International Journal of Applied Linguistics & English Literature
ISSN 2200-3592 (Print), ISSN 2200-3452 (Online)
Vol. 5 No. 4; July 2016
Australian International Academic Centre, Australia

Examining Five Behaviors Conducted by Two Groups of Novice and Experienced Raters in Two Rating Processes

Saheb Mostofee (Corresponding author)
Persian Gulf University, Buoshehr, Iran
E-mail: emostofee@gmail.com

Nasim Ghanbari
Persian Gulf University, Buoshehr, Iran

Fateme Nemati
Persian Gulf University, Buoshehr, Iran

Received: 09-03-2016 Accepted: 10-05-2016 Advance Access Published: May 2016
Published: 01-07-2016 doi:10.7575/aiac.ijalel.v.5n.4p.199 URL: http://dx.doi.org/10.7575/aiac.ijalel.v.5n.4p.199

Abstract

This study aims at comparing five rating behaviors of 8 raters; four novice raters and four experienced raters. The five specific behaviors including number and frequency of referring to the rating scale (Jacobs' et al. EFL Composition Profile), number of interpretation (justification), total rating time, total score, and number of pauses longer than 5 seconds are compared between the two groups. The 8 raters were asked to rate two essays written by two B.A. students of English Literature attending their 4 semester at Persian Gulf university of Bushehr, Iran. Using TAPs the behaviors conducted by the eight raters were transcribed, then analyzed. It was found that although a similar pattern was observed in both groups' total scores assigned to the two essays, there was found no consistent trend in both the experienced and novice raters' number of referring to the rating scale. In addition, we found that the novice raters' number of referring to the rating scale, and number of pauses were higher than those of the experienced ones, while the experienced raters' number of interpretation (justification) and total rating time were higher than the novices'. The findings while supporting the findings by the previous research, paves the way for future researchers in this regard.

Keywords: Experienced raters, Novice raters, Rating behavior, Rating scale, Rating time

1. Introduction

Since the second half of the twentieth century, the research on writing, especially essay writing, as a main skill in language learning has received great attention, and a number of studies have certified its value as a field of research in teaching as well as testing. Within the field of language testing, the assessment of oral and written proficiency have been considered as the most common types of performance-based assessment (Swartz et al., 1999).

Assessing writing performance, as a productive skill is a complex and time-consuming process which has been the focus of attention by many testing specialists and testing journals. One area of modern research which has been the focus of great attention is the comparison of behaviors conducted by different raters with various expertise and experience in rating a written piece of writing (Barkaoui, 2007, 2008; Cumming et al., 2002; Li & He, 2015). Here, the role of raters as scorer who should attempt to assign as fair scores as possible becomes more identifiable. That is why many researchers have dealt with different behaviors of raters with various expertise and experience mostly between novice and experienced raters (Barkaoui, 2007, 2008, 2010 a, b; Erdosy, 2004; Li & He, 2015; Lumley, 2002; Winke & Lim, 2015; Yan, 2014).

In recent years, considerable attention has been paid to the raters and what they do when involved in the rating process (Ghanbari et al., 2012; Li & He, 2015; McNamara, 1996; Yan, 2014). In other words, many researchers have attempted to investigate if there are any differences between novice and experienced raters, identifying whether rating scales or contextual factors have any effect on raters' rating process and so many other research topics (Barkaoui 2008, 2010b; Cumming et al., 2002; Lumley, 2002, 2005; Winke & Lim, 2015).

The present study was mostly motivated by a desire to improve EFL rating practices in Iran. Generally speaking, this study is an attempt for gaining more insights about the differences between novice and experienced raters in terms of the five behaviors including the number of times the raters refer to the rating scale_ which in the current study Jacobs et al.'s EFL Composition Profile is selected, the number of interpretation (justification), total rating time, total score, and number of pauses longer than 5 seconds. In addition, the current study aims to determine the possible effects of raters' experience and expertise on the rating process. In simpler words, our guiding purpose in this research is to describe the behaviors that raters conduct while evaluating compositions (rather than the characteristics of the written texts that examinees produce).

To put it in a nutshell, it could be said that the previous studies (Barkaoui, 2010b) comparing novice and experienced raters did not consider the number and frequency of raters' referring to the rating scale (the effect of scale interaction between novice and experienced raters), nor did they consider the number of interpretations and pauses in their studies of comparing the novice and experienced groups.

2. Literature Review

2.1 Writing assessment

Among the four major language skills, creating a coherent and extended piece of writing has always been considered as the most difficult task to do in a language. Writing is a skill that even most native speakers of a language can hardly master. Furthermore, Behizadeh and Engelhard (2011) posited that writing is a crucial aspect of academic literacy and communicative competence. In the same vein, the profound number of studies on writing assessment implies that articulating sound and fair assessment of writing ability is a significant, difficult and time-consuming duty.

It can be noted that modern writing assessment was born in 1961 in Princeton, New Jersey. That year, Diederich, French, and Carlton of the Educational Testing Service (ETS) published Factors in Judgments of Writing Ability (ETS Research Bulletin 61-15) (cited in Broad, 2003). Basing their two-decade work on writing assessment, Diederich, French, and Carlton (1961) declared that their study intended to be served as a stepping stone toward closer agreement among judges of student writing by revealing common causes of disagreement.

The 1990s brought dramatic changes to the fields of psychometrics and writing assessment. In an attempt to provide a more valid picture of the construct of writing ability, there has been a major shift in language testing towards the development and use of performance test within the past decades. According to Abedi (2010) one of the main goals of performance assessment is to judge and assess the level of competency students achieve in doing functions such as reading or language arts, science, and mathematics (Parker et al., 2009). Therefore, performance assessments can also produce useful information for diagnostic purposes to assess what students know.

A significant issue in this regard is that writing assessment has always been considered as a kind of performance assessment mainly focusing on the evaluation of learners in the process of performing the assigned tasks. However, writing assessment procedures in academic contexts are a long way off from the pure form of performance assessment (Maftoon & Akef, 2010). Taking the concept of context into account, some researchers claimed that the main issue in the field of language assessment is to hold the notion of performance assessment as a means of achieving a close correlation between the test context and authentic language use (Lynch & McNamara, 1998).

2.2 Differences among Raters

Many factors have been indicated to have some kinds of influence on the raters' rating behavior such as prior experience in rating a written script, rating scale, background and expertise as ESL/EFL or English first language (L1) teacher or students. Rating styles have also been shown to have some kinds of impacts on the rating process. Rating style refers to how a rater reads the essay, interprets the rating scale, and assigns a score (Lumley, 2005; Sakyi, 2003; Smith, 2000). In other words, rating style refers to the differences among various behaviors conducted by raters.

However, most of the variations in rating processes are often ascribed to individual differences in raters' characteristics such as expertise and experience (Barkaoui 2008, 2010b; Cumming et al., 2002; Lumley, 2005; Winke & Lim, 2015). Only a small number of empirical studies have examined the differences between novice and experienced raters and their interaction with the scale.

In this regard, some researchers have implicitly assumed that less agreement on final score assignments among raters is due to variations in rating behaviors (e.g., Lumley, 2002). Researchers have suggested that low agreement may be due to contrasts between expert and novice raters (e.g., Barkaoui, 2010b). Li and He (2015) argued that the fact that there was considerable individual variability in text focus suggests that raters interacted with the scales in different ways. In this regard, Eckes (2008) made a claim that equally prepared and expert (trained) raters' inconsistency is due to their individual differences in how they read essays and which criteria they perceive as important.

Regarding the impact of rater characteristics on writing assessment, factors that affect the assessment of ESL writing have been categorized into three major categories: student-related, task-related, and rater-related (e.g., Brown, 1991; Huang, 2009; Kobayashi, 1992; Sakyi, 2000; Santos, 1988; Weigle, 1994). Of the three categories, rater-related factors are the most precarious in efforts to achieve fairness in assessment. This is because rater-related factors may jeopardize the reliability and in turn validity and fairness of writing assessment (Gamaroff, 2000; Vann et al., 1984; Johnson et al, 2005; Huang & Foote, 2010; Huang, 2008, 2009, 2012). Some of the rater-related factors included raters' academic disciplines (Brown, 1991; Santos, 1988; Song & Caruso, 1996; Vann et al., 1984; Weigle, Boldt, & Valsecchi, 2003), professional experiences (e.g., Barkaoui, 2010; Hamp-Lyons, 1994; Vaughan, 1991), linguistic backgrounds, tolerance for error, perceptions and expectations, and rater training (Huang, 2009).

One of the rater elements that seems to play a key role in the rating process is rater experience as well (e.g., Barkaoui, 2008, 2010b, Cumming, 1990; Lumley, 2005; Schoonen, Vergeer, & Eiting, 1997; Wolfe, 2006). Schoonen et al. (1997), for instance, argued that "the expertise and knowledge that raters bring to the rating task are essential for a reliable and valid rating" (p. 158). Comparing the behaviors of experienced raters and inexperienced raters, Cumming (1990) observed that experienced raters, compared with novice raters, typically showed greater capacity for consistently integrating multiple criteria and knowledge sources in their judgments on L2 writing. Also, Weigle (1998) noticed that inexperienced raters tended toward more extreme severity and less consistency in their ratings.

There is a relatively extensive literature on the effects of rater expertise in ESL essay rating processes (Barkaoui, 2007, 2010b; Cumming, 1990; Erdosy, 2004; Sakyi, 2003; Weigle, 2002). Barkaoui (2010b) indicated that experienced and novice raters employ qualitatively different rating processes. Also, Cumming (1990) found that experienced teachers had a much fuller mental representation of the essay assessment task and used a large and varied number of criteria, self-control strategies, and knowledge sources to read and judge ESL essays. Novice raters, by contrast, tended to evaluate essays with only a few of these component skills and criteria, using skill that may derive from their general reading abilities or other knowledge they have acquired previously (e.g., editing).

From the above review considering the above literature, we can easily come to this understanding that rater expertise which is directly related to the raters' experience plays a significant role in rater behavior and may affect rater performance while rating a written sample in profound way. We also understood that consistency among judgments of different raters is a matter of concern in EFL writing instruction which is highly dependent on the expertise, experience and other characteristics the raters bring to threating process.

2.3 Jacob's et al. (1981) ESL Composition profile

ESL Composition Profile by Jacobs et al. (1981) is one of the most well-known and widely-used analytic rating scale for assessing a written sample which has been adopted by many ESL assessment programs around the world (see Brakaoui, 2010a; Connor-Linton & Polio, 2014; Janssen et al., 2015 as cited in Winke & Lim, 2015) and it is "one of the best-known multi-trait rubrics in ESL" (Lee et al., 2010, p. 394).

More specifically, this Profile is divided into five major writing components: content, organization, vocabulary, language, and mechanics with each one having four rating levels of very poor, poor to fair, average to good, and very good to excellent. Each component and level has clear descriptors of the writing proficiency for that particular level as well as a numerical scale. For example, very good to excellent content has a minimum rating of 27 and a maximum of 30 indicating essay writing which is ''knowledgeable — substantive — thorough development of thesis — relevant to assigned topic'', while very poor content has a minimum of 13 and a maximum of 16 indicating essay writing that ''does not show knowledge of subject — nonsubstantive- not pertinent — or not enough to evaluate'' (Jacobs et al., 1981). The range and also weight for each of the writing skills are content 13–30 (30% of the total score), organization 7–20 (20% of the total score), vocabulary 7–20 (20% of the total score), language use 5–25 (25% of the total score) and mechanics 2–5 (5% of the total score).

The Jacobs' et al. (1981) ESL Composition Profile criteria was selected among other similar ones in the present study since it has been used successfully in evaluating the essay writing proficiency levels of students in ESL/EFL programs by the previous researchers (Brakaoui, 2010a; Connor-Linton & Polio, 2014; Janssen et al., 2015).

2.4 Think Aloud Protocols

Alongside survey questionnaires and interviews, think-aloud protocols (TAPs) have been widely used to examine rating processes of a writing performance (e.g., Barkaoui, 2007, 2010a, 2010b; Cumming et al., 2002; Huot, 1993; Lumley 2002, 2005; Wolfe, Kao, & Ranney, 1998; Vaughan, 1991). TAPs are often preferred over the retrospective ways of collecting data (e.g., questionnaires and interviews), in that concurrent verbalizations are believed to better reflect raters' thinking (Green, 1997). The proponents of TAPs claim that even if TAPs are not a perfect method, they still provide valid information regarding which aspect of writing raters paid attention to, or how raters made their decisions on scores (Ericsson & Simon, 1993, cited in Winke & Lim, 2015).

A number of studies on rating processes have legitimized the use of TAPs on the basis of Ericson and Simon's study or prior rating research (Cohen, 1996; Matsumoto, 1993; Huot, 1993; Vaughan, 1991). As a result, Think-aloud protocols (TAPs) are frequently used in research on essay rating processes. Think aloud protocols essentially require raters to verbalize their thought processes, impressions and feelings into a taping device as they rate designated scripts. These verbalizations are subsequently transcribed to allow analysis and interpretation by the researcher. This technique gains its validity mostly on the basis of Ericsson and Simon's (1993) work.

2.4.1 Advantages

In terms of the merits of think aloud data collection strategy, the literature points to several advantages of TAPs (e.g., Barkaoui 2010a; Cohen, 1998; Faerch and Kasper, 1987; Green, 1998; Kormos, 1998). Faerch and Kasper (1987), for example, argued that TAPs are 'particularly informative about informants' global approach to a task, the levels of decision making they operate on, and the considerations that govern their decisions' (p. 16). Furthermore, compared to other self-report methods, such as interviews and questionnaires, TAPs have the added advantage of being immediate, thus avoiding problems of information retrieval and/or filtering (Green, 1998). Additionally, TAPs are more likely to reflect what raters actually do and are concerned about as they read and rate essays, rather than what they believe they do and are concerned about as is usually revealed in interviews and questionnaires (Huot, 1993). Finally, while interviews and questionnaires provide generalized statements about behaviors, TAPs inspect specific instances of actual behaviors (Connor-Linton, 1995; Ericsson and Simon, 1987).

2.4.2 Disadvantages

Many of the previous studies which have made use of verbal protocol analysis admit that Think aloud strategy is not without its fair share of criticisms on its validity (See for example Stratman & Hamp-Lyons, 1994).

In spite of their being used frequently by many researchers, TAPs have their own limitations, however. First, TAPs are difficult to administer because participants are often not used to verbalizing their internal thoughts while focusing on the

completion of a task (Smagorinsky, 1994). In addition, the process of gathering, transcribing, coding, and analyzing data from TAPs is time-consuming and labor-intensive (Green, 1998; Smagorinsky, 1994). The main criticism of TAPs, however, concerns their veridicality and reactivity. Veridicality concerns whether the TAPs accurately report and represent the participants' true and complete thinking and rating processes, while reactivity concerns whether the requirement to report the rating process alters the process being observed and/or its outcomes (Barkaoui, 2010a; Ericsson & Simon, 1984/93; Lumley, 2005; Russo et al., 1989; Stratman & Hamp-Lyons, 1994). In terms of veridicality, Ericsson and Simon agreed that TAPs are incomplete because automatized processes, non-verbal states, and long-term memory contents are inaccessible to verbalization. As for reactivity, Ericsson and Simon (1984/93) argued that this depends on the type of verbalization asked from the rater.

Further, Sasaki (2003) found that TAPs do contain interactive and social features and that participants orient to a listener and are selective about what information to report while thinking aloud. More recently, Barkaoui (2010a) found that verbalization may draw participants' attention to certain elements of the task which is consistent with the literature (e.g., Russo et al., 1989).

3. Present study

3.1 Research questions

Specifically, the following question guides the present study:

3.1.1 Do novice raters refer to the rating scale (Jacobs' et al. ESL composition in this study) more than the experienced raters?

3.1.2 Is experienced raters' number of interpretations (justifications) higher than the novices'?

3.1.3 Is experienced raters' total rating time higher than the novices'?

3.1.4 Is novice and experienced raters' total score close to each other?

3.1.5 Is novice raters' number of pauses longer than 5 seconds higher than the experienced ones?

4. Methodology

4.1 Context of the study

The study, which is part of a thesis for M.A. degree was conducted at English Department in Persian Gulf University of Bushehr between spring 1393 and summer 1394. The researcher devoted more than one year for the study to be conducted.

4.2 Participants

Generally, 8 participants (i.e., raters) participated in this study. 4 out of these 8 participants were M.A. TEFL students at Persian Gulf University in Bushehr sharing somehow similar backgrounds in terms of qualifications, teaching experience and experience as raters of EFL, as shown in Table 3.1. These M.A. students are referred to by these pseudonyms during the study: Zohre, Sara and Mona as three M.A. female raters and Fardin as one M.A. male rater. Four experienced raters including Niki, Fata, Rama and Mola are assistant professors having taught English for many years, they were also teaching different courses of English as a foreign language at the time of the study.

The study included 2 essays on two tasks written by 2 intermediate B.A. EFL university students of English Literature under exam conditions. The essays were rated by these 8 EFL raters. Table 3.1 describes the four teachers by their pseudonyms.

Feature / Name	Gender/ Age	Education	Affiliation	Years of teaching & Assessing writing	Years of teaching English
Zohre	F*/ 26	M.A. st*	PGU*	0	1.5
Sara	F/25	M.A. st.	PGU	0	1
Mona	F/24	M.A. st.	PGU	0	1
Fardin	M*/26	M.A. st.	PGU	0	2
Niki	F/34	Ph.D.	PGU	6	10
Fata	F/39	Ph.D.	PGU	5	20
Rama	M/53	Ph.D.	PGU	6	20
Mola	M/49	Ph.D.	Teacher-training university	6	18

* In this Table, in gender section, F stands for female and M stands for male, and in affiliation section, PGU stands for Persian Gulf University, also st. stands for student.

Experienced raters were assistant professors having taught and rated EFL writing for at least 5 years. They have a Ph.D degree in TESOL, TEFL or Linguistics, all have received specific training in assessment and essay rating, and rated

themselves as competent or expert raters. Novice raters were MA graduates in TEFL who were not enrolled in or had not completed a pre-service or teacher training program in TEFL, had no TEFL writing teaching and rating experience at all at the time of data collection.

4.3 Sampling procedure

Sampling of the selected participants was quite non-randomized. The participants enrolled in this study were selected mostly based on the needs of the study, convenience of selection done by the researcher and also the intuitions and experiences gained from studying some previous studies in the field. In fact, the type of sampling used in the study was a convenient sampling procedure, the generalizations from which can be limited if not misleading (McMillan & Schumacher, 2001).

Regarding the selection of the essays, a stratified random sample of two essays was selected from a corpus of thirty-six final exam essays written by English major students at the end of a 4-month semester in the essay writing course, a two-credit course presented in the fourth semester of the English B.A. syllabus.

4.4 Data collection procedure

Four novice and four experienced raters rated a sample of two EFL essays while thinking aloud. All raters received an initial demonstration (training) and description of thinking aloud protocols and Jacob's et al. (1981) analytic ESL composition profile, respectively. This 20-minute training session involved some instruction on what they should do while thinking aloud and how they should make use of Jacob's et al analytic rating scale, but raters were left to their own devices, strategies and judgments as to how to rate the compositions.

Because, it is not just rubric design that may affect scoring; rubric training (in this study better to say instructing) patterns may affect scoring (see Johnson & Lim, 2009), the only training raters received was on TAPs, and how should they verbalize their feelings (cf. Hamp-Lyons & Henning, 1991). In other words, each think-aloud rater attended a 20-minute session where they received detailed instructions (adopted from Cumming et al., 2001, pp. 83–85) and careful training, following procedures in Ericsson and Simon (1984/93), on how to think aloud when rating the essays. During this session, the technique was explained and demonstrated to the participant following Barkaoui (2007). After this, each rater completed a background questionnaire concerning their name (optional), gender, education, affiliation, years of TEFL teaching, rating and writing assessment experience.

As in previous studies (e.g., Barkaoui, 2008, 2010b; Cumming, 1990; Cumming et al., 2002; Wolfe, 2006; Wolfe, Kao, & Ranney, 1998), the focus in this study is on comparing the frequency of the behaviors experienced and novice raters conduct. For this reason, think aloud protocols were utilized as the data collection procedures. The raters were asked to verbalize whatever they did (e.g., referring to the scale or topics, expressing their feelings and so on) while rating each essay as naturally as possible.

The verbal protocols were taped and subsequently transcribed. Each tape-recorded audio files of think-aloud protocols were transcribed by researcher into word-processing files. The think-aloud protocols were extensive, ranging in length from 10 to 21 typed pages per rater, or between 1-6 page(s) per composition. The final verbal protocols from think-aloud were lengthy, totaling 15,879 words. Afterward, to assure accuracy, the transcriptions were verified and/or changed by the person who had originally produced the verbal report. Finally, it is worth mentioning that sixteen think-aloud protocols were collected for this study (8 raters × 2 essays ×1 rating scale). So, the results concern sixteen think-aloud protocols. It should be said that the results of this study should be interpreted in recognition of its limitations, stated in the limitation section of conclusion chapter (see 6.4).

4.5 Data analysis

The study mainly used think-aloud protocols that the raters were trained to produce. Raters' comments along the margin were also inspected to provide detailed insights into the rating process. Before everything, the decision-making behaviors displayed in the think-aloud protocols were analyzed. Then the original protocols were segmented into "idea units" using the definition by Brown, et al., (2005) stating that an idea unit includes "a single or several utterances, either continuous or separated by other talk, but falling within the same turn, with a single aspect of the performance as the focus" (p. 14, cited in Le & He, 2015). Such a unit may consist of one clause or many clauses but each of them centered on a dominant behavior.

The outputs achieved through TAPs were analyzed separately, and the findings were declared in some charts and tables. In fact, the data analysis used in this study is qualitative in nature. Quantitative results (the number, frequency and type of rating behavior among novice and experienced EFL raters) are also triangulated with qualitative rater comments (TAPs) to arrive at a more representative picture of rater performance.

The present study combined both quantitative and qualitative methods to better understand possible effects of raters' experience and expertise on the rating process, the times at which the raters refer to the rating scale (Jacobs' et al. EFL Composition Profile), number of interpretation (justification), total rating time, total score, and number of pauses longer than 5 seconds among novice and experienced raters, all happening in an EFL context.

5. Results and Discussion

5.1 Results

In order to compare the rating behaviors of the 8 raters, five general features indicated by the raters are summarized in the following table. These features are: 1- The number of times both novice and experienced raters refer to the rating

scale regarding their assigning of scores and for understanding the descriptors given in the scale as well, 2- The number of interpretations (justifications) raters show in the rating process for assigning the scores, 3- The rating period in which both novice and experienced raters completed their rating, 4- The total score each rater assigned to the two essays, and 5- The number of pauses more than 5 seconds indicated by each rater during the rating process, a component used by Cumming et al. (2002) to segment the TAPs of their study comparing the decision-making behaviors of the raters.

5.1.1 Some general information indicated by four experienced raters in their rating process

In the following table, five general features mentioned above are indicated for four experienced raters in this study.

Table 5.1 Features shown by four novice raters: (Ns)

Rater / Essay	N1	N2	N3	N4	feature
Higher level essay**	4(2nd)*	3(1st)*	4(1st)	6(2nd)	Referring to the rating scale
Lower level essay	3(1st)	3(2nd)	6(2nd)	3(1st)	
Higher level essay	5	1	0	1	Number of interpretation (Justification)
Lower level essay	1	0	0	5	
Higher level essay	20 (2nd)	7 (1st)	9(1st)	7(2nd)	Rating period
Lower level essay	10(1st)	11(2nd)	12(2nd)	14(1st)	
Higher level essay	76	84	76	79	Total score
Lower level essay	39	47	52	63	
	31	30	17	26	Number of pauses more than 5 seconds

* 1st shows that this essay is the first essay rated by the rater.

* 2nd shows that this essay is the second essay rated by the rater.

** Higher level essays are better in quality receiving higher marks, but lower level essays are poor in quality receiving lower marks.

Table 5.1 indicates that there is not a consistent trend in the novice raters' number of referring to the rating scale taking the level of essay into account; specifically, novice raters 1 and 4 have referred to the rating scale in their rating of higher level essay more than the rating of lower level essay, while novice rater 3 has referred to the rating scale in her rating of lower level essay more than the rating of higher level essay. Further, novice rater 2 has referred to the rating scale twice in both essay rating processes. But regarding the order of the rating, it should be mentioned that three out of four novice raters have referred to the rating scale in their second rating more than their first rating of the essays.

Regarding the number of interpretations, again, there is not a consistent pattern in the novice raters' number of interpretations and justifications. For example, novice rater 1 showed five justifications for his assigning scores in the higher level essay and just one justification in the lower level essay while the novice rater 4 functioned exactly in the reverse patter showing one justification for her assigning scores in the higher level essay and five justifications in the lower level essay. It is worth mentioning that novice rater 3 conducted no interpretation for both higher and lower essay at all.

Furthermore, regarding the rating period, 3 out of 4 novice raters rated the second essay in a longer period than the first one, but novice rater 4 allocated a period to the first essay rating process two times higher than the second essay rating process. As a result, it can be said that three novice raters considered more times for assessing the second essay than the first one considering their two rating processes.

Finally, a consistent pattern was seen in this table taking the scores given by 8 raters into consideration; all four novice raters assigned higher scores to the higher level essays and lower scores to the lower level essay. Interestingly, two novice raters 1 and 3 assigned the same score of 76 to the higher level essay. It is worth mentioning that the range of difference (variance) between the scores given to the higher level essays by four novice raters was 8, while the range of difference (variance) among the scores given to the lower level essays by the same raters was 24. The novice raters'

number of pauses more than 5 seconds will be compared with that of experienced raters in the following section of the comparison of novice and experienced raters' general features (see 5.1.3).

5.1.2 Some general information indicated by four experienced raters in their rating process

In the following table, five general features mentioned in the above table (5.1) for novice raters are restated this time for four experienced raters in this study. After this section, a section of comparison between novice and experienced raters regarding these general features will be presented (see 5.1.3).

Table 5.2 Features shown by four experienced raters: (Es)

Rater / Essay	E1	E2	E3	E4	Feature
Higher level essay	3(1st)*	3(1st)	3(2nd)*	4(1st)	Referring to the rating scale
Lower level essay	1(2nd)	5(2nd)	4(1st)	2(2nd)	
Higher level essay	5	5	5	7	Number of interpretation (Justification)
Lower level essay	8	4	3	3	
Higher level essay	37(1st)	19(1st)	9(2nd)	21(1st)	Rating time
Lower level essay	37(2nd)	15(2nd)	20(1st)	12(2nd)	
Higher level essay	75	70	73.50	58	Total score
Lower level essay	65	41	67	36	
	8	2	2	3	Number of pauses more than 5 seconds

* 1st shows that this essay has been the first essay rated by the rater.

* 2nd shows that this essay has been the second essay rated by the rater.

According to table 5.2, there is not a consistent trend in the experienced raters' number of referring to the rating scale the same as novice raters; experienced raters 1 and 4 have referred to the rating scale in their rating of higher level essay more than the rating of lower level essay, while experienced raters 2 and 3 have more referred to the rating scale in their rating of lower level essay than the rating of higher level essay. Regarding the experienced raters order of rating, two out of four experienced raters have referred more to the rating scale in their first ratings and the remaining two raters have referred more to the rating scale in their second ratings.

Regarding the number of interpretations, there is an approximate consistent pattern in the experienced raters' number of interpretations and justifications. Three experienced raters showed higher justifications for their assigning scores in the higher level essay and just the experienced rater 1 indicated more justifications in the lower level essay than higher level essay, showing five justifications for her assigning scores in the higher level essay and eight justifications in the lower level essay.

Furthermore, regarding the rating period, 3 out of 4 the experienced raters rated the first essay in a longer period, but interestingly experienced rater 1 allocated the same period of 37 minutes to the first essay rating process and the second essay rating process. As a result, it can be said that three experienced raters considered more times for assessing the first essay regarding their two rating processes. Such a trend could be because of their physical tiredness or their familiarity with the scale components.

Finally, similar to the total score given by novice raters, a consistent pattern was also seen in the experienced raters' total scores assigned to the two essays. According to table 5.2, all four experienced raters assigned higher scores to the higher level essays and lower scores to the lower level essay. An interesting finding was that the experienced rater 3 assigned a decimal score of 73.50 to a higher level essay, which he gave 4.50 out of 5 score of mechanics in the rating process. It is worth mentioning that the range of difference (variance) between the scores given to the higher level essays by four experienced raters was 17, while the range of difference (variance) among the scores given to the lower level essays by the same raters was 29. The experienced raters' number of pauses more than 5 seconds will be compared with that of experienced raters in the following section of the comparison of novice and experienced raters' general features (see 5.1.3).

5.1.3 The comparison of novice and experienced raters' general features stated in tables 5.1 and 5.2

According to tables 5.1 and 5.2, there is not a consistent trend in both the experienced and novice raters' number of referring to the rating scale. Two novice raters referred to the rating scale in their rating of higher level essay more than the rating of lower level essay, and one novice rater referred to the rating scale in her rating of lower level essay more than the rating of higher level essay. Also one novice rater referred to the rating scale twice in both essay rating processes. For experienced raters, two raters have more referred to the rating scale in their rating of higher level essay than the rating of lower level essay, and two experienced raters have more referred to the rating scale in their rating of lower level essay than the rating of higher level essay. Regarding the order of the rating, it should be mentioned that three out of four novice raters have referred to the rating scale in their second rating more than the first rating of the two essays. In contrary, three out of four experienced raters have referred more to the rating scale in their first rating and only one rater has referred more to the rating scale in his second rating than the first one.

Regarding the number of interpretations, there was an approximate consistent pattern in the experienced raters' number of interpretations and justifications, while there was not any consistent pattern in the novice raters' number of interpretations and justifications. Three experienced raters showed higher justifications for their assigning scores in the higher level essay and just one experienced rater indicated more justifications in the lower level essay than higher level essay. However, novice rater 1 showed five justifications for her assigning scores in the higher level essay and just one justification in the lower level essay while the novice rater 4 functioned exactly in the reverse patter while novice rater 3 did not show any interpretation in neither of the essays.

Furthermore, regarding the rating period, it can be mentioned that three experienced raters considered more times for assessing the first essays regarding their two rating processes. However, regarding the rating period of novice raters, it can be said that three novice raters considered more times for assessing the second essay regarding their two rating processes. As a result, the rating period assigned by experienced raters is in obvious contrast with the rating period assigned by novice raters in terms of the order of essays rated.

Finally, similar to the total score given by novice raters, a consistent pattern was also seen in the experienced raters' total scores assigned to the two essays. According to table 5.2, all four experienced raters assigned higher scores to the higher level essays and lower scores to the lower level essay, in the same manner, all four novice raters assigned higher scores to the higher level essays and lower scores to the lower level essay. In addition, the range of difference (variance) between the scores given to the higher level essays by four experienced raters was 17, while the range of difference (variance) among the scores given to the lower level essays by the same raters was 29, while the range of difference (variance) between the scores given to the higher level essays by four novice raters was 8, while the range of difference (variance) among the scores given to the lower level essays by the same raters was 24. Therefore, the range of difference (variance) between the scores given to the higher level essays by four experienced raters was higher than its counterpart assigned by four novice raters. In the same regard, the range of difference (variance) between the scores given to the lower level essays by four experienced raters was also higher than its counterpart assigned by four novice raters.

The sum of the number of pauses more than 5 seconds shown by the four novice raters was 104 with novice rater 1 (N1) showing the highest value of 31 and novice rater 3 (N3) showing the least value of 17, but the sum of the number of pauses more than 5 seconds shown by the four experienced raters was 15 with the experienced rater 1 (E1) showing the highest value of 8 and experienced raters 2 and 3 (N2, N3) showing the least value of 2. Therefore, the novice raters' number of pauses was almost seven times higher than that of the number of pauses indicated by experienced raters.

5.2 Discussion

This study attempted find answer to the five research questions stated. It should be added that none of the features stated in the research questions have been examined by the previous researchers except for (Barkaoui 2010b) and Total score (Barkaoui, 2008, 2010b; Huang & Foote, 2010; and Johnson & Lim, 2009). Also the number of pauses more than 5 seconds has been first considered as a factor by Cumming et al (2002) to segment the think-aloud protocols into separate, comparable units of decision making, while no further examination has been conducted in terms of the differences between novice and experienced raters in this regard.

Regarding the first research question, generally, the four novice raters referred to the scale more than the experienced raters did and no consistent trend was found in both the experienced and novice raters' number of referring to the rating scale in term of the essay level. Also, regarding the order of the rating, it should be mentioned that three out of four novice raters have referred to the rating scale in their second rating more than their first rating of the two essays. In contrary, three out of the four experienced raters have referred more to the rating scale in their first rating and only one experienced rater has referred more to the rating scale in his second rating. Therefore, the novice raters have referred to the rating scale in their second rating process more than their first rating process.

One possible reason for this could be their gaining experience from the first rating and wanting to seem more judicial bringing more fairness to their scores assigned. Further, the experienced raters have referred to the rating scale in their first rating process more than their second rating process mostly because of their prior experience in this regard. According to table 5.2, also there was not a consistent trend in the experienced raters' number of referring to the rating scale the same as the novice raters. This finding is in line with Barkaoui's (2010b) finding that novice raters tended to refer to the rating scales, "the source of the evaluation criteria, more frequently than did the experienced raters who referred more often to the essay, the focus of the assessment, regardless of the rating scale used" (p. 64).

In the researcher's perspective, one reason for the novice raters' higher number of reference to the rating scale could be their lack of experience with the rating process, the reason also stated by Barkaoui (2010b). Furthermore, some reasons behind novices' lack of referring to the rating scale for learning it could be their first reading of it in the training session, then memorizing it well, or their consideration of the rating scale examination as something which is not that much important in the rating process.

Regarding the second research question, the study showed that there was an approximate consistent pattern in the experienced raters' number of interpretations and justifications, while there was not any consistent pattern in the novice raters' number of interpretations and justifications. Therefore, it can be concluded that the experienced raters showed higher justifications for their assigning scores in the higher level essay than the novices did while the novices did not show any specific trend in this regard.

Regarding the third question, the study found that the experienced raters tended to allocate more time to reading and assessing the essays overall, than the novices did (Bukta, 2007; Cumming, 1990; Barkaoui, 2010b; Milanovic et al., 1996). Contrary, some researchers found that the novice raters tend to spend more time interpreting and/or editing text than the experienced raters did (cf. Cumming, 1990; Sakyi, 2003). More specifically, it can be mentioned that the experienced raters considered more times for assessing the first essay while the novice raters considered more times for assessing the second essay. As a result, the rating period assigned by experienced raters is in obvious contrast with the rating period assigned by novice raters in terms of the order of essays rated. In addition, novices tended to focus on specific, local aspects of writing more often than the experienced raters did (Barkaoui, 2010b). Furthermore, Bukta (2007) stated that although fatigue of attention is significantly observable in large scale testing context (Congdon & McQueen, 2000, cited in Bukta, 2007), a decrease in number of words by the end of the rating process in some of the protocols indicates that raters' attention may decrease with time. She added that this tendency was more obvious for the novice raters than for the experienced ones (ibid).

In terms of the forth research question, the study indicated that similar to the total score given by novice raters, a consistent pattern was observed in the experienced raters' total scores assigned to the two essays. The experienced raters assigned higher scores to the higher level essays and lower scores to the lower level essay, in the same manner, all four novice raters assigned higher scores to the higher level essays and lower scores to the lower level essay.

Finally, regarding the sum of the number of pauses more than 5 seconds (the fifth research question), it was shown that the novice raters' number of pauses was almost seven times higher than that of the number of pauses indicated by experienced raters. Therefore, it can be concluded that the novice raters' speed of rating was highly less than the experienced ones', which could be because of their lack of experience and expertise in writing assessment process.

6. Conclusion

The study found no consistent trend in both the experienced and novice raters' number of referring to the rating scale. As stated earlier, the novice raters have referred to the rating scale in their second rating process more than their first rating process. Further, the experienced raters have referred to the rating scale in their first rating process more than their second rating process.

Regarding the number of interpretations, it was found an approximate consistent pattern in the experienced raters' number of interpretations and justifications, while there was not any consistent pattern in the novice raters' number of interpretations and justifications (Bukta, 2007). We found that the experienced raters showed higher justifications for their assigning scores in the higher level essay than the novices did while the novices did not show any specific trend in this regard.

Regarding the rating period, it was found that the experienced raters tended to allocate more time to reading and assessing the essays overall, than the novices did (Bukta, 2007; Cumming, 1990; Barkaoui, 2010b; Milanovic et al., 1996) while this finding is in contrast with some researcher's findings that the novices tended to spend more time interpreting and/or editing text than the experienced raters did (cf. Cumming, 1990; Sakyi, 2003). Furthermore, previous research stated that fatigue of attention is significantly observable in large scale testing context indicating that raters' attention may decrease with time, a tendency more apparent for the novice raters than for the experienced ones (Congdon & McQueen, 2000, cited in Bukta, 2007; Bukta (2007).

In addition, a consistent pattern was observed in the experienced raters' total scores assigned to the two essays. It was also found that the range of difference between the scores given to both lower and higher level essays by four experienced raters was higher than its counterpart assigned by four novice raters.

Finally, regarding the sum of the number of pauses more than 5 seconds, it was shown that the novice raters' number of pauses was almost seven times higher than that of the number of pauses indicated by experienced raters. Therefore, it can be concluded that the novice raters' speed of rating was highly less than the experienced ones', which could be because of their lack of experience in the writing assessment process (Barkaoui, 2010b; Bukta, 2007).

6.1 Some lateral findings of the study

Some lateral findings of the current study which some of them are supported by the previous research and some need further research are as followings:

Based on the analysis of the TAPs, it was found that the most influencing factors which the raters, especially the experienced raters made their decisions upon them were the raters' experiences in the writing assessment and their overall impression even made only by the first looking at the essays or by reading of the first paragraph. In the same

vein, Lumley (2000) concludes that "the raters' overall impression, based on their professional judgment, is the primary influence in their assessments" (p. 281). Barkaoui (2007) confirming the existence of some hidden effective elements in the rating process, has also referred to the raters' impression and reaction to the essay as 'internal criteria' (p. 101). Also, in an article aiming at finding the impact of rater experience on both essay holistic scores and these associations, Barkaoui (2010b) indicated rater experience is effective regarding the raters' rating process.

Furthermore, Lumley (2005) emphasizes the centrality of raters' knowledge and experiences stating that people assess what they believe, have learned, and value. Therefore, professional and educational experience as referred to by some raters in this study, are among the very elements that influence the judgment of raters profoundly (Barkaoui, 2007, 2010b; Mendelsohn & Cumming, 1987; Santos, 1988).

As another lateral finding we can refer to the observation of differences between novice and experienced raters' strategies regarding the problem-seeking in the study. It was observed that the novice raters' main strategy in this regard was that they just read the essays to find the problems and faulty parts. That is they spend a considerable amount of time seeking the problems to be found almost without any justification and interpretation of the problems' causes. But, the experienced raters' main strategy was seeking for the problems' causes interpreting and justifying their rating decisions giving some kinds of remedies in most cases. After conceptualizing the type of problem that has been presented, the expert solves the problem quickly and accurately.

As the last lateral finding of the current study we can refer to the observation that none of the novice raters created, not even thought of, a new rating scale based on which to assess the two essays. But in the other side of the coin, three out of four experienced raters (E1, E3 and E4) devised their own rating scale changing the value order of different components mostly because of their prior experience in the assessment of writing scripts. Again, it can be concluded that the role of experience and expertise is very considerable in the rating process to be given special weight.

6.2 Implications

The results of this study provide several pedagogical implications for teachers in developing appropriate argumentative essay instruction.

First, it follows that finding answers to the designed research questions can give us more understandings of the language assessment and testing domain in general and essay rating process and writing assessment in particular. It might also give us deeper insights about assigning more accurate and fair scores to any piece of writing.

Moreover, one of the major objectives of the study is to make clear the existing discrepancies between experienced and novice raters and identifying different aspects, strategies and behaviors that these two groups of raters use and conduct in assessing an essay. Furthermore, findings of the current study indicate that triangulating different sources of data on rater behaviors in writing assessment using a mixed-methods approach would be very effective, especially in local testing contexts.

6.3 Suggestions for further research

It should be admitted that some unresolved concepts emerged during this research process, which can serve as directions for future research. First, experienced raters 1 and 4 who were female conducted 27 and 22 rating stages, respectively. But, experienced raters 2 and 3 who were male conducted 13 and 17 rating stages, respectively. As a result, it may be inferred from this observation that gender, a factor not examined in this study, may also affect the rating process.

In addition, further research needs to examine higher number of raters in terms of the differences between novice and experienced raters in an EFL context and examine the cultural assumptions. Also, further researchers can conduct this study in an ESL context to see if the findings are similar to the findings of this study conducted in an EFL context.

6.4 Limitations

The current study has limitations that future research should address. The following limitations need to be taken into account when interpreting the findings and conclusions of this study.

First, the finding that the completeness and reactivity of thinking aloud were not uniform across individuals and contexts makes comparison of TAP data across contexts, individuals and groups problematic.

Second, the think-aloud protocols were analyzed quantitatively. In addition, the terminology used to describe rating behaviors is open to different interpretations; raters might have meant different things by using the same term or meant the same thing by different terms (Cumming et al., 2001, p. 71).

Third, the training session on the use of TAPs may not be enough for the 8 raters to be able to handle the rating processes using think aloud protocols completely.

Fourth, the impact of gender on the rating process was not examined in this study while some discrepancies in terms of rating behaviors were found between female and male raters.

References

Abedi, J. (2010). *Performance Assessments for English Language Learners.* Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing. Elsivier*.12: 86–107.

Barkaoui, K. (2008). Effects of scoring method and rater experience on ESL rating outcomes and processes (Unpublished doctoral dissertation). University of Toronto, Toronto, Canada.

Barkaoui, K. (2010a). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. *Language Testing*, 28(1), 51–75.

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

Behizadeh, N., & Engelhard, G. (2011). Historical view of the influences of measurement and writing theories on the practice of writing assessment in the United States. *Assessing writing, 16*(3), 189-211. http://dx.doi.org/10.1016/j.asw.2011.03.001.

Broad, B. (2003). *What We Really Value: Beyond Rubrics in Teaching and Assessing Writing.* All USU Press Publications. Book 140.

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

Bukta, K. (2007). Processes and outcomes in L2 English written performance assessment: Raters' decision-making processes and awarded scores in rating Hungarian EFL learners' compositions. (Unpublished doctoral dissertation). Hungary.

Cohen, A. D. (1996). Verbal reports as a source of insights into second language learner strategies. *Applied Language Learning,* 7(1–2), 5–24.

Cohen, A. D. (1998). *Strategies in learning and using a second language*. London: Longman.

Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement*, 37(2), 163-178.

Connor-Linton, J. (1995). Looking behind the curtain: What do L2 composition ratings really mean? *TESOL Quarterly, 29,* 762–765.

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

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

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

Diederich, P. B., French, J., and Carlton, S. (1961). *Factors in judgments of writing ability*. ETS Research Bulletin 61-15. Princeton, NJ: Educational Testing Service.

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

Erdosy, U. (2004). Exploring variability in judging writing ability in a second language: A study of four experienced raters of ESL compositions (TOEFL Research Report No. RR-03-17). Princeton, NJ: Educational Testing Service.

Ericsson, K. & Simon, H. (1984). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Ericsson, K. & Simon, H. (1987). Verbal reports on thinking. In C Faerch and G Kasper (eds), Introspection in second language research (pp. 24–53). Clevedon: Multilingual Matters.

Ericsson, K. & Simon, H. (1993). *Protocol analysis: Verbal reports as data* (rev. ed.). Cambridge, MA: MIT Press.

Faerch, C. and Kasper, G. (1987). From product to process: Introspective methods in second language research. In C Faerch and G Kasper (Eds), *Introspection in second language research* (pp. 5– 23). Clevedon: Multilingual Matters.

Gamaroff, R. (2000). Rater reliability in language assessment: The bug of all bears. *System, 28,* 31-53.

Ghanbari, B. Barati, H. and Moinzadeh, A. (2012a). Rating Scales Revisited: EFL Writing Assessment Context of Iran under Scrutiny. *Language Testing in Asia*, *2* (1), 83-100.

Green, A.J.K. (1997). *Verbal protocol analysis in language teaching research*. Cambridge University Press and University of Cambridge Local Examinations Syndicate.

Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook*. Cambridge: Cambridge University Press.

Hamp-Lyons, L. (1994). Rating non-native writing: The trouble with holistic scoring. *TESOL Quarterly*, *29*(4), 759–762.

Hamp-Lyons, L., & Henning, G. (1991). Communicative writing profiles: An investigation of the transferability of a multiple-trait scoring instrument across ESL writing assessment contexts. *Language Learning*, *41* (3), 337–373.

Huang, J. (2008). How accurate are ESL students' holistic writing scores on large-scale assessments?—A generalizability theory approach. *Assessing Writing, 13,* 201-218.

Huang, J. (2009). Factors affecting the assessment of ESL students' writing. *International Journal of Applied Educational Studies, 5*(1), 1-17.

Huang, J. and Foote, C. J. (2010). Grading Between the Lines: What Really Impacts Professors' Holistic Evaluation of ESL Graduate Student Writing?. *Language Assessment Quarterly*, 7:3, 219-233.

Huang, J. (2012). Using generalizability theory to examine the accuracy and validity of large-scale ESL writing assessment. *Assessing Writing*, 17, 123–139.

Huot, B. (1993). The influence of holistic scoring procedures on reading and rating students essays. In M. Williamson & B. Huot (Eds.), *Validating holisticscoring for writing assessment* (pp. 206–236). Cresskill, NJ: Hampton Press.

Jacobs, H. L., Zinkgraf, S. A., Wormouth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). *Testing ESL composition: a practical approach.* Rowely, MA: Newbury House.

Janssen, G., Meier, V., & Trace, J. (2015). Building a better rubric: Mixed methods rubric revision. *Assessing Writing*. DOI:10.1016/j.asw.2015.07.002

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

Johnson, R. L., Penny, J., Gordon, B., Shumate, S. R., & Fisher, S. P. (2005). Resolving score differences in the rating of writing samples: Does discussion improve the accuracy of scores? *Language Assessment Quarterly, 2*(2), 117-146.

Kobayashi, T. (1992). Native and nonnative reactions to ESL compositions. *TESOL Quarterly*, 26(1), 81-111.

Lee, Y.W., Gentile, C., & Kantor, R. (2010). Toward automated multi-trait scoring of essays: Investigating links among holistic, analytic, and text feature scores. *Applied Linguistics*, 31(3), 391– 417.

Li, H & He, L. (2015). A Comparison of EFL Raters' Essay-Rating Processes across Two Types of Rating Scales. *Language Assessment Quarterly*, 12: 178–212.

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

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

Lynch, B. K. & McNamara, T. F. (1998). Using G-Theory and many facet Rasch measurements in the development of performance assessments of ESL speaking skills of immigrants, *Language Testing,* 15(2), 158-188.

Maftoon, P. & Akef, K. (2010). Developing rating scale descriptors for assessing the stages of writing process: The constructs underlying students' writing performances. *Journal of language and translation*, volume 1, number1, pp. 1-18.

Matsumoto, K. (1993). Verbal-report data and introspective methods in second language research. *RELC Journal*, 24(1), 32–60.

McMillan, J. H., & Schumacher, S. (2001). *Research in education: A conceptual introduction* (5th ed.). New York: Longman.

Mendelsohn, D., & Cumming, A. (1987). Professors' ratings of language use and rhetorical organization in ESL compositions. *TESL Canada Journal, 5*, 9-26.

Milanovic, M., Saville, N. & Shuhong, S. (1996*). A study of the decision-making behavior of composition markers.* In M. Milanovic & N. Saville (Eds.), Performance testing, cognition and assessment. Selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem (pp. 92-1 1 1). Cambridge: Cambridge University Press.

Parker, C. E., Louie, J., & O'Dwyer, L. (2009). *New measures of English language proficiency and their relationship to performance on large-scale content assessments.* (Issues & Answers Report,REL 2009–No. 066). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast and Islands. Retrieved from http://ies.ed.gov/ncee/edlabs.

Russo, J.E., Johnson E.J. and Stephens D.L. (1989). The validity of verbal protocols. *Memory and Cognition.* 17, 759–769.

Sakyi, A. A. (2000). Validation of holistic scoring for ESL writing assessment: How raters evaluate composition. In M. Milanovic, & A. J. Kunnan (Eds.), *Fairness and validation in language assessment*: Selected papers from the 19th Language Testing Research Colloquium (pp. 129–152). Cambridge: Cambridge University Press.

Sakyi, A. A. (2003). *A study of the holistic scoring behaviors of experienced and novice ESL instructors. Unpublished doctoral dissertation*, University of Toronto, Toronto, Canada.

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

Sasaki, T. (2003). Recipient orientation in verbal report protocols: Methodological issues in concurrent think-aloud. *Second Language Studies*. 22, 1–54.

Schoonen, R., Vergeer, M., & Eiting, M. (1997). The assessment of writing ability: Expert readers versus lay readers. *Language Testing*, *14*, 157–184.

Smagorinsky, P. (1994). Think-aloud protocol analysis: Beyond the black box. In P Smagorinsky (ed.), *Speaking about writing: Reflections on research methodology* (pp. 3–19). Thousand Oaks, CA: Sage.

Smith, D. (2000). Rater judgments in the direct assessment of competency-based second language writing ability. In Studies in immigrant English language assessment, Vol. 1, ed. G. Brindley, 159–89. Sydney: National Centre for English Language Teaching and Research, Macquarie University.

Song, B., & Caruso, I. (1996). Do English and ESL faculty differ in evaluating the essays of native English-speaking and ESL students? *Journal of Second Language Writing, 5,* 163-182.

Stratman, J. F., & Hamp-Lyons, L. (1994). Reactivity in concurrent think-aloud protocols: Issues for research. In P. Smagorinsky (Ed.), *Potential problems andproblematic potentials of using talk about writing as data about writing processes*. (pp. 89-114).

Swartz, C.W., Hooper, S.R., Mongomery, J. W., Wakely, M. B., De-Kruif, R.E.L., Reed, M., Brown, T.T., Levine, M.D. and White, K.P. (1999). *Using generalizability theory to estimate the reliability of writing scores derived from holistic and analytical scoring methods*. Educational and Psychological Measurement, 59, 492_506.

Vann, R., Meyer, D., & Lorenz, F. (1984). Error gravity: A study of faculty opinion of ESL errors. *TESOL Quarterly, 18*, 427-440.

Vaughan, C. (1991). Holistic assessment: What goes on in the raters' minds? In L. Hamp-Lyons (Ed.), *Assessingsecond language writing in academic contexts* (pp. 111–126). Norwood, NJ: Ablex.Winke, P. and Lim, H. (2015). *ESL essay raters' cognitive processes in applying the Jacobs et al. rubric: An eye-movement study*. Assessing Writing, 25, 37–53.

Weigle, S.C. (1994). Effects of training on raters of English as a second language compositions. Quantitative and qualitative approaches. Unpublished PhD dissertation, University of California, Los Angeles.

Weigle, S.C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*, 263–287.

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

Weigle, S. C., Boldt, H., & Valsecchi, M. I. (2003*). Effects of task and rater background on the evaluation of ESL student writing: A pilot study.* TESOL Quarterly, *37*, 345-354.

Wolfe, E. W. (2006). Uncovering rater's cognitive processing and focus using think-aloud protocols. *Journal of Writing Assessment, 2*, 37–56**.**

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