Linguistic Research 36(Special Edition), 25-57
DOI: 10.17250/khisli.36..201909.002

Analysis of rater effect in the evaluation of second language grammatical knowledge in the context of writing: Application of a generalized linear model*

Hyun Jung Kim[a] · Junkyu Lee[a] · Hyun-Jo You***[b]**
(Hankuk University of Foreign Studies[a] · Seoul National University[b])

Kim, Hyun Jung, Junkyu Lee, and Hyun-Jo You. 2019. Analysis of rater effect in the evaluation of second language grammatical knowledge in the context of writing: Application of a generalized linear model. *Linguistic Research* 36(Special Edition), 25-57. Despite ample testing research involving the impact of rater background characteristics (e.g. raters' native language) on assessment, relatively little has been known about how raters' linguistic knowledge influences their scoring decision-making process. By associating second language acquisition research with regard to two types of linguistic knowledge (i.e. implicit and explicit knowledge), this study aims to explore the influence of raters' linguistic knowledge and other rater factors (including teaching experience) on second language writing assessment, particularly in relation to the grammatical aspect of performance. Forty-two raters' grammatical knowledge was measured with an untimed paper-and-pencil grammar test (explicit knowledge) and a timed computer-based experiment (implicit knowledge). The raters also responded to a background questionnaire. For scoring, the raters evaluated five examinees' responses on two writing tasks (a personal essay and an argumentative essay), based on a six-point analytic scoring rubric. An analysis of a cumulative link mixed model revealed that the raters' scores could be significantly related to their grammatical knowledge and teaching experience. In addition, different effects emerged across the examinees' ability levels. These findings contribute to a better understanding of the predictors with respect to rating performance and provide practical implications for rater selection and training in second language performance assessment. **(Hankuk University of Foreign Studies · Seoul National University)**

Keywords rater effects, writing test, grammatical knowledge, teaching experience, cumulative link mixed model

* We would like to thank the two anonymous reviewers for their valuable comments on the earlier version of this article. This work was supported by Hankuk University of Foreign Studies Research Fund.

** First author

*** Corresponding author

1. Introduction

Performance assessment, a topic of great interest for test developers and users alike, has been widely used to measure test-takers' second language (L2) productive skills. As evident in the term itself, performance assessment requires test-takers to actually perform or verbally produce language (e.g. speaking and writing); subsequently, it necessarily involves raters and rating scales for evaluating language performance (McNamara 1996; Bachman 2002; Schaefer 2008; Johnson and Lim 2009). Due to its direct nature, it is considered the most valid way to evaluate test-takers' language ability (Schoonen, Vergeer, and Eiting 1997). At the same time, however, the involvement of raters (another defining characteristic of performance assessment) has constantly raised the issue of test score validity and reliability, due to rater effects (Dunbar, Koretz, and Hoover 1991; Bachman 2004; Eckes 2005).

Raters have been recognized as one of the key factors affecting test-takers' test performance and their test scores. That is, raters' interpretations of performance features and use of rating criteria during the scoring process determine test scores, from which test-takers' language ability is inferred (McNamara 1995). McNamara argued that performance assessment "introduces a new type of interaction, that between the rater and the scale; this interaction mediates the scoring of the performance" (1996: 121). Therefore, test-takers are not the only figure responsible for their test result (McNamara 1997). Indeed, the problem lies in the fact that there may exist variations in test scores due to rater variability, which can severely damage the precision of ratings (Linacre 1989; Bachman 1990; Johnson, Penny, and Gordon 2000). Due to such rater variability in the rating process, previous studies have also reported that the same or similar ratings do not necessarily guarantee a similar quality of performance (e.g. Douglas 1994; Meiron and Schick 2000). Therefore, rater effects have been studied in L2 performance assessment because variations *among* and *within* raters contribute to measurement error, and ultimately threaten the fairness of assessments (Popham 1990; AERA, APA, and NCME 1999; Reed and Cohen 2001; Bachman 2004).

1.1 Rater effects in L2 writing assessments

Validation studies of L2 writing assessments have actively investigated rater effects, both statistically and qualitatively. For statistical analyses of rater behavior or rating patterns, many-facet Rasch measurement (MFRM) analysis has often been employed, which makes it possible to measure an examinee's language ability, controlling for the effects of other variables involved in a testing situation (e.g. test item difficulty). In addition, MFRM analysis identifies the sources of examinees' score variances, such as an individual's ability, item difficulty, and rater severity (Linacre 1989; Schaefer 2008). Through the use of MFRM analysis, empirical studies have shown evidence or sources of rater variability, mostly in one of the following three areas: (1) rater severity/leniency, defined as the overall harshness in rating and interpretations of rating scale thresholds (e.g. Engelhard 1994; Lumley and McNamara 1995); (2) consistency/inconsistency, which indicates whether raters are internally consistent throughout their rating or over time (e.g. Cho 1999; Wiseman 2008; Lim 2011); and (3) interaction with other facets (e.g. bias toward certain examinee groups/levels, gender, and rating scales) involved in the rating situation (e.g. Wigglesworth 1993, 1994; Du, Wright, and Brown 1996; Kondo-Brown 2002; Lumley 2002, 2005; Schaefer 2008). In addition to the statistical analysis of ratings, raters' scoring processes have also been examined qualitatively in order to identify differences in rater behavior, oftentimes using think-aloud protocols or verbal reports. These studies compared raters' reading strategies and styles, as well as their overall decision-making processes in L2 writing assessments (e.g. Vaughan 1991; Milanovic, Saville, and Shen 1996; Wolfe, Kao, and Ranney 1998; Sakyi 2000; Smith 2000; Cumming, Kantor, and Powers 2001, 2002). For example, Milanovic et al. (1996) devised a model of raters' decision-making process in composition rating. In this model, seven steps of reading behavior were identified: (1) pre-marking (internalizing the marking scheme and interpreting the tasks); (2) scanning (length, format, handwriting, and organization); (3) reading quickly (establishing the overall level of comprehension); (4) rating (e.g. assessing the relevance, topic development, coherence, organization, syntactic complexity, appropriateness of the lexis, and mechanics); (5) modifying; (6)

reassessing/revising; and (7) deciding the final mark. Similarly, Cumming et al. (2001) also identified three stages of raters' decision-making process: scanning for surface-level identification (e.g. length and format of the compositions), engaging in interpretation strategies (e.g. classifying error types and interpreting rhetorical strategies), and making a scoring decision.

While examining rater behavior through psychometric and qualitative analyses, oftentimes different rater groups are defined a priori or a posteriori, according to rater characteristics; their rating behavior is then compared across groups or is interpreted in relation to the characteristics that each group possesses. Rater characteristics identified as factors attributing to rater variability include raters' teaching experience (e.g. Vaughan 1991; Hamp-Lyons 1996), language or professional background (e.g. Shi 2001; Hamp-Lyons and Davies 2008; Johnson and Lim 2009), and expertise in rating or prior rater training experience (e.g. Cumming 1990; Shohamy, Gordon, and Kraemer 1992; Schoonen et al. 1997; Weigle 1998; Wolfe et al. 1998; Barrett 2001). These studies contribute to explaining why and how raters differ in scoring writing performances. However, the same or similar level/degree of a factor that raters bring into the rating context (e.g. experience in teaching L2 writing) does not necessarily guarantee the same or similar rating pattern across different rating contexts. For example, Cumming (1990) found that expert and novice raters showed a significant difference in scoring the content and rhetorical organizational aspects of 12 adult ESL learners' compositions. On the other hand, in Schoonen et al. (1997), expert and novice raters did not show a significant difference in reliability when scoring the content component of sixth-grade children's compositions across different task types. Such varying results might be attributed to a number of reasons. First, rater characteristics have been defined inconsistently across different studies; consequently, 'expert raters' might imply different meanings that are not comparable across studies. Moreover, previous research has tended to focus on the surface features of rater characteristics (e.g. number of years teaching and language background), and has less paid attention to more direct construct-relevant characteristics (such as raters' knowledge of writing ability) in explaining their rating behavior and variability. Another possible reason can be found in the analytical methods used to examine rater behavior. MFRM has often been used for the analysis of rating patterns;

however, the unidimensionality assumption of MFRM analysis only allows a descriptive explanation in relation to rater characteristics. In other words, MFRM has difficulty in capturing multiple characteristics within an individual rater. Therefore, the direct relationships between individual rater characteristics and actual ratings have not been statistically analyzed.

In order to figure out the factors (both directly and less directly construct-relevant) that affect raters' scoring behavior, raters' background characteristics and rating ability need to be understood comprehensively within a systematic framework. For this purpose, Kim (2011) suggested a model of rating ability, wherein the *rater* factor that had previously been limited to part of the model of performance assessment (e.g. Kenyon 1992; McNamara 1995) was further explored in relation to rater background characteristics and other related factors. In the model, four variables (rating experience, TESOL experience, rater training, and relevant coursework) and the interactions among these four variables were hypothesized to build *raters' knowledge of performance rating* (knowledge about the test construct and understanding of the rating scales) and *strategic competence* (the ability to apply raters' knowledge to a given rating context); ultimately, the combination of *knowledge of performance rating* and *strategic competence* determines a rater's rating ability. It is the raters' rating ability that leads their rating performance, in addition to interactions with their personal attributes (e.g. gender and native language), rating contexts, and rating task characteristics. Earlier, other researchers also argued for similar factors and a comprehensive consideration of such factors as contributing elements to the development of rater expertise in L2 writing assessments (e.g. Kim 2018; Lim 2011).

Among the many factors involved in scoring, the current study aimed to examine the effects of raters' linguistic knowledge as a first attempt to examine separate rater characteristics using a more advanced statistical analysis. The focus has been placed on raters' linguistic knowledge because its impact on rating or the decision-making process has rarely been investigated, even though it is the most construct-relevant factor for evaluating test-takers' language ability. Moreover, raters' linguistic knowledge (instead of their own L2 or writing ability) has been chosen as a part of raters' knowledge in order to make a direct, specific connection between raters' knowledge and their ratings on the grammar component. Since none of the different aspects of raters' knowledge has been singled out for an analysis of their potential impact on scoring, many different aspects other than grammar could also be considered; however, considering the target rater participants (i.e. Korean English language teachers), who were expected to have high grammatical knowledge due to their prior language learning experience, centered on grammar in their school years, the grammatical aspect became the focus of the study.

1.2 Measurement of linguistic knowledge

Despite extensive arguments regarding the exact nature of linguistic knowledge (e.g. Chomsky 1965; Rumelhart and McClelland 1986; Gregg 2003), it has extensively been characterized in a dichotomous fashion (explicit and implicit) within the domain of second language acquisition research (Bialystok 1982; Green and Hecht 1992; Han and Ellis 1998; Dienes and Perner 1999; Hu 2002; Ellis 2005, 2006; Ellis et al. 2009; Lee 2011; Jang and Lee 2015). Ellis (2005) has clearly distinguished the two types of knowledge with respect to awareness, types of knowledge, systematicity, accessibility, use of L2 knowledge, and self-report learnability. For example, implicit knowledge involves intuitive/non-verbalizable awareness of linguistic forms while explicit knowledge entails conscious/verbalizable awareness of linguistic forms (see Ellis 2005 for more details).

Of particular relevance regarding the two types of knowledge to the current study is the availability of various measurements that tap into either explicit or implicit knowledge. Ellis (2005) argued that the two types of knowledge can be tested[1], depending on the degree of awareness, time availability, focus of attention (i.e. focus on meaning vs. focus on linguistic form), consistency of test-takers' responses, use of metalinguistic knowledge, and learnability (i.e. early learning vs. late learning). A timed grammaticality judgment test (a type of implicit knowledge measurement), for instance, is performed in a situation where test-takers evaluate the grammaticality of a series of sentences in a limited

[1] The construct of grammatical knowledge proposed by Ellis (2005) is the 'sentence-level' but not the 'discourse-level' grammar.

amount of time. Under time pressure, the test-takers would depend on their feelings/intuition (not rules in their minds) and would not use their verbalizable metalinguistic knowledge. With no time limitation, in contrast, a metalinguistic knowledge test (a type of explicit knowledge measurement) typically asks test-takers to evaluate or explain the grammaticality of a series of sentences. No time pressure means that test-takers are able to examine the series of sentences repeatedly, in which they are more likely to (1) utilize their rules or verbalizable explicit knowledge in their minds; and (2) pay attention to both the meanings and linguistic forms of the given sentences.

2. The current study

This study investigated whether raters' grammatical knowledge made a difference in their scoring behavior with respect to L2 writing assessments. It further examined whether other factors have an effect on raters' scoring of writing assessments. The scope of the investigation was limited to ratings of the grammatical aspect of writing performance in order to directly relate raters' knowledge to the domain of the rating. The following two research questions were addressed:

1. Does raters' grammatical knowledge have an effect on the scoring of the grammatical aspect of writing performance?
2. Are there any other factors that have an effect on raters' scoring of the grammatical aspect of writing performance?

3. Method

3.1 Design

The current study was designed to examine whether or not raters who had varying degrees of grammatical knowledge (in addition to other background variables identified in the literature) showed differences in assessing writing performance, specifically in regard to the assessment of the grammatical control component. Thus, the main focus of the research was on rater behavior; consequently, a large group of raters were intentionally involved for scoring while the minimum number of examinees representing different levels of writing ability was selected for the raters' scoring. For the writing samples used in the scoring process, a group of high school students voluntarily completed two writing tasks. After screening the students' responses for the two tasks, the researchers purposefully selected five students' responses that represented high, intermediate, and low writing ability levels across the two tasks. This was done to provide the raters with chances to score different performance levels. A fully crossed design allowed each of the raters to evaluate 10 writing samples involving the five students' responses (two high-, two intermediate-, and one low-level examinees' responses) on the two tasks.

3.2 Participants

The current study included two groups of participants: students who completed two writing tasks and raters who scored their written responses. A total of 60 second-year high school students from Seoul, Korea voluntarily participated in the research. All of them were female students who were enrolled in the humanities track at the same high school. They represented different levels of L2 ability, including writing ability.

The raters included 42 graduate students enrolled in an M.A. TESOL program at a Korean university, who had also been teaching English as a Foreign Language (EFL) at different levels (e.g. children, adolescents, and adults) and in different contexts (e.g. public vs. private educational sectors). Since the raters included M.A. students who were willing to participate voluntarily in the current study, they represented different rater characteristics. In order to understand the raters' characteristics and backgrounds, a background questionnaire was developed and distributed to the raters. Questions about demographic information (gender, native/second languages, and overseas experience), experience in teaching EFL, experience in rating English speaking/writing assessments and related rater training, and educational background were asked.

The majority of the raters were female (N = 35, 83%) while only a small number of male raters (N = 7, 17%) participated in the scoring. The raters' experience in living/studying in English-speaking countries varied to a large extent. While 13 raters (31%) had no experience, and 23 raters (55%) had one- to three-years of experience living in an English-speaking country, four raters (9%) had grown up in the U.S., spending six to nine years of their childhood there. There were even two raters (5%) who had moved to Canada and England, respectively, with their families and who had lived there for more than 10 years. As a result, different levels of exposure to an English-speaking environment might have affected the raters' English ability to varying degrees.

The raters also presented a wide range of experience in teaching English. Approximately 55 percent of the raters had taught English for less than three years (none and $0 < x \leq 3$ in Table 1) while approximately nine percent of the raters had taught EFL for more than nine years ($9 < x \leq 12$ and $12 < x \leq 15$ in Table 1). The raters' varying teaching experience is summarized in the following table:

Table 1. Raters' teaching experience

Number of years teaching	Number of raters			Percentage (%)
None	7	(F: 7	M: 0)	17
$0 < x \leq 3$	16	(F: 12	M: 4)	38
$3 < x \leq 6$	8	(F: 7	M: 1)	19
$6 < x \leq 9$	7	(F: 6	M: 1)	17
$9 < x \leq 12$	3	(F: 2	M: 1)	7
$12 < x \leq 15$	1	(F: 1	M: 0)	2
Total	42	(F: 35	M: 7)	100

* F: female; M: male

** Discretizing teaching experience was merely done for the descriptions. In the actual analysis, teaching experience was used as a continuous variable.

Along with their experience in teaching English, the raters reported that they had evaluated and scored students' speaking and writing performance during class and/or for achievement tests. However, they had not been systematically trained for such classroom assessments. Only a few high school teachers had been trained for a large-scale performance test; however, the amount of time was limited, and the training was provided during a single occasion as part of an in-service teacher education program. Therefore, the raters' rating experience was relatively limited to classroom assessment.

3.3 Instruments

Writing test. The writing test included two tasks. The first task was a short 60- to 80-word writing about everyday life, and the second task involved writing an 80- to 120-word argumentative essay. Both tasks were sample items provided by the Korea Institute for Curriculum and Evaluation (KICE) for high school students to help secondary school students improve their communicative language ability in a more balanced manner (Korea Institute for Curriculum and Evaluation 2011).

The first task asked examinees to write the most memorable place they had visited, addressing the required conditions of a prompt (name of the place, time of the visit, and reason for choosing it). The second task required examinees to choose one of the positions for a college education (agree vs. disagree) and to support their position. They were provided with two reasons for each of the two positions in the prompt (e.g. new experience and knowledge for the advantages of a college education, and too much money and too much time for the disadvantages of a college education). In addition to these two reasons, the task required examinees to come up with an additional third reason and to write an essay, including an introduction and conclusion. For these two tasks, 15 and 20 minutes were given, respectively.

Scoring rubric. The raters evaluated the five examinees' written responses using an analytic scoring rubric. The four analytic criteria (task completion, content, organization, and language use) and accompanying descriptors were adapted from the scoring guideline provided by KICE for the high school English writing assessment (Korea Institute for Curriculum and Evaluation 2011). The first component of the rubric, *task completion,* focused on the extent to which an examinee carried out the requested task following the given instructions and provided the relevant information. *Content* was evaluated in terms of elaborating the supporting details. *Organization* considered the logical development of information and the connection of sentences. The last component, *language use,* evaluated the use of correct and suitable grammar and vocabulary for the given situation. Since the focus of the study was on the raters' evaluation of the grammatical aspect of the examinee responses, the *language use* component was further divided into its two separate components of *grammar* and *vocabulary use.* As a result, the raters were asked to assign five analytic ratings instead of four to each examinee response. A six-point scale was used for each analytic component, ranging from 0 for no control or too little evidence for evaluation, to 5 for full, adequate control.

Grammar test. Two types of grammatical knowledge tests in this study were developed based on Ellis (2005). Explicit grammatical knowledge was measured with a paper-and-pencil metalinguistic knowledge test (MKT) while implicit grammatical knowledge was estimated via a computerized timed grammaticality judgment test (GJT). These two exams were designed to test 17 grammatical structures, including the past regular tense and relative clauses, both of which are traditionally known to be difficult for L2 learners to acquire (see Ellis 2005 for more details). Each grammar test was scored by giving 1 point for each correct response and 0 points for each incorrect response.

More specifically, in this study, MKTs, intended to measure explicit grammatical knowledge, consisted of two parts, as in Ellis (2005). In part 1, called MKT1 in this study, the raters were presented with a series of 17 ungrammatical structures, such as *Yesterday the gentleman go to the store,* with four choices of grammatical explanations for each item (e.g. (a) article error, (b) past tense error, (3) relative clause error, and (4) adverbial misplacement). The raters were instructed to find the best rule explanation choice out of the four possibilities. Part 2, called MKT2, was made up of two sections. In the first section of MKT2, the raters were asked to identify 21 specific grammatical items, such as gerunds, from a short passage. In the second section of MKT2, the raters were instructed to find the grammatical parts in a set of sentences.

Implicit grammatical knowledge was measured with a timed GJT, which was created by using E-prime 2.0, a computer software for psycholinguistic experiments. The timed GJT contained 34 grammatical and 34 ungrammatical sentences, each of which was presented as a whole sentence. The duration of each sentence was determined by five native speakers of English, ranging from

1800 milliseconds (ms) to 6500 ms. The raters were asked to judge the grammaticality of each sentence by pressing either a Yes (j key) or No (f key) button as quickly and as accurately as possible. The reliability of the three tests was fairly high when estimated with the KR-20 coefficients (MKT1 = 0.92, MKT2 = 0.91, Time GJT = 0.82).

3.4 Data collection

In order to obtain writing samples for scoring, 60 high school students' written responses to the two tasks were collected. They were given 15 minutes for the first task (writing about the most memorable travel location) and another 20 minutes for the second task (opinion writing about a college education). After reviewing the responses across the two tasks, the five students' responses (two high-, two intermediate-, and one low-level responses) were selected[2]. Raters were expected not to show differences in scoring for very low-level responses because there is usually insufficient evidence to evaluate in such responses. Therefore, only one student was selected for a low level, unlike the other two levels. For instance, low-level responses included a small amount of writing (e.g. less than 30 words when 60-80 and 80-120 words were required). More specifically, in terms of the grammatical aspect, low-level responses failed to include even a single correct sentence.

High school students were recruited to obtain writing samples, although not all raters in the current study had been teaching adolescents. Some of the raters had been working with younger learners while others had been teaching adult EFL learners. Despite this wide range of teaching contexts, high school students were recruited because they were able to provide a fair amount of writing on a topic. While the subject of English is first introduced and taught at the elementary level in Korea, the national curriculum does not emphasize writing at lower grade levels (Ministry of Education, Science and Technology 2011). Therefore, it appeared inappropriate to ask young learners to complete extended

[2] The five students' responses were initially selected from screening the responses. Later, the five students' writing ability was estimated based on the raters' ratings across the five components, using MFRM analysis. The results indicated that the five students were separated into the three writing ability groups as intended, which confirmed the initial screening.

writing tasks. Similarly, most teachers working only with children and adolescents might not have had opportunities to evaluate adult learners′ longer compositions using a complicated rubric (e.g. iBT TOEFL writing tasks). Due to these reasons, high school students′ writing samples were collected for the raters′ scoring.

Once the five students′ written responses to the two tasks were selected, the raters participated in a norming session before the actual rating. During the norming session, they were introduced to the two tasks and the analytic scoring criteria. In addition, they practiced evaluating sample responses, considering each analytic component of the rubric. They discussed as a group which response features they focused on from the sample responses while rating them, and which descriptors they referred to in the rubric to make their scoring decisions. For accurate scoring of the grammar component, which was the focus of the current study, the raters were asked to consciously distinguish among the analytic criteria during the norming session, and they discussed how to differentiate the criteria while scoring the sample responses. Then the raters individually evaluated the 10 responses using the given rubric. A set of five examinees′ responses to the two tasks was distributed, and each rater was given the freedom to decide whether to score examinee-by-examinee or task-by-task. In order to provide the raters with a natural rating environment and to prevent them from paying extra attention to the grammar component, they were asked to assign all five ratings for each response, as instructed in the rubric. However, the focus was limited only to their grammar ratings for later analysis.

In addition to ratings, the raters were asked to fill out a background questionnaire (see the Raters section for more details). As noted earlier, explicit grammatical knowledge (i.e. MKT 1 and MKT 2) was evaluated with a paper-and-pencil format while implicit grammatical knowledge (i.e. a timed GJT) was measured in the form of a computerized test. The explicit knowledge tests were administered with no time limits in a classroom. The raters, however, were not allowed to use any resources, such as grammar references or dictionaries. The two tests lasted approximately 20 minutes to complete. In contrast, the implicit knowledge test was individually performed with time limits in a laboratory. The raters were informed of the time-outs for each sentence. Before the main test of 34 sentences, the raters had 10 practice trials to become accustomed to the process of the main test. The implicit knowledge test took approximately 10 minutes to complete.

3.5 Data analysis

We used a Generalized Linear Mixed Model (GLMM) in order to figure out the factors affecting raters' scoring behavior, raters' background characteristics, and rating ability. MFRM has often been used for the analysis of rating patterns; however, due to the limitation regarding the nature of MFRM analysis (i.e. the unidimensionality assumption), rater behavior has only been descriptively explained in relation to rater characteristics. Since multiple characteristics within an individual rater cannot be separated or tested in relation to his/her ratings using MFRM, direct relationships between individual rater characteristics and actual ratings have not been statistically analyzed.

We used a context-specific implementation of a larger statistical approach toward treating IRT models as GLMM/Non-linear MM, as suggested by Rijmen, Tuerlinckx, De Boeck, and Kuppens (2003), De Boeck and Wilson (2004), and Doran, Bates, Bliese, and Dowling (2007). We introduced GLMM, more specifically a Cumulative Link Mixed Model (CLMM). A 'linear' model allows the analysis of a rater's individual characteristics as explanatory variables, thereby avoiding the unidimensionality problem. The 'generalized' linear model is a generalization of the linear model to allow for response variables with a non-normal error distribution, such as binary, nominal, or ordinal responses. A 'mixed' model with random effects allows us to deal with the dependency within the raters and examinees. There are several GLMMs, according to the type of response variable. As the ratings were coded into six-level ordinal variables (on a 6-point scale), an ordinal regression model–CLMM–was applied in order to examine the effects of raters' grammatical knowledge and background characteristics on their ratings. The logit link function was used, which is equivalent to the proportional odds model. The models were fitted with the CLMM function of the ordinal package (Christensen 2013) in R.

4. Results

4.1 Baseline characteristics of raters

The 42 raters varied in terms of their backgrounds and grammatical knowledge. A brief summary of the rater variables can be seen in Table 2. As stated earlier with regard to rater characteristics, the raters showed a great deal of variety in their overseas and teaching experience. The average number of years living/studying in English-speaking countries was 2.60, with a range from 0 to 30 years. The average number of years teaching English was 4.06, with a 14-year range. In regard to the raters' explicit grammatical knowledge, the mean score of MKT1 was 14.90 (range from 10 to 17), and the standard deviation was 2.00. The mean of MKT2 was 16.20 (range from 8 to 21), and the standard deviation was 3.86. For the raters' implicit grammatical knowledge, the mean score of GJT was 51.50 (range from 36 to 66), and the standard deviation was 7.38. The mean of the GJT response time (RT) was 3,232 ms (range from 1,626 to 4,041), and the standard deviation was 514 ms. For the log-transformed RT, the mean was 8.07 (range from 7.39 to 8.30), and the standard deviation was 0.181.

Table 2. Descriptive statistics for rater variables (N=42)

	Mean	SD	Median	Minimum	Maximum
Background					
Overseas experience (years)	2.60	5.14	1.00	0	30
Teaching experience (years)	4.06	3.61	3.00	0	14
Explicit knowledge					
MKT1 score	14.90	2.00	15.00	10	17
MKT2 score	16.20	3.86	17.50	8	21
Implicit knowledge					
GJT score	51.50	7.38	51.00	36	66
GJT RT (ms)	3,232	514	3,354	1,626	4,041

* MKT: Metalinguistic Knowledge Test, GJT: Grammatical Judgment Test, GJT RT: GJT Response Time

These rater variables did not show strong correlations. Spearman's correlation coefficients are summarized in Table 3. Note that as a measure of metalinguistic knowledge, Ellis (2005) used the collapsed scores of MKT1 and MKT2. This study, however, ended up using the separate scores of MKT1 and MKT2, as the two scores were not strongly correlated. Also, the grammatical (implicit and explicit) knowledge scores were obviously uncorrelated with the background experience variables.

Table 3. Spearman' s correlation coefficients among rater variables

	Teaching experience	MKT1 score	MKT2 score	GJT score	GJT RT
Overseas experience	0.144	-0.097	-0.174	0.143	-0.177
Teaching experience		0.087	0.005	-0.049	-0.081
MKT1 score			0.558	0.550	-0.133
MKT2 score				0.498	-0.063
GJT score					-0.379

* MKT: Metalinguistic Knowledge Test, GJT: Grammatical Judgment Test, GJT RT: GJT Response Time

In order to examine the extent of agreement among the raters on the ratings of grammar, Kendall's coefficient of concordance was calculated for 39 raters who scored 10 items (i.e. two tasks × five examinees). Three raters who had missing ratings were not included in the analysis. The result indicates that the raters showed high agreement when assessing the grammatical aspect of writing performance ($W = 0.805$, $\chi^2_{(9)} = 283$, $p < 0.001$).

4.2 Variations in examinees and tasks for grammar

The average scores on *grammar* assigned by the 42 raters were calculated for each examinee and task. Two examinees classified in the high group (E1 and E3) had almost identical test results on Task 1 while they showed a slight difference on Task 2. E1 (mean score of 4.366) outperformed E3 (mean score of 3.950) on Task 2, both of whom received a higher score on *grammar* for Task 2 than for Task 1. Two intermediate-level examinees (E4 and E5) showed almost no difference in their scores across the two tasks (3.619 and 3.575 for E4; 2.905 and 2.769 for E5). However, E4's *grammar* scores on both tasks appeared to be much higher than E5's, and even closer to the two high-level examinees' scores. This result might be because the groups were determined based on the overall test performance, in which grammatical knowledge is only part of the overall writing ability. In other words, E4 received a rather high score on *grammar* compared to the other intermediate-level examinee while their overall performance turned out to be similar. However, this difference within the intermediate group was not problematic because the examinees were modeled as random effects, and individual differences were counted during the CLMM analysis. The low-level examinee (E2) performed better on Task 1 while the difference was minimal.

Table 4. Descriptive statistics for grammar ratings for each examinee and task (N=42)

		Task 1		Task 2		Total	
Group	Examinee	Mean	SD	Mean	SD	Mean	SD
High	E1	3.857	0.521	4.366	0.488	4.108	0.563
	E3	3.881	0.593	3.950	0.552	3.915	0.571
Intermediate	E4	3.619	0.623	3.575	0.636	3.598	0.626
	E5	2.905	0.726	2.769	0.706	2.834	0.715
Low	E2	1.286	0.554	1.025	0.423	1.159	0.508
Total		3.110	1.150	3.145	1.312	3.127	1.231

* The total ratings of 410 were included due to 10 missing ratings.

4.3 Model selection

In order to find the best model, the backward elimination procedure was used. We started with a full model, including seven rater background characteristics (gender, overseas experience, teaching experience, GJT log-transformed response time, GJT score, MKT1 score, and MKT2 score), the examinee group and all of its possible interactions with the rater variables, rater random effects, and examinee random effects[3]. We then selected important variables using the backward elimination procedure, testing the significance of every term in the model, determining the least informative term among non-significant terms, removing it from the model, choosing the next best model, and repeating the procedure step by step until reaching the final model, which

[3] For ease of exposition, the full model was presented in the form of formulas in R software, but not in mathematical equations. The colons refer to interaction terms. clmm (GrammarRating ~ Gender + Gender:ExamineeGroup + YearsTeaching + YearsTeaching:ExamineeGroup + YearsOverSeasTrain + YearsOverseasTrain:ExamineeGroup + GJTscore + GJTscore:ExamineeGroup + GJTlogRT + GJTlogRT:ExamineeGroup + MKT1 + MKT1:ExamineeGroup + MKT2 + MKT2:ExamineeGroup + (1|Examinee) + (1|Rater))

cannot be reduced any more. Akaike's Information Criterion (AIC) was used for model comparisons. The model with the smallest AIC indicates the best model. Table 5 summarizes the results of fitting and comparing several CLMMs to the grammar rating data. The table also presents the results of a likelihood-ratio test (LRT) used to check the significance of each term within the model. That is, the -2 log-likelihood ratio (-2LLR) of the models was assessed using a chi-squared distribution, with the degree of freedom being equal to the difference in the number of parameters between the null and the alternative.

Table 5. Results of fitting CLMMs through the backward elimination procedure

Step	Term	df	AIC	LRT	p-value
1 removal	GJTlogRT:ExamineeGroup	2	689.410	0.330	0.848
2 removal	Gender:ExamineeGroup	2	686.540	1.126	0.570
3 removal	GJTlogRT	1	684.550	0.012	0.914
4 removal	Overseas experience:ExamineeGroup	2	682.880	2.330	0.312
5 removal	MKT2:ExamineeGroup	2	680.570	1.698	0.428
6 removal	MKT2	1	678.580	0.007	0.935
7 removal	Overseas experience	1	677.200	0.618	0.432
8 removal	Gender	1	677.690	2.490	0.115
9 selected	Teaching experience:ExamineeGroup	2	683.840	10.153	0.006
	GJTscore:ExamineeGroup	2	686.830	13.138	0.001
	MKT1score:ExamineeGroup	2	682.520	8.835	0.012

* MKT: Metalinguistic Knowledge Test, GJTscore: Grammatical Judgment Test Score, GJTlogRT: log-transformed GJT Response Time

The full model included seven rater variables and their interaction with the examinee group factor. All models tested in Table 5 included random terms for the rater effect and the examinee effect. The rater random effect controlled for individual differences between raters. The examinee random effect controlled for individual differences between the examinees for each task. Note that the examinees were assigned into groups according to their overall ability before the experiment (refer to footnote 2). Grouping was not made based on the grammar scores. Individual differences in the grammar scores were modeled by the examinee random effect.

The log-transformed GJT RT, overseas experience, MKT2 score, gender, and their interactions with the examinee group factor were not statistically significant (refer to the 1st to 8th removal steps in Table 5). On the other hand, teaching experience, the GJT score, the MKT1 score, and their interactions with the examinee group factor were statistically significant (refer to the 9th final selected step in Table 5).

The random effects were then tested after the selection of fixed effects. The rater random effect was highly significant (-2LLR = 66.68, df = 1, p < 0.001). This result indicates that the raters were basically different in their severity/leniency. In addition, the term for the examinee-task random effect was highly significant (-2LLR = 19.924, df = 1, p < 0.001). This result was expected because it is obvious that high-level examinees receive high scores. There was a significant interaction between the examinees and tasks (refer to Table 4). Therefore, the final model indicated that teaching experience, the MKT1 score, and the GJT score had effects on the ratings, but differently so, according to the level of the examinees' writing ability.

4.4 Results of the CLMM analyses

In order to analyze the effects of individual variables on the ratings, fixed effects were estimated. Table 6 summarizes the fixed effects estimation of CLMM by applying dummy coding for the examinee group variable. The high-level group was coded as the baseline category. The slopes of teaching experience (β = -0.024, z = -0.295, p = 0.768), the GJT score (β = 0.042, z = 0.947, p = 0.344), and the MKT1 score (β = -0.035, z = -0.209, p = 0.834) for the baseline group (high-level group) were not significantly different from zero. This result means that the ratings of the high-level students' writings were stable over the change of raters' teaching experience and grammatical knowledge (GJT score and MKT1 score). Overall, the slopes for the intermediate- and low-level groups were significantly different from the slope for the high-level group. This finding suggests that the ratings of the intermediate- and low-level students' writings tended to vary, depending on the raters' teaching experience and grammatical knowledge (refer to the following subsections for more details on the effects of each variable).

Table 6. Results for fixed effects estimation

Coefficient	Estimate	Std. Error	z-value	Pr(>|z|)
Teaching experience (years)	-0.024	0.080	-0.295	0.768
GJTscore	0.043	0.045	0.947	0.344
MKT1score	-0.035	0.166	-0.209	0.834
ExamineeGroup(Int)	3.850	2.338	1.647	0.100
ExamineeGroup(Low)	-1.691	3.336	-0.507	0.612
Teaching experience:ExamineeGroup(Int)	-0.087	0.069	-1.249	0.211
Teaching experience:ExamineeGroup(Low)	0.222	0.100	2.230	0.026
GJTscore:ExamineeGroup(Int)	-0.140	0.039	-3.534	0.000
GJTscore:ExamineeGroup(Low)	-0.093	0.055	-1.693	0.090
MKT1score:ExamineeGroup(Int)	0.015	0.144	0.107	0.915
MKT1score:ExamineeGroup(Low)	-0.537	0.205	-2.619	0.009

* MKT: Metalinguistic Knowledge Test, GJTscore: Grammatical Judgment Test Score

Teaching experience. According to Table 6, the slope of teaching experience for the high-level group was slightly negative, but it was not statistically significant (β = -0.024, z = -0.295, p = 0.768). The difference in the slopes between the high- and intermediate-level groups was not significant (β = -0.087, z = -1.249, p = 0.212). However, the difference in the slopes between the high- and low-level groups was statistically significant (β = 0.222, z = 2.230, p = 0.026). The results indicate that the experienced raters tended to be more lenient and tended to give higher scores to the low-level writing.

In Figure 1, the ascending lines refer to harsher ratings. On the other hand, the descending lines refer to more lenient ratings. The high- and intermediate-level examinee graphs demonstrate the ascending trend, whereas the low-level examinee shows a descending trend. Figure 1 illustrates the cumulative proportions of each rating scale (0 to 5) over the years of teaching experience, which are discretized into six groups: the no-experience group and five other levels with regard to the number of years. The cumulative proportions of ratings were calculated at each discretized[4] teaching experience group after adding 0.5 to zero cells in order to avoid computational problems. For the high- and intermediate-level examinee groups, the proportions of lower ratings tended to

[4] Teaching experience (years) is treated as continuous data in the modeling. It is discretized only for the graphs.

increase with longer teaching experience while they decreased with longer teaching experience for the low-level examinee. The slope of teaching experience for the high-level group was not significantly different from zero (β = -0.024, z = -0.295, p = 0.768). Undeniably, trend analysis is merely an exploratory analysis, and thus is not tested statistically. The statistical test of the regression slope is provided in Table 6. The slope of teaching experience for the high-level group was not significantly different from zero (β = -0.024, z = -0.295, p = 0.768).

The difference in the slopes between the high- and intermediate-level groups was not significant (β = -0.087, z = -1.247, p = 0.211). The difference between the high- and low-level groups was significant (β = 0.222, z = 2.230, p = 0.026). That is, the ascending pattern was not statistically significant in the high- or intermediate-level examinee graphs while the descending pattern was statistically significant in the low-level examinee graph.

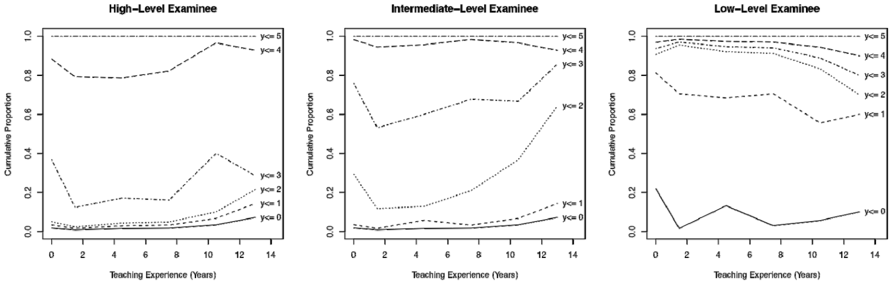

Figure 1. Cumulative proportions over teaching experience[5]

[5] The x-axis represents the independent variable. The ascending lines indicate that a rater tends to be harsher in proportion to his/her teaching experience. The descending lines indicate a tendency of a lenient trend. The y-axis represents the proportion. The line labeled as '$y \leq x$' represents the change of the proportions of ratings that are no higher than x at a given teaching experience point. For example, the line with the label $y \leq 3$ in the figure for the intermediate-level examinees represents the proportion rating of scores that are no greater than 3 at a given teaching experience point. There are 7 raters with zero teaching experience, and each rater evaluated 4 writing samples. These responses are summarized as follows: 7 responses of 2 points, 13 responses of 3 points, 6 responses of 4 points, and 2 missing responses. Consequently, the proportion of scores no greater than 3 (i.e. $y \leq 3$) is approximately 0.77 (20 responses out of 26), and the proportion of $y \leq 2$ is approximately 0.27 (7 out of 26). The actual lines in the figures are slightly, but unnoticeably different from these values because zero frequencies are replaced with 0.5 in order to avoid computational problems.

Explicit knowledge: MKT1 score. Referring to the results of the fixed effects estimation of CLMM (Table 6), it is evident that a rater with a higher MKT1 score gave lower ratings to the writings of the low-level examinee. The slope of the MKT1 score for the high-level group was slightly negative, but it was not statistically significant (β = -0.035, z = -0.209, p = 0.834). The difference in the slopes between the high- and intermediate-level groups was not significant (β = 0.015, z = 0.107, p = 0.915). However, the difference in the slopes between the high- and low-level examinees was statistically significant (β = -0.537, z = -2.619, p = 0.009). Figure 2 also shows the same rating patterns for different groups. The cumulative proportions of ratings for the low-level examinee increased with the raters' higher MKT1 scores while no obvious pattern was found for the high- or intermediate-level examinees. Therefore, the raters with higher MKT1 scores tended to be harsh when scoring the low-level examinee's writings. On the other hand, they did not show any differences in their rating patterns while evaluating the high- or intermediate-level examinees' responses.

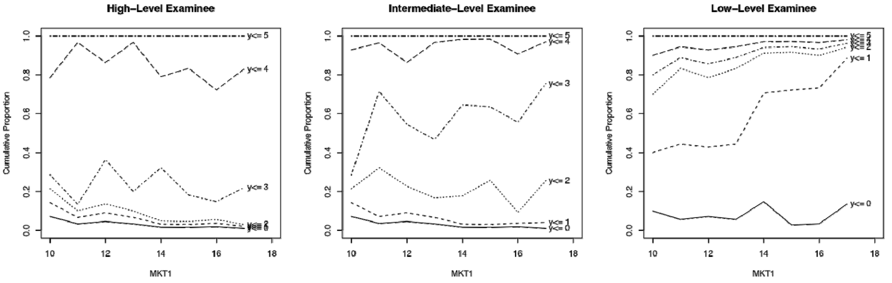

Figure 2. Cumulative proportions over MKT1 scores

* Since the range of the MKT1 score was not large, raw scores were used instead of discretizing the score range.

GJT score: Implicit knowledge. The fixed effects estimation of the CLMM and Wald test failed to reject the null hypothesis that the slope of the GJT score for the high-level group is zero (β = 0.042, z = 0.947, p = 0.344). The difference in the slopes between the high- and intermediate-level examinee groups was highly significant (β = -0.139, z = -3.534, p < 0.001). The difference in the slopes between the high- and low-level examinee groups was smaller and not significant (β = -0.093, z = -1.693, p = 0.090). In other words, the raters' implicit grammatical knowledge did not affect their severity/leniency pattern when scoring the high- or low-level examinees. However, it made a difference when the raters scored the intermediate examinees' writings. Those who had more implicit grammatical knowledge (higher GJT scores) tended to be harsher than those who had less implicit knowledge.

The effect of the GJT scores on the ratings is illustrated in Figure 3. The GJT scores are discretized with eight equally spaced intervals. The lines show a relatively steep increasing trend for the intermediate-level examinees and probably an increasing cumulative proportion over the GJT score for the low-level examinee. However, no specific pattern was observed for the high-level examinee group.

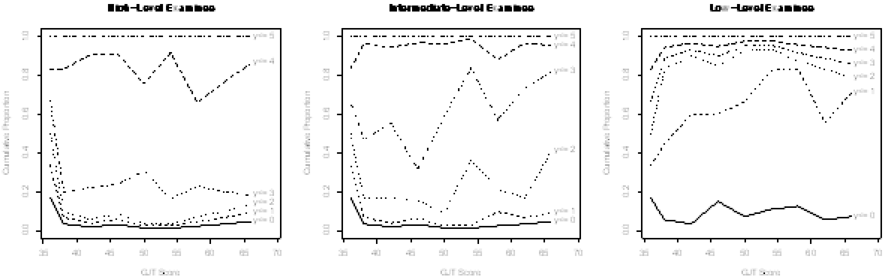

Figure 3. Cumulative proportions over GJT scores

5. Discussion

The current study examined rater effects on a writing performance assessment using CLMM in order to make up for the shortcomings of MFRM analyses, which have been dominantly used to investigate rater effects in previous studies. While MFRM analyses cannot separate different characteristics that raters bring into the rating context, the CLMM analyses enabled rater characteristics to be separated and linked individually to the raters' ratings. As a result, separate rater characteristics were tested to explain their individual effects on ratings. Since raters normally come from various backgrounds, and multiple variables jointly explain each rater with varying degrees of importance/impact, raters cannot be treated or expected to be experts or novices according to a single characteristic. A newly introduced analysis of rater behavior using CLMM contributes to a better understanding of rater effects.

Initial preliminary analyses of rater characteristics found that raters' background variables (i.e. overseas experience and teaching experience) and their explicit and implicit grammatical knowledge were not correlated with one another (refer to Table 3). This result suggests the need for further investigation of rater behavior, specifically in relation to raters' grammatical knowledge, given that raters' knowledge has rarely been considered in rater selection and training, despite its potential direct impact on scoring.

In fact, raters' teaching experience, the MKT1 score (explicit knowledge), the GJT score (implicit knowledge), and their interactions with the examinee group factor were found to be significant (refer to Table 5). This finding supports previous studies that treat teaching experience as one of the most important predictors for rater expertise. Oftentimes, raters' experience in teaching English has been used as a criterion to characterize or select raters in previous studies (e.g. Cumming 1990; Schoonen et al. 1997). At the same time, however, the significant effects of the MKT1 and GJT scores provide new perspectives regarding rater characteristics, in that raters' grammatical knowledge has largely been ignored in studies on rater behavior. Whereas teachers' own knowledge in a specific domain has been emphasized in teacher education for the development of teaching expertise (Wallace 1991), the importance of raters' knowledge concerning target language grammar, as well as other aspects of writing ability, has rarely been addressed in the literature.

In terms of the first research question, "Does raters' grammatical knowledge have an effect on the scoring of the grammatical aspect of writing performance?", the raters' explicit and implicit grammatical knowledge had an effect on their rating of grammar, and the effects varied across the different levels of examinees' written responses. The interaction terms (GJTscore:ExamineeGroup and MKT1score:ExamineeGroup in Table 5) were statistically significant. Interestingly, as shown in Table 6, the raters' explicit grammatical knowledge (MKT1 score) did not explain the differences in their ratings when scoring the high- or intermediate-level examinees' grammar.

However, the MKT1 score predicted the differences in their ratings when evaluating the low-level examinee's grammar; specifically, the raters who had more explicit grammatical knowledge (higher MKT1 scores) tended to be harsh when evaluating the low-level examinee's grammar. In other words, explicit grammatical knowledge played an important role in predicting differences in the ratings of the low-level essays, but not in evaluating the high- or intermediate-level essays. Related to the aforementioned characteristics of explicit grammatical knowledge (e.g. the use of metalinguistic knowledge and a focus on linguistic form), the raters used their metalinguistic knowledge (i.e. explicit grammatical knowledge) when rating the low-level essay. Alternatively, the raters were more likely to focus on linguistic form errors that were more numerous in the low-level essay than in the high- or intermediate-level essays. Their verbalizable explicit grammatical knowledge exerted a key role mainly in explaining differences in their ratings of the low-level essays.

Implicit grammatical knowledge (operationalized as the correct scores in the timed GJT) was associated with different rating patterns from explicit grammatical knowledge. Implicit knowledge was related to the ratings of both the intermediate- and low-level essays. The raters who had more implicit grammatical knowledge (higher GJT scores) were harsh when evaluating the intermediate- and low-level examinees' grammar. More interestingly, as shown in Table 6, the effect of implicit grammatical knowledge on ratings was greater when predicting differences in the ratings of the intermediate- versus low-level essays. Note that explicit knowledge played a significant role mainly in rating the low-level essays. A notable difference is that explicit knowledge relies on rules while implicit knowledge depends on unconscious, non-verbalized *feelings* (Ellis 2005). Given the characteristics of implicit knowledge, the greater effect of implicit knowledge on the ratings of the intermediate-level essays is likely to stem from the reliance on *feelings* rather than the dependence on explicit *rule* application.

To sum up, the raters tended not to be either strict or generous for higher-level groups, regardless of their grammatical knowledge while raters with high scores on MKT1 and GJT were additionally strict when assigning ratings to the lower-level examinee's responses. These results indicate that raters with more grammatical knowledge appear to be less tolerant toward grammatical errors.

These findings provide somewhat different insights on rater behavior from previous studies. Research that compared raters′ reactions to linguistic forms has reported that native speaker raters, who are supposed to have higher grammatical knowledge (at least higher implicit knowledge), were not different from non-native speaker raters (e.g. Kim 2009). Although native raters tended to pay less attention to linguistic forms and made more positive comments on L2 learners′ use of language, native and non-native raters did not show statistical differences in severity while evaluating L2 learners′ oral or written performance holistically (e.g. Shi 2001; Zhang and Elder 2011). Alternatively, the findings of the current study might relate to the nature of lower-level essays. It is possible that raters with high scores on the explicit and implicit measures activated their grammatical knowledge when scoring lower-level responses because grammatical features may be the only aspect to which they can pay attention due to low-level learners′ limited writing ability. On the other hand, they might have been able to evaluate other aspects of higher-level responses (e.g. content and organization); as a result, less grammatical knowledge might have been required or activated while scoring. Therefore, the relative contribution of grammar might have been different across the different examinee groups.

In addressing the second research question, "Are there any other factors that have an effect on raters′ scoring of the grammatical aspect of writing performance?", the raters′ teaching experience partly affected their scoring of the examinees′ grammar. The raters′ teaching experience did not explain the differences in scoring the high- or intermediate-level examinees′ grammar. However, it had different effects on the evaluation of the low-level examinee′s responses. Raters with more experience in teaching L2 were more lenient than those with less teaching experience. Contrary to the effects of raters′ grammatical knowledge, raters with more teaching experience were more tolerant toward grammatical errors made by the student with very limited writing ability. This finding supports the results from prior research on the effects of raters′ teaching experience. Previous studies have reported that less experienced teachers or raters with no TESOL experience tend to pay more attention to linguistic features, including sentence-level features and syntax, than more experienced teachers while rating examinees′ written responses (e.g. Brown 1991; Sweedler-Brown 1993; Cumming et al. 2002). The current study, however, provides further insights into raters' teaching experience by showing that their rating behavior (leniency/severity) toward grammatical features could vary across different levels of examinees' grammatical knowledge.

The findings of this study have several implications for rater selection and training in L2 performance assessment. First, rater behavior is not static; rather, it changes across what raters score. Previous research has also reported different degrees of rater severity/leniency in general, and in relation to a certain examinee ability group or task (e.g. Wigglesworth 1993, 1994; Du et al. 1996; Kondo-Brown 2002; Lumley 2002, 2005; Schaefer 2008). However, the current research advances previous studies by separating the effects of individual rater characteristics. In other words, it found that rater behavior can be changeable, according to a single attribute that a rater has. Diverse attributes might not determine rater behavior altogether. Therefore, this analysis of rater behavior in relation to separate rater characteristics contributes to making an accurate and detailed diagnosis of raters' strengths and weaknesses, which ultimately leads to customized rater training and education to meet the specific needs of individual or certain groups of raters.

In addition, this study found that raters were not balanced in implementing their explicit or implicit knowledge in the evaluations of the lower-level essays. Implicit knowledge may not be trained, given the nature of its unconsciousness. During the process of rater training, however, raters may be informed of their tendency to rely on explicit grammatical knowledge on lower-level essays, which is not the case for higher-level essays. Such information/feedback can help raters become aware of their increased attention to a certain aspect of responses for a particular examinee group. Consequently, such feedback can provide them with an opportunity to monitor their rating behavior consciously, thereby contributing to the development of their rating ability.

6. Conclusion

The results of this study indicate that the raters' grammatical knowledge and teaching experience had significant effects on their ratings. However, rater effects were different across the examinees' writing ability levels. Rater effects were not significant in scoring high-level writing responses while they were significant and varied in scoring the intermediate- and low-level responses.

Despite the findings of separate rater effects, there are a few limitations of the study that should be addressed. First, only grammatical knowledge (explicit and implicit) was included, which comprises only part of raters' knowledge concerning L2 writing. Other aspects of writing ability, such as organizational and pragmatic knowledge, need to be operationalized and tested individually, as well as together in order to better represent raters' knowledge about writing. In addition, the raters who participated in the current study did not present various backgrounds other than their experience in living/studying in English-speaking countries and experience in teaching English. This limitation was partly due to the fact the raters were recruited from the same graduate school. While their teaching experience had significant effects on the ratings, more studies are needed to test the effects of diverse backgrounds (e.g. experience in rating L2 writing for classroom assessment and/or large-scale, standardized assessment). At the same time, more restrictive criteria should be applied in selecting future rater participants. The current study included raters who were teaching EFL at different levels (e.g. children, adolescents, and adults) and in different contexts (e.g. public vs. private educational sectors). It would be more meaningful to investigate the extent to which teachers or raters who teach at a similar level and context exhibit similar rating behavior. Finally, only a very limited number of examinee responses and writing tasks were included for the raters' scoring in the present study. Although the focus of the study was on rater effects, the small number of examinee responses (particularly, only one low-level examinee) for a restricted number of tasks made it difficult to generalize any findings, particularly concerning rater behavior in relation to the different levels of examinee ability and the different types/topics/difficulty levels of tasks. Moreover, the three different levels of examinee responses were selected based on their overall writing performance, instead of grammatical control. Therefore, a sufficient number of examinees representing diverse grammatical knowledge levels, as well as a wide range of writing tasks, should be included in future studies.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 1999. *Standards for educational and psychological testing.* 2^{nd} edition. Washington, D.C.: American Educational Research Association.

Bachman, Lyle F. 1990. *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, Lyle F. 2002. Some reflections on task-based language performance assessment. *Language Testing* 19(4): 453-476.

Bachman, Lyle F. 2004. *Statistical analyses for language assessment.* Cambridge: Cambridge University Press.

Barrett, Steven. 2001. The impact of training on rater variability. *International Education Journal* 2(1): 49-58.

Bialystok, Ellen. 1982. On the relationship between knowing and using linguistic forms. *Applied Linguistics* 3(3): 181-206.

Brown, James D. 1991. Do English and ESL faculties rate writing samples differently? *TESOL Quarterly* 25(4): 587-603.

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

Chomsky, Noam. 1965. *Aspects of the theory of syntax.* Cambridge, MA: The MIT Press.

Christensen, Rune H. B. 2013. *Analysis of ordinal data with cumulative link models-estimation with the R-package ordinal.* R package version 2013.

Cumming, Alister. 1990. Expertise in evaluating second language compositions. *Language Testing* 7(1): 31-51.

Cumming, Alister, Robert Kantor, and Donald E. Powers. 2001. *Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: An investigation into raters' decision making and development of a preliminary analytic framework* (TOEFL Monograph Series, MS-22). Princeton, NJ: Educational Testing Service.

Cumming, Alister, Robert Kantor, and Donald E. Powers. 2002. Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal* 86(1): 67-96.

De Boeck, Paul and Mark Wilson (eds.). 2004. *Explanatory item response models: A generalized linear and nonlinear approach.* New York: Springer.

Dienes, Zoltan and Josef Perner. 1999. A theory of implicit and explicit knowledge. *Behavioral and Brain Sciences* 22(5): 735-808.

Doran Harold, Douglas Bates, Paul Bliese, and Maritza Dowling. 2007. Estimating the multilevel Rasch model: With the lme4 package. *Journal of Statistical Software* 20(2): 1-18.

Douglas, Dan. 1994. Quantity and quality in speaking test performance. *Language Testing*

11(2): 125-144.

Du, Yurong, Benjamin D. Wright, and William L. Brown. 1996. *Differential facet functioning detection in direct writing assessment.* Paper presented at the Annual Meeting of the American Educational Research Association, New York.

Dunbar, Stephen B., Daniel M. Koretz, and Hiram D. Hoover. 1991. Quality control in the development and use of performance assessments. *Applied Measurement in Education* 4(4): 289-304.

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

Ellis, Rod. 2005. Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition* 27(2): 144-159.

Ellis, Rod. 2006. Modelling learning difficulty and second language proficiency: The differential contributions of implicit and explicit knowledge. *Applied Linguistics* 27(3): 432-455.

Ellis, Rod, Shawn Loewen, Christopher P. Elder, Rosemary Erlam, Jenefer Philip, and Jan Hendrik Reinders (eds.). 2009. *Implicit and explicit knowledge in second language learning, testing and teaching.* Bristol, UK: Multilingual Matters.

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

Green, Peter S. and Karlheinz Hecht. 1992. Implicit and explicit grammar: An empirical study. *Applied Linguistics* 13(2): 168-184.

Gregg, Kevin R. 2003. The state of emergentism in second language acquisition. *Second Language Research* 19(2): 95-128.

Hamp-Lyons, Liz. 1996. The challenges of second language writing assessment. In Edward White, William Lutz, and Sandra Kamusikiri (eds.), *Assessment of writing: Policies, politics, practice.* New York: Modern Language Association.

Hamp-Lyons, Liz and Alan Davies. 2008. The Englishes of English tests: Bias revisited. *World Englishes* 27(1): 26–39.

Han, Youngju and Rod Ellis. 1998. Implicit knowledge, explicit knowledge and general language proficiency. *Language Teaching Research* 2(1): 1-23.

Hu, Guangwei. 2002. Psychological constraints on the utility of metalinguistic knowledge in second language production. *Studies in Second Language Acquisition* 24(3): 347-386.

Jang, Juhyun and Junkyu Lee. 2015. Comparing two types of explicit pronunciation instructions on second language accentedness. *Linguistic Research* 32(Special Edition): 15-32.

Johnson, Jeff S. and Gad S. Lim. 2009. The influence of rater language background on writing performance assessment. *Language Testing* 26(4): 485-505.

Johnson, Robert L., James Penny, and Belita Gordon. 2000. The relation between score resolution methods and interrater reliability: An empirical study of an analytic rating rubric. *Applied Measurement in Education* 13(2): 121-138.

Kenyon, Dorry. 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, Hyun Jung. 2011. *Investigating raters' development of rating ability on a second language speaking assessment*. EdD Dissertation. Teachers College, Columbia University.

Kim, Sun-Young. 2017. ESL college learners' perspective and its influence on reading-writing practices and development. *Linguistic Research* 34(Special Edition): 1-24.

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

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

Korea Institute for Curriculum and Evaluation. 2011. *Directions for the revisions of the national English ability test and English national curriculum*. Seoul: Korea: Korea Institute for Curriculum and Evaluation.

Lee, Junkyu. 2011. The activations of relational structures in processing second language noun-noun compound. *Linguistic Research* 28(1): 143-157.

Lim, Gad S. 2011. The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing* 28(4): 543-560.

Linacre, John M. 1989. *Many-faceted Rasch measurement*. Chicago: MESA Press.

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

Lumley, Tom. 2005. *Assessing second language writing: The rater's perspective*. Frankfurt, Germany: Peter Lang.

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

McNamara, Tim F. 1995. Modelling performance: Opening pandora's box. *Applied Linguistics* 16(2): 159-179.

McNamara, Tim F. 1996. *Measuring second language performance*. London: Longman.

McNamara, Tim F. 1997. 'Interaction' in second language performance assessment: Whose performance? *Applied Linguistics* 18(4): 446-466.

Meiron, Beryl E. and Laurie S. Schick. 2000. Ratings, raters and test performance: An exploratory study. In Antony John Kunnan (ed.), *Fairness and validation in language assessment. Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida*. Cambridge: Cambridge University Press.

Milanovic, Michael, Nick Saville, and Shaohua Shen. 1996. A study of the decision-making behavior of composition markers. In Michael Milanovic and Nick Saville (eds.), *Studies in language testing 3: Performance testing, cognition and assessment. Selected papers form the 15th Language Testing Research Colloquium*. Cambridge: Cambridge University Press.

Ministry of Education, Science and Technology. 2011. *An introduction to the elementary*

English national curriculum. Seoul, Korea: Ministry of Education, Science and Technology.

Popham, James. 1990. *Modern educational measurement: A practitioner's perspective.* 2nd edition. Englewood Cliffs, NJ: Prentice-Hall.

Reed, Daniel J. and Andrew D. Cohen. 2001. Revisiting rater and ratings in oral language assessment. In Tim McNamara, Kieran O'Loughlin, Catherine Elder, and Annie Brown (eds.), *Studies in language testing 11: Experimenting with uncertainty: Essays in honor of Allan Davies.* Cambridge: Cambridge University Press.

Rijmen, Frank, Francis Tuerlinckx, Paul De Boeck, and Peter Kuppens. 2003. A nonlinear mixed model framework for item response theory. *Psychological Methods* 8(2): 185-205.

Rumelhart, David E., James L. McClelland, and the PDP research group. 1986. *Parallel distributed processing: Explorations in the microstructure of cognition. Volume I.* Cambridge, MA: The MIT Press.

Sakyi, Alfred A. 2000. Validation of holistic scoring for ESL writing assessment: How raters evaluate compositions. In Antony John Kunnan (ed.), *19th Language Testing Research Colloquium. Fairness and validation in language assessment.* Cambridge: Cambridge University Press.

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

Schoonen, Rob, Margaretha Vergeer, and Mindert Eiting. 1997. The assessment of writing ability: Expert readers versus lay readers. *Language Testing* 14(2): 157-184.

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

Shohamy, Elana, Claire M. Gordon, and Roberta Kraemer. 1992. The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal* 76(1): 27-33.

Smith, David. 2000. Rater judgments in the direct assessment of competency based second language writing ability. In Geoff Brindley (ed.), *Studies in immigrant English language assessment* (vol. 1). Sydney: National Centre for English Language Teaching and Research, Macquarie University.

Sweedler-Brown, Carol O. 1993. ESL essay evaluation: The influence of sentence-level and rhetorical features. *Journal of Second Language Writing* 2(1): 3-17.

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

Wallace, Michael J. 1991. *Training foreign language teachers - A reflective approach.* Cambridge: Cambridge University Press.

Weigle, Sara Cushing. 1998. Using FACETS to model rater training effects. *Language Testing* 15(2): 263-287.

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

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

Wiseman, Cynthia S. 2008. *Investigating selected facets in measuring second language writing ability using holistic and analytic scoring methods.* EdD Dissertation. Teachers College, Columbia University.

Wolfe, Edward W., Chi-Wen Kao, and Michael Ranney. 1998. Cognitive differences in proficient and nonproficient essay scorers. *Written Communication* 15(4): 465-492.

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

Hyun Jung Kim
Associate Professor
Graduate School of TESOL
Hankuk University of Foreign Studies
107, Imun-ro, Dongdaemun-gu,
Seoul, 02450, Rep. of Korea
E-mail: hkim@hufs.ac.kr

Junkyu Lee
Associate Professor
Graduate School of Education
Hankuk University of Foreign Studies
107, Imun-ro, Dongdaemun-gu,
Seoul, 02450, Rep. of Korea
E-mail: junkyu@hufs.ac.kr

Hyun-Jo You
Teaching Associate Professor
College of Humanities
Seoul National University
1 Gwanak-ro, Gwanak-gu,
Seoul, 08826, Rep. of Korea
E-mail: youhyunjo@snu.ac.kr

Received: 2019. 04. 16.
Revised: 2019. 08. 14.
Accepted: 2019. 08. 16.