of Table 4.1 is repeated; the means of criterion importance rating were 3.62 for Content, 3.12 for Organization, 2.27 for Vocabulary, 2.12 for Language use, and 1.54 for Mechanics. No raters in the study assigned 1 rating to Content and 4 to Mechanics. Therefore, the judgment based on the absolute value of the importance rating would not be appropriate in such a way that only the importance rating of 4 is interpreted as important and the rating of 1 as less important. Rather, it would be more sensible to consider the mean of the criterion importance ratings, which were calculated across all 30 raters when examining the relationship between criteria perception and criterion-related bias.

Four biases were observed in only the two rating criteria, Content and Mechanics. As for Content, severity bias (negative 0.56 logits) was observed in ORT 3 with the corresponding criterion importance rating of 4. All raters in ORT 3 assigned 4 to Content. Leniency bias (positive 1.01 logits) toward Content was identified in ORT 5 and the corresponding importance rating was 3.3. Considering that more than half of raters in the criterion importance survey assigned 4 to Content and that the mean criterion importance rating of Content across all 30 raters was 3.62, it can be possibly said that the criterion perceived as less important (i.e., 3.3) than the average importance perception was rated more leniently. Similarly, it can also be said that the criterion perceived as more important (i.e., 4.0) than the average importance perception was rated more severely. Thus, it seemed that these results concerning Content may confirm the finding of Eckes (2012) that “criteria perceived as highly important were differentially associated with a severity bias.”

However, the patterns between bias measures and criterion importance ratings shown in Content were reversed towards Mechanics. Severity bias (negative 0.64 logits) was displayed in ORT 4 toward Mechanics with the corresponding criterion importance rating of 1.4. Leniency bias (positive 0.60 logits) toward Mechanics was exhibited in ORT 6 with the corresponding criterion importance rating of 1.7. These results concerning Mechanics seemed to contradict the finding of Eckes (2012) that severity bias is only combined with high importance and that leniency is only aligned with less importance. In this study, Mechanics of ORT 6, which received a criterion importance rating higher than the average, was rated more leniently. Similarly, Mechanics of ORT 4, which received a criterion importance rating lower than the average, was rated more severely. Thus, it can be said that the nature of the link between perceived criterion importance and scoring behavior was different depending on which criteria were involved.

Due to the attribute of the mean values as being sensitive to outlier values, which thus may have interfered with the investigation into the exact relationship between the bias measure and the criterion importance rating, the following section is based on the data from individual raters, which does not involve the mean values.

4.6.2 Individual Rater-Based Investigation

As stated previously, employing the bias measure and the criterion importance ratings from individual raters to probe the relationship between perceived criterion importance and scoring behavior can defy the limitation of the mean values as not specifying an individual data case. However, at the same time, it has also a limitation with generalizing the finding if there are only a few cases involved. MFRM analyses on individual raters based on both RSM and the PCM discovered that the results of the two Rasch models were very similar in terms of the number of bias cases and the size of each bias, as presented in Table 4.12 and Table 4.13. It was also found that severity and leniency biases were observed toward all five rating criteria, which thus differed from the ORT observations that criterion-related biases were shown towards only the two criteria, Content and Mechanics.

Table 4.12 Criterion-Related Bias Measures under the Rating Scale Model and the Criterion Importance Ratings (N=27)

Criteria Rater No.	Content	Organization	Vocabulary	Language use	Mechanics
1		0.91(4)			
2			*0.75(2)		-0.75(1)
3	*1.61(3)		-0.66(2)	-0.59(1)	*0.66(1)
4					-0.75(1)
5					*-0.52(2)
6	*0.94(3)				
8	0.54(4)				-0.58(1)
9					*0.59(1)
10					*-0.65(2)
12		-0.9(3)		*0.65(2)	*1.66(1)
13					0.53(2)
16				*-0.71(3)	0.69(2)
17	*-0.71(4)				
20	*-0.64(4)				
26	*-0.63(4)			*0.54(2)	
27					0.54(2)
28					*-0.78(2)
30		-0.77(3)			0.59(3)

Note: Parenthesized are the corresponding criterion importance ratings of the corresponding rater. Asterisked are the indications of severity bias combined with importance higher than the average perception and of leniency bias aligned with importance lower than the average perception.

Table 4.13 Criterion-Related Bias Measures under the Partial Credit Model and the Criterion Importance Ratings (N=27)

Criteria / Rater No.	Content	Organization	Vocabulary	Language use	Mechanics
1		0.85(4)			-0.53(1)
2			*0.84(2)		-0.73(1)
3	*1.64(3)		-0.63(2)	-0.6(1)	*0.56(1)
4		*0.5(3)			-0.66(1)
6	*0.97(3)				
8	0.55(4)				
9					*0.57(1)
10					*-0.63(2)
12		-0.9(3)		*0.67(2)	*1.56(1)
16				*-0.71(3)	0.64(2)
17	*-0.72(4)				
20	*-0.65(4)				
26	*-0.67(4)			*0.54(2)	
27					0.5(2)
28					*-0.69(2)
30		-0.77(3)			0.53(3)

Differences in the bias measures obtained from both the RSM and the PCM were so insignificant as not to affect the analysis of the relation between the perceived criterion importance and scoring behavior. As the mean criterion importance ratings, which were calculated across all raters, were applied to the group-based analysis to determine the level of criterion importance, the level of a rater's criterion perception was judged as either higher or lower in comparison to the average criterion importance rating. Asterisked are the indications of the link between perceived criterion importance ratings and criterion-related biases, accounting for 83% (5/6) in Content, 0% (0/3) in Organization, 50% (1/2) in Vocabulary, 75% (3/4) in Language use, and 46% (6/13) in Mechanics under the RSM. Similarly, this pattern took up 83% (5/6) in Content, 25% (1/4) in Organization, 50% (1/2) in Vocabulary, 75% (3/4) in Language use, and 46% (5/11) in Mechanics. As the number of bias cases in Organization, Vocabulary, and Language use was not sufficient to draw a conclusion as to the link between the criteria perception and scoring behavior, the discussions concerning these three rating criteria were excluded.

Taken together, the analysis based on the individual raters under two Rasch models similarly revealed that the relationship between perceived criterion importance and scoring behavior varied depending on the rating criteria involved. In Content, severity bias tended to be aligned with importance higher than the average perception, and leniency bias was more likely to be combined with importance lower than the average perception. The same pattern, however, was not displayed in Mechanics in which severity bias was combined with either higher or lower criterion importance than the average criterion perception, and leniency bias was also aligned with either higher or lower criterion importance than the average criterion perception. Thus, the commonality of the outcomes between group-based and individual rater-based measures was that the alignment of severity bias with higher criterion importance was observed in Content whereas the same phenomenon was not displayed in Mechanics. Therefore, as for the third research question asking ‘To what extent is perceived criterion importance related to scoring behavior?’, it can be concluded that the effects of perceived criterion importance on scoring behavior vary depending on which rating criteria were involved.

The disparity in the findings between the two studies could possibly be attributed to the differences in the procedures of the two research. First of all, as previously indicated as a limitation in Eckes (2008, 2012), there was a lapse of time between the data collection for criterion importance ratings and the live scoring sessions, thus potentially not taking into account the possibility that the perceived criterion importance of the raters could have changed in the time between these two procedures. To address this shortcoming, the present study aimed to conduct the data collection and essay scoring without a gap in time, which may have generated different results from those of Eckes.

Secondly, it should be noted that not all the participants involved in the cognition study (Eckes, 2008) participated in the following study (Eckes, 2012); whereas the data from all 64 participants were analyzed as the source of the formation of CRTs, only the data obtained from 18 participants were analyzed to identify ORTs. The analysis of the link between criteria perception and scoring behavior, thus, was based on a different pool of subjects, which can potentially undermine the validity of the link Eckes found between criteria perception and scoring behavior. Responding to this limitation, the present study involved the same participants in the formation of CRTs and ORTs, generating a different result; that is, the effect of perceived criterion importance on scoring behavior varied depending on which rating criteria were involved.

Third, when analyzing the relation between criteria perception and scoring behavior, whereas Eckes (2012) used the common importance marker to identify the correspondence between CRTs and ORTs, the present study opted to use the mean value of importance ratings among raters in each ORT because the mean value of importance ratings could better represent the nature of criterion perception. Thus, choosing the mean criterion importance value of the raters who belonged to the same ORT, but came from different CRTs did not allow the investigation of the direct relationship between CRTs and ORTs.

CHAPTER 5

CONCLUSION

This study investigated the possibility of clustering raters in terms of their perceptions of criterion importance (i.e., CRTs) and their rating bias (i.e., ORTs) and analyzed the relation between criteria perception and scoring behavior, which is ultimately believed to provide insight into how rater cognition affects scoring behavior. Chapter 5 summarizes the major findings and implications of the present study and suggests limitations and areas of further research.

5.1 Findings and Implications

The primary aim of the study was to investigate how rater cognition, specifically rater perceived criterion importance, can affect scoring behavior based on the hypothesis that criteria perceived as highly important are likely to be rated more severely and criteria viewed as less important are likely to be rated leniently. To investigate the relationship between rater cognition and scoring behavior, two types of clusters (i.e., CRTs and ORTs) were identified and the nature of the link between criteria perception and scoring behavior was analyzed.

CRTs were derived from hierarchical clustering through which raters sharing similar perceptions of criterion importance were grouped and differentiated from raters from other CRTs. ORTs were created based on the bias analysis between raters and the rating criteria, so that raters in the same ORT displayed a similar bias pattern and differed from those of other ORTs.

The finding indicated that even though each five CRTs was distinctive in its criterion importance profile, a lack of rater variability of perceived criterion importance was observed. On the surface, it seemed that most participants attached substantial importance to Content and less importance to Mechanics as the mean of the importance ratings for Content was the highest, which was followed by Organization, Vocabulary, Language use, and Mechanics. This suggests that Korean English secondary teachers generally tend to put more importance on Content than Mechanics.

A finding in relation to ORTs was that raters can be differentiated based on criterion-related bias; the bias profile of each ORT was distinguished from that of other ORTs. Notable was that among 30 possible interaction cases (i.e., five rating criteria and six ORTs) only four interactions were indicated as biased. Interestingly, all four biases pertained to Content and Mechanics and there were no other bias cases displayed toward the other three rating criteria (i.e., Organization, Vocabulary, and Language use). Similarly, based on the analysis of individual raters, it was revealed that approximately 68% and 63% of the total biases under the RSM and the PCM respectively were derived from Content and Mechanics. From this observation, it could be tentatively concluded that raters are homogeneous in terms of interpreting and applying Organization, Vocabulary, and Language use criteria, whereas they differ in the interpretation and the application of Content and Mechanics.

Finally, concerning the relation between criteria perception and scoring behavior, the hypothesized link between these variables was only substantiated towards Content in a group-based investigation; severity bias was aligned with the importance rating higher than the average criterion perception and leniency bias was combined with the importance rating lower than the average criterion perception. However, this pattern was not confirmed in Mechanics since the opposite pattern was observed. To address the limitation of the mean values of a group-based analysis, which are sensitive to outliers and unable to explain an individual rater's measures, the analysis between the bias measure and the criterion importance rating was conducted on an individual rater basis. A similar result to that of the group-based analysis was obtained; the link between perceived criterion importance and scoring behavior was only confirmed in Content whereas the same link was not identified in Mechanics. In Mechanics, both severity and leniency bias were combined with the criterion importance which was either higher or lower than the average criterion perception. Even though bias cases were shown in Organization, Vocabulary, and Language use in the individual rater-based investigation, these were not subject to the analysis due to the limited number of bias cases in these three rating criteria. Therefore, evidence gathered both from groups of raters and individual raters demonstrated that the effects of perceived criterion importance on scoring behavior vary depending on the rating criteria involved.

The present study provides methodological and practical implications in the field of performance assessment, and in particular, writing assessments. First, concerning methodological implications, the current study empirically examined the relationship between rater criterion perception and scoring behavior. Previous research studying rater effects has been centered around identifying the sources of rater variability and describing scoring patterns (e.g., Caban, 2003; Eckes, 2012; Johnson & Lim, 2009; Kondo-Brown, 2002; McNamara, 1996; Park, 2012; Park & Shim, 2014; Schaefer, 2008; Shin, 2010; Wigglesworth, 1993 to name a few). To the researcher’s knowledge, Eckes (2012) is the sole study that explained rater scoring bias from a cognitive perspective.

Second, when it comes to practical implications, the finding of the study can be employed in the context of rater training. The number of criterion-related biases varied according to rating criteria. In a group-based investigation displaying four biases, two were from Content, and the rest of the two were from Mechanics. In an individual rater-based investigation, biases in Mechanics accounted for the largest portion of the total biases (i.e., 46% and 41% in the RSM and PCM, respectively), followed by the number of biases in Content (i.e., 21% and 22 % in the RSM and PCM, respectively). Thus, it can be said that extreme criterion importance can potentially introduce rater bias on the grounds that the mean criterion importance was highest in Content and lowest in Mechanics. Reflecting on this evidence, rater training could be enriched by helping raters to take more balanced views on criterion importance, which can ultimately lead to a decrease in the occurrences of criterion-related biases.

A probe into the scoring patterns of three rater misfits (i.e., rater 7, 18, 29) can also suggest the importance of the rater training aimed at addressing raters' inconsistencies in applying the rating scale. The commonality of these three misfit raters was that they all displayed uncharacteristically lower correlations between a single rater to the rest of the raters than what FACETS expected. Low correlations signal random rater effects, which indicate that those raters may not have been able to differentiate examinees' performances on the trait being measured (Myford &Wolfe, 2004). To be specific, towards Language use showing the highest criteria difficulty, all three raters assigned higher ratings than would have been expected, given how the other raters used the rating scale. These outcomes demonstrated the need to conduct rater training or formal norming sessions before the evaluations to prevent random ratings from being assigned to examinees.

5.2 Limitations and Suggestions for Further Research

The current study encompasses shortcomings. The first has to do with not providing in-depth explanations behind the occurrence of criterion-related biases. Even though the employment of the FACETS program enabled the researcher to pinpoint the biased interactions between particular raters and rating criteria, quantitative research methods alone may not delineate all the cognitive processes involved in the evaluation or decision-making process raters undergo in assigning test scores. Thus, multiple research methodologies combined with the process-oriented approach such as verbal protocol analysis may shed more light on the relation between rater cognition and scoring behavior.

The second limitation is associated with a methodological issue about controlling the conditions in which participants rated 30 essays. During an individual Zoom session with the researcher, each participant was notified of the time constraints for the completion of rating essays, which should not exceed 30 days. However, detailed specifications, for example, the acceptable number of essays to be rated per day or the time of rating, were not informed and were left at the discretion of participants since the researcher judged that placing rather strict limitations on the rating context could cause difficulty in recruiting prospective research participants. However, the lack of control concerning the conditions of evaluating essays could have potentially resulted in irrelevant variables influencing rater scoring. Factors other than the perception of criterion importance such as rater fatigue caused by crammed ratings or the late time of rating and rater inconsistency resulting from a long interim between ratings could possibly have affected the results. Thus, an attempt to place specific conditions as to the rating of the essays might have minimized the unwanted effects of other possible variables.

The third limitation arose from the fact that participants in the study were not trained raters for assessing English writing compositions. Of course, participants, as English teachers at middle or high schools, can have had experience in rating the written products of students. However, to the researcher’s knowledge, there is no rater training provided exclusively for classroom writing evaluations. Additionally, since few numbers of writing performance tests are administered in class, participants’ previous rating experiences can be limited and vary from rater to rater. Thus, a lack of training experience can have affected the results of the study. Given that participants of Eckes (2008, 2012) were trained raters working on the writing section of the Test of German as a Foreign Language, the different results between the present study and Eckes (2012) can possibly be attributed to the differences in the amount of rating experience or training among the participants of two studies. With participants having considerable rating experiences, it could have been possible to obtain different research results.

When it comes to suggestions for further research, it should be noted that the scope of rater cognition can vary tremendously (Bejar, 2012; Han, 2015). Various ways of clustering raters based on cognitive aspects other than perceived criterion importance can be employed, for example, raters' decision-making strategies, metacognitive strategies, and reading strategies. Similarly, as rater bias can arise between raters and other facets of the evaluative process, such as the time of rating, the types of tasks or prompts, and examinees, operational rater types can also be generated depending on these various types of interactions involving raters. Considering the paucity of research aiming to bridge the gap between rater cognition and scoring behavior, research along this line holds much promise. Thus, relating cognition-based rater types to operation-based rater types is expected to provide more insight into investigating rater variability.

REFERENCES

Andrich, D. (1998). Thresholds, steps and rating scale conceptualization. *Rasch Measurement Transactions*, *12*(3), 648–649.

Bacha, N. (2001). Writing evaluation: what can analytic versus holistic essay scoring tell us? *System*, *29*(3), 371-383.

Bachman, L. F. (1990*). Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F., & Savignon, S. J. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL oral interview. *The Modern Language Journal*, *70*(4), 380-390.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests.* Oxford: Oxford University Press.

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.

Barkaoui, K. (2014). Multifaceted Rasch analysis for test evaluation. In A. J. Kunnaan (Ed.), *The companion to language assessment: Evaluation, methodology, and interdisciplinary themes* (Vol. 3, pp. 1301-1322). Chichester: Wiley.

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

Bereiter, C., & Scardamalia, M. (1987*). The psychology of written composition*. Lawrence Erlbaum Associates, Inc.

Breland, H. M., & Robert, J. J. (1984). Perceptions of Writing Skills. *Written Communication*, *1*, 101-109.

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

Caban, H. L. (2003). Rater bias in the speaking assessment of four LI

Japanese ESL. *Second Language Studies*, *12*, 1-15.

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. Richards & R. Schmidt (Eds.), *Language and communication* (pp. 2-27). London: London Group Ltd.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching. *Applied Linguistics*, *1*, 1-47.

Carr, N. T. (2000). A comparison of the effects of analytic and holistic rating scale types in the context of composition tests. *Issues in Applied Linguistics*, *11*(2), 207–241. https://doi.org/10.5070/L4112005035

Carroll, J. B. (1961). Fundamental considerations in testing for English language proficiency of foreign students. In H. B. Allen & R. N. Campbell (Eds.), Teaching English as a second language: A book of readings (2nd ed.) (pp. 313-321). New York: McGraw-Hill.

Cho, D. (1999). A study on ESL writing assessment: intra-rater reliability of ESL compositions, *Melbourne Papers in Language Testing*, *8*(1), 1-24.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Clark, J. L. D. (1972). *Foreign language testing: theory and practice*. Philadelphia: Center for Curriculum Development.

Clark, J. L. D. (1975). Theoretical and technical considerations in oral proficiency testing. In R. L. Jones & B. Spolsky (Eds.), *Testing language proficiency* (pp. 10-28). Arlington, VA: Center for Applied Linguistics.

Clark, L. A., & Watson, D. (2019). Constructing Validity: New Developments in Creating Objective Measuring Instruments. *Psychological Assessment*, *31*, 1412-1427. https://doi.org/10.1037/pas0000626

Connor-Linton, J., & Polio, C. (2014). Comparing perspectives on L2 writing: Multiple analyses of a common corpus. *Journal of Second Language Writing*, *26*, 1–9.

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

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

Davies, A. (1968). Introduction. In A. Davies (Ed.*), Language testing symposium: a psycholinguistic approach* (pp. 1-18). London: Oxford University Press.

Davies, A. (1977). The construction of language tests. In J. P. B. Allen & A. Davies (Eds.), *Testing and experimental methods. The Edinburgh Course in Applied Linguistics* (Vol. 4, pp. 38-104). Oxford: Oxford University Press.

Deane, P., Odendahl, N., Quinlan, T., Fowles, M., Welsh, C., & Bivens-Tatum, J. (2008). Cognitive models of writing: Writing proficiency as a complex integrated skill. *ETS Research Report Series*(2), i–36.

Draba, R. E. (1977). *The identification and interpretation of item bias* (Research Memorandum No. 25). Chicago, IL: The University of Chicago, Department of Education, Education Statistics Laboratory.

Du, Y., Wright, B. D., & Brown, W. L. (1996). *Differential facet functioning detection in direct writing assessment*. Paper presented at the Annual Meeting of the American Educational Research Association, New York.

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

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

Eckes, T. (2009). On common ground? How raters perceive scoring criteria in oral proficiency testing. In A. Brown &K. Hill (Eds.), *Tasks and criteria in performance assessment: Proceedings of the 28th Language Testing Research Colloquium* (pp. 43–73). Frankfurt, Germany: Lang.

Eckes, T. (2011). *Introduction to many-facet Rasch measurement: analyzing and evaluating rater-mediated assessments*. Frankfurt, Germany: Lang.

Eckes, T. (2012). Operational rater types in writing assessment: linking rater cognition to rater behavior. *Language Assessment Quarterly, 9*(3), 270–292.

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: analyzing and evaluating rater-mediated assessments* (2nd ed.). Frankfurt: Peter Lang.

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal & T. M. Haladyna (Eds.*), Large-scale assessment programs for* *all students: Validity, technical adequacy, and implementation* (pp. 261–287). Mahwah, NJ: Erlbaum.

Engelhard, G., Wang, J., & Wind, S. A. (2018). A tale of two models: Psychometric and cognitive perspectives on rater-mediated assessments using accuracy ratings. *Psychology Science, 60*(1), 33-52.

Fahim, M., & Bijani, H. (2011). The effects of rater training on raters' severity and bias in second language writing assessment. *Iranian Journal of Language Testing*, *1*(1), 1-16.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication*, *32*(4), 365–387.

Freedman, S. W. (1979). How Characteristics of Students' Essays Influence Teachers' Evaluation. *Journal of Educational Psychology*, *71*, 328-338.

Ghanbari, B., Barati, H., & Moinzadeh, A. (2012). Rating scales revisited: EFL writing assessment context of Iran under scrutiny. *Language Testing in Asia*, *2*(1), 1-18.

Grabe, W., & Kaplan, R. B. (1996). *Theory and practice of writing: An applied linguistics perspective*. Longman.

Han, Q. (2015). Rater cognition in L2 speaking assessment: a review of the literature. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics*, *16*(1), 1-24.

Heidari, N., Ghanbari, N., & Abbasi, A. (2022). Raters' perceptions of rating scales criteria and its effect on the process and outcome of their rating. *Language Testing in Asia*, *12*(20). https://doi.org/10.1186/s40468-022-00168-3

Hijikata-Someya, Y., Ono, M., & Yamanishi, H. (2015). Evaluation by native and non-native English teacher raters of Japanese students' summaries. *English Language Teaching*, *8*(7), 1–12.

Hill, C. E., O'Grady, K. E., & Price, P. (1988). A method for investigating sources of rater bias. *Journal of Counseling Psychology*, *35*, 346-350.

Huot, B. A. (1990). Reliability, Validity, and Holistic Scoring: What We Know and What We Need to Know. *College Composition and Communication*, *41*(2), 201-213.

Huot, B. A. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M. M. Williamson and B.A. Huot (Eds.), *Validating holistic scoring for writing assessment: theoretical and* *empirical foundations* (pp. 206-236). Cresskill, NJ: Hampton Press.

Hymes, D. (1967). Models of the interaction of language and social setting. *Journal of Social Issues, 23*(2), 8-38.

Hymes, D. (1972). *On Communicative Competence.* In J. B. Pride and J. Holmes (Eds.), *Sociolinguistics: selected readings* (pp. 269-293). Harmondsworth: Penguin.

Ishikawa, S. (2018). Comparison of three kinds of alternative essay-rating methods to the ESL Composition Profile. *International Journal of Computer-Assisted Language Learning and Teaching*, *8*(4), 32-44.

Jacobs, H., Zingraf, S., Wormuth, D., Hartfiel, V., & Hughey, J. (1981). *Testing ESL composition: a practical approach. English composition program*. Rowley: Newbury House.

Jeong, H. (2019). Writing scale effects on raters: An exploratory study. *Language Testing in Asia*, *9*(20), 1–19. https://doi.org/10.1186/s40468-019-0097-4

Johnson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing*, *26*(4), 485-505.

Johnson, S. C. (1967) Hierarchical Clustering Schemes. *Psychometrika*, *32*, 241-254.

Kellogg, R. (2001). Competition for working memory among writing processes. *The American Journal of Psychology*, *114*, 175–191. https://doi.org/10.2307/1423513

Kenyon, D. M. (1992). Introductory remarks at symposium on *Development and use of rating scales in language testing*. Paper presented at the 14th Language Testing Research Colloquium, Vancouver.

Kim, H. (2020). Effects of rating criteria order on the halo effect in L2 writing assessment: a many-facet Rasch measurement analysis. *Language Testing in Asia*, *10*(16). https://doi.org/10.1186/s40468-020-00115-0

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing*, *12*, 26-43. https://doi.org/10.1016/j.asw.2007.04.001

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese L2 writing performance. *Language Testing*, *19*(1), 3–31.

Kuiken, F., & Vedder, I. (2014). Rating written performance: What do raters do and why? *Language Testing*, *31*(3), 329-348. https://doi.org/10.1177/0265532214526174

Lado, R. (1961). *Language testing: the construction and use of foreign language tests*. London: Longman.

Levy, C. M., & Ransdell, S. (1995). Is writing as difficult as it seems?. *Memory & Cognition*, *23*(6), 767-779.

Li, H., & He, L. (2015). A comparison of EFL raters& essay-rating processes across two types of rating scales. *Language Assessment Quarterly*, *12*(2), 178–212. https://doi.org/10.1080/15434303.2015.1011738

Linacre, J. M. (1998). Rating, judges and fairness. *Rasch Measurement Transactions*, *12*, 630-631.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago, IL: MESA Press.

Linacre, J. M. (2014). *A user's guide to FACETS: Rasch-model computer programs*. Chicago: Winsteps.com. Retrieved from http://www.winsteps.com/facets.htm

Lumley, T. (2002). Assessment criteria in a large-scale writing test: what do they really mean to the raters? *Language Testing, 19(3), 246–276*.

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

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

McNamara, T. F. (1995). Modelling performance: Opening pandora's box. *Applied Linguistics*, *16*, 159-179.

McNamara, T. F. (1996*). Measuring second language performance*. NewYork: Longman.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13–103). New York: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13–23. https://doi.org/10.3102/0013189X023002013

Milanovic, M., Saville, N., & Shen, S. (1996). A study of decision-making behaviour of composition markers. In M. Milanovic & N. Saville (Eds*.),* *Studies in Language Testing 3: Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem* (pp. 92-115). Cambridge, UK: Cambridge University Press.

Morrow, K. (1979). Communicative language testing: revolution or evolution? In C. J. Brumfit & K. Johnson (Eds.), *The communicative approach to language teaching* (pp. 143-157). Oxford: Oxford University Press.

Murtagh, F., & Legendre, P. (2014) Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *Journal of Classification*, *31*, 274-295. http://dx.doi.org/10.1007/s00357-014-9161-z

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*(4), 386–422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, *5*(2), 189-227.

North, B. (2003). Scales for rating language performance: Descriptive models, formulation styles, and presentation formats. *TOEFL Monograph, 24.*

Odell, L. (1993*). Theory and practice in the teaching of writing: Rethinking the discipline*. SIU Press.

Ono, M., Yamanishi, H., & Hijikata, L. (2019). Holistic and analytic assessments of the TOEFL iBT® Integrated Writing Task. *JLTA Journal*, *22*, 65–88.

Park, M.-Y. (2012). Exploring the Raters' Bias on an EFL Writing Assessment Using Multi-faceted Rasch Measurement. *Studies in English Education*, *17*(2), 175-202.

Park, M.-Y., & Shim, J.-W. (2014). An investigation into pre-service and in-service teachers' judgment of English writing performance. *Modern English Education*, *15*(2), 71-90.

Paula, J. J., & Huot, B. A. (1993). A model of background influences on holistic raters. In M. M. Williamson & B.A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 237-265). Cresskill, NJ: Hampton.

Plakans, L., & Gebril, F. (2017). Exploring the relationship of organization and connection with scores in integrated writing assessment. *Assessing*

Writing, 31, 98–112.

Raimes, A. (1987). Language Proficiency, Writing Ability, and Composing Strategies: A Study of ESL College Student Writers. *Language Learning, 37*(3), 439-468.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedagogiske Institute.

Ruetten, M. K. (1991). Reading problematical ESL placement essays. *College English*, *1*, 37-47.

Sakyi, A. A. (2000). Validation of holistic scoring for ESL writing assessment: How raters evaluate compositions. In: A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (pp.129-152). Cambridge, UK: Cambridge University Press.

Sasaki, M., & Hirose, K. (1999). Development of an analytic rating scale for Japanese L1 writing. *Language Testing*, *16* (4), 457-478.

Savignon, S. J. (1972). *Communicative competence: an experiment in foreign language teaching*. Philadelphia: The Center for Curriculum Development.

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

Setyowati, L., Sukmawan, S., El-Sulukiyyah, A. A. (2020). Exploring the use of ESL Composition Profile for college writing in the Indonesian context. *International Journal of Language Education, 4*(2), 171-182.

Shin, Y. (2010). A FACETS analysis of rater characteristics and rater bias in measuring L2 writing performance. *English Language & Literature Teaching*, *16*(1), 123-142.

Shohamy, E. (1988). A proposed framework for testing the oral language of second/foreign language learners. *Studies in Second Language Acquisition*, *10*(2), 165-179.

Shohamy, E. (1995). Performance assessment in language testing. *Annual Review of Applied Linguistics, 15*, 188–211.

Skar, G. B., & Jølle, L. J. (2017). Teachers as raters: An investigation of a long-term writing assessment program. *L1-Educational Studies in Language and Literature*, *17*, 1-30. https://doi.org/10.17239/L1ESLL-2017.17.01.06

Spolsky, B. (1985). The limits of authenticity in language testing. *Language Testing*, *2*(1), 31-40.

Stiggnins, R. J. (1987). Design and development of performance assessments. *Education Measurement: Issues and Practice*, *6*(3), 33-42.

Tedick, D. J., & Mathison, M. A. (1995). Holistic scoring in ESL writing assessment: What does an analysis of rhetorical features reveal? In D. Belcher & G. Braine (Eds.), *Academic writing in a second language: Essays on research and pedagogy* (pp. 205-230). Norwood, NJ: Ablex.

Turner, C. E. (2000). Listening to the voices of rating scale developers: Identifying salient features for second language performance assessment. *Canadian Modern Language Review*, *56*, 555-584.

Valdes, G., Haro, P., & Arriarza, M. P. E. (1992). The development of writing abilities in a foreign language: Contributions toward a general theory of L2 writing. *The Modern Language Journal*, *76*, 333-352.

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.

Wang, J., Engelhard, G., Raczynski, K., Song, T., & Wolfe, E. D. (2017). Evaluating rater accuracy and perception for integrated writing assessments using a mixed-methods approach. *Assessing Writing*, *33*, 36–47.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, *10*(3), 305–335.

Wigglesworth, G. (1994). Patterns of rater behaviour in the assessment of an oral interaction test. *Australian Review of Applied Linguistics*, *17*(2),77–103.

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, *6*(2), 145-178. https://doi.org/10.1016/S1075-2935(00)00010-6

Weigle, S. C. (2002). *Assessing Writing*. Cambridge, UK: Cambridge University Press.

Winke, P., & Lim, H. (2015). ESL essay cognitive raters& processes in applying the Jacobs et. al. rating scale: An eye movement study. *Assessing Writing*, *25*, 38–54.

Wolfe, E. W., Kao, C. W., & Ranney, M. (1998). Cognitive Differences in Proficient and Nonproficient Essay Scorers. *Written communication, 15*(4), 465-492.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Zamel, V. (1982). Writing: The process of discovering meaning. *TESOL Quarterly, 16*(2), 195-209.

Zamel, V. (1987). Recent Research on Writing Pedagogy. *TESOL Quarterly, 21*(4), 697-715.

APPENDICES

APPENDIX 1: Questionnaire

영작문 채점 기준인 Content, Organization, Vocabulary, Language use, Mechanics 에 대하여 평소 얼마나 중요하게 생각하고 있었는지 평가해주세요. 해당되는 칸에 클릭하면 됩니다. 단, 특정 평가 상황을 가정하지 않고, 평소 갖고 있던 채점 기준에 대한 중요성을 평가합니다.

Content : 'Content'가 갖춰진 글은 논지(thesis)의 전개가 논리적이고, 주장을 뒷받침하는 세부 내용이 충분함. 또한 논지와 상관없는 내용이 없으며, 그 내용이 흥미로움.
□ 극도로 중요함 (extremely important) □ 매우 중요함 (very important) □ 중요함 (important) □ 덜 중요함 (less important)

Organization : 'Organization'이 갖춰진 글은 논지(thesis)를 담은 중심 문장이 있으며, 도입-전개-결론에 이르는 구성이 명확함. 또한 연결사(transition word)의 사용이 적절하고 단락 간의 연결이 매끄러움.
□ 극도로 중요함 (extremely important) □ 매우 중요함 (very important) □ 중요함 (important) □ 덜 중요함 (less important)

Vocabulary : 수준 높은 어휘를 구사하며, 그 사용에 오류가 없음. 또한 어휘 사용의 범위가 넓고, 문맥에 맞음. academic writing의 모습을 띰.

□ 극도로 중요함 (extremely important)

□ 매우 중요함 (very important)

□ 중요함 (important)

□ 덜 중요함 (less important)

Language use : 좋은 'Language Use ' 능력은 효과적인 복문(complex sentence)을 사용하며, 그 사용에 오류가 없음. 일치, 시제, 관사, 대명사, 전치사의 사용에 문법적 오류가 없어 의미 전달에 방해를 주지 않음. 다양한 종류의 문장을 구사함.

□ 극도로 중요함 (extremely important)

□ 매우 중요함 (very important)

□ 중요함 (important)

□ 덜 중요함 (less important)

Mechanics : 'Mechanics'가 좋은 글은 철자, 구두법(punctuation), 대/소문자 구사에 오류가 없음.

□ 극도로 중요함 (extremely important)

□ 매우 중요함 (very important)

□ 중요함 (important)

□ 덜 중요함 (less important)

APPENDIX 2: Rating Scale

Content	5 (Excellent)	Thorough and logical developments of thesis; sufficient supporting details; no irrelevant information at all; interesting; a substantial number of words for the amount of time given
	4 (Good)	Good and logical development of thesis; adequate supporting details, but not sufficient; almost no irrelevant information; somewhat interesting; an adequate number of words for the amount of time given
	3 (Fair)	Limited development of thesis; somewhat insufficient supporting details; several irrelevant information; somewhat uninteresting; an adequate number of words for the amount of time given
	2 (Poor)	Very limited development of thesis; a lack of supporting details; a substantial amount of irrelevant information; uninteresting; a limited number of words for the amount of time given
	1 (Very poor)	No development of thesis; no supporting details; almost all information irrelevant; very few words for the amount of time given; not enough to evaluate

Organization	5 (Excellent)	Excellent overall organization; clear thesis statement; strong introduction and conclusion; excellent use of transition words; excellent connections between paragraphs; unity within every paragraph
	4 (Good)	Good overall organization; clear thesis statement; good introduction and conclusion; good use of transition words; good connections between paragraphs; unity within most paragraphs
	3 (Fair)	Some general coherent organization; adequate thesis statement or main idea; rather weak introduction and conclusion; occasional use of transitions words; some disjointed connections between paragraphs; some paragraphs may lack unity
	2 (Poor)	Very loose organization; weak thesis statement or main idea; weak introduction and conclusion; a lack of transitions words; many disjointed connections between paragraphs; most of the paragraphs may lack unity
	1 (Very poor)	No coherent organization; no thesis statement or main idea; no introduction and conclusion; no use of transition words; disjointed connections between paragraphs; all paragraphs lack unity; not enough to evaluate

Vocabulary	5 (Excellent)	Very sophisticated vocabulary; excellent choice of words with no errors; excellent range of vocabulary; effective word/ idiom choice and usage; academic register
	4 (Good)	Somewhat sophisticated vocabulary; attempts, even if not completely successful, at sophisticated vocabulary; good choice of words with some errors that do not obscure meaning; an adequate range of vocabulary but some repetition; approaching academic register
	3 (Fair)	Unsophisticated vocabulary; limited word choice with some errors that do not obscure meaning; repetitive choice of words; having a resemblance to academic register
	2 (Poor)	Simple vocabulary; very limited word choice with frequent errors that often obscure meaning; dominance of repetitive choice of words; no resemblance to academic register
	1 (Very poor)	Very simple vocabulary; severe errors in word choice that obscure meaning and thus do not communicate at all; no variety in word choice; no resemblance to academic register; not enough to evaluate

Language use	5 (Excellent)	Effective and frequent complex constructions; no major errors in word order or complex structures; few errors in agreement, tense, number, word order, articles, pronouns, or prepositions that do not interfere with comprehension; excellent sentence variety
	4 (Good)	Effective and frequent complex constructions; a few minor errors in word order or complex structures; a few errors in agreement, tense, number, word order, articles, pronouns, or prepositions that do not interfere with comprehension; good sentence variety
	3 (Fair)	Effective but dominated by simple constructions than complex ones; several minor errors in word order or complex structures; several errors in agreement, tense, number, word order, articles, pronouns, or prepositions that do not obscure meaning; weak sentence variety
	2 (Poor)	Few complex sentences, major problems in word order and sentence constructions; frequent errors in negation, agreement, tense, number, word order, articles, pronouns, or prepositions and/or fragments, run-ons, deletions; meaning confused or obscured; little sentence variety

	1 (Very poor)	virtually no mastery of sentence construction rules; dominated by errors; does not communicate; not enough to evaluate

Mechanics	5 (Excellent)	Demonstration of mastery of conventions; few errors in spelling, punctuation, and capitalization
	4 (Good)	A few errors in spelling, punctuation, and capitalization
	3 (Fair)	Occasional errors in spelling, punctuation, and capitalization that do not obscure meaning
	2 (Poor)	Frequent errors in spelling, punctuation, and capitalization; meaning confused and obscured
	1 (Very poor)	No mastery of convention; dominated by errors in spelling, punctuation, and capitalization; not enough to evaluate

국 문 초 록

의사소통 중심 교수법(Communicative Language Teaching)의 도입으로 학습자의 실제 영어 사용 능력을 평가하는 수행평가에 대한 중요성이 강조되어왔다. 수행평가가 기존의 선다형 평가와 구별되는 점은 학습자가 구성한 답안을 채점하기위해 사용하는 채점기준표(rating rubric)의 존재함에 있다. 다시 말해서, 어떻게 채점자가 채점기준표를 해석하고 적용하는지가 학습자가 받게 될 점수에 지대한 영향을 끼친다. 따라서 채점자와 채점기준(rating criteria) 사이의 상호작용의 속성을 연구한 많은 선행 연구가 있었고, 이들 연구는 채점자의 비일관적인 채점 기준의 적용을 줄임으로써 보다 타당도 높은 평가가 이루어지게 하는데 도움을 주기 위한 목적이 있었다. 채점자 오류(rater effects)를 연구한 기존의 연구는 채점자가 어떠한 채점 기준에 대하여 채점의 엄격성 혹은 관대함을 나타내는지를 기술적으로 증명하였다. 그러나 채점자 오류가 일어나는 근본적인 원인을 인지적인 측면에서 규명하려는 시도가 부족하였다. 즉, 채점자의 채점 기준에 대한 인식이 채점의 엄격성 혹은 관대함에 어떠한 영향을 미치는지를 알아보는 것이 필요하다.

따라서 본 연구는 영어 쓰기 평가 채점 기준에 대한 중요성 인식이 실제 채점에 어떠한 영향을 미치는지를 탐색하는 것이다. 이를 통해 채점 기준에 대한 채점자의 인식을 고찰하고, 채점 기준에 대한 편향 없는 인식을 갖는데 도움을 주기 위함이다.

본 연구를 위해 한국의 중학교 혹은 고등학교에 근무하는 한국인 영어 교사 30명이 참여하였다. 이들은 다섯 가지 채점 기준(Content, Organization, Vocabulary, Language use, Mechanics)에 부여하는 중요성의 정도에 관한 인식을 묻는 설문조사에 참여하고, 30개의 영어 작문을

채점하였다. 다국면라쉬모형과 계층적 군집 분석을 사용하여 채점 기준에 대한 중요성 인식과 채점 기준에 대한 오류를 바탕으로 채점자 인지 유형(Cognitive Rater Types: CRTs)과 채점자 행동 유형(Operational Rater Types: ORTs)을 구성하였다. 이후 두 채점자 유형 사이에 어떠한 관련성이 있는지를 분석하였다.

연구 결과 채점 기준에 대한 중요성 인식에 따라 5가지 채점자 인지 유형이 형성되었고, 채점 기준에 대한 채점자 오류에 따라 6가지 채점자 행동 유형이 구성되었다. 채점자 행동 유형은 상이한 채점 기준에 관한 중요성 인식을 가진 채점자들로 구성이 되었기에 채점자 인지 유형과 채점자 행동 유형 사이에 직접적인 비교가 가능하지 않았다. 따라서 같은 채점자 행동 유형에 속하는 채점자들의 채점 기준에 대한 중요도의 평균 점수와 해당 채점 기준에 보인 편향 수치를 비교한 결과 채점 기준에 대한 중요성 인식이 채점 행동에 미치는 영향은 채점 기준 별로 차이가 있음을 발견하였다. Content와 Mechanics에서 채점의 엄격성과 관대함이 모두 발견되었는데, 채점자 오류가 채점 기준에 대한 중요성 인식과 결합하는 패턴은 이 두 가지 채점 기준에서 차이가 있었다. Content에서 채점의 엄격성은 평균보다 높은 채점 기준 중요도와 결합되어 나타났고, 채점의 관대함은 평균보다 낮은 채점 기준 중요도와 결합되어 관찰되었다. 그러나 이 결합의 패턴은 Mechanics에서 반대로 나타났는데, 채점의 엄격성은 평균보다 낮은 채점 기준 중요도와 결합되어 나타났고, 채점의 관대함은 평균보다 높은 채점 기준 중요도와 결합되어 관찰되었다. 이렇듯, Content와 Mechanics에서 채점 기준 중요도와 채점자 오류의 상이한 결합 패턴은 개별 채점자를 대상으로 한 데이터에서도 관찰되었다.

본 연구는 쓰기 평가를 위해 훈련된 참여자를 대상으로 하지 않은 한계가 있으나 쓰기 평가 채점 기준에 관한 중요성 인식과 채점자 행동의 관계를 파악함으로써 채점자 오류 연구에 관한 지평을 넓히는데 도

움이 될 것으로 기대한다.

주요어: 채점자 오류, 채점 기준에 대한 중요성, 영어 쓰기 평가, 채점자 유형, 채점자 인식, 수행 평가

학 번: 2021-28726