UC Davis
Journal of Writing Assessment

Title
Using Appraisal Theory to Understand Rater Values: An Examination of Rater Comments on ESL Test Essays

Permalink
https://escholarship.org/uc/item/1hm9j7pk

Journal
Journal of Writing Assessment, 6(1)

Authors
Hall, Carla
Sheyholislami, Jaffer

Publication Date
2013

Copyright Information
This work is made available under the terms of a Creative Commons Attribution-NonCommercial-NoDerivatives License, availalbe at https://creativecommons.org/licenses/by-nc-nd/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Using Appraisal Theory to Understand Rater Values: An Examination of Rater Comments on ESL Test Essays

by Carla Hall and **Jaffer Sheyholislami**

Abstract

This study is an illustration of the value of appraisal theory in studies of writing assessment. To demonstrate its functionality and value, appraisal theory is used to examine rater variability and scoring criteria through an analysis of the evaluative nature of rater comments. This allows an exploration of the values raters bring to the scoring task of rating second language writing proficiency. The written comments of three raters scoring the same sixteen writing tests were analyzed through appraisal theory and correlated to each test score. The analysis of the comments suggests that textual features external to the scoring rubric influenced raters' scoring decisions. The findings shed light on raters' perception of the construct of "good writing" and show how raters bring their own interpretations to the rating task. The findings also suggest that there may be unidentified shared rater values, as was evidenced when all raters awarded the same score but disagreed on the quality of specific features of a text. These observations may have implications for rater monitoring, rater training, and scoring rubric revision. This study illustrates that appraisal theory may offer a systematic means to analyze rater comments as they relate to the rating process.

Introduction

An important part of building the validation argument for a writing test is the gathering of evidence for the validity and reliability of the scoring processes, including rater reliability. Research exploring inter-rater reliability has focused mainly on rater decision-making processes (Cumming, 1990; Smith, 2000; Sakyi 2001; Cumming, Kantor, & Powers, 2002), rater variability (Vaughn, 1991; Huot, 1993; Weigle, 1994; McNamara, 1996; Schoonen, 2005; Harsch & Rupp, 2011), and scoring criteria (Freedman & Calfee, 1983; Charney, 1984; Pula & Huot, 1993; DeRemer, 1998). None of these studies, however, has employed appraisal theory in their investigation. In the past two decades, appraisal theory has proven quite productive in a wide range of inquiries such as foreign language writing (Abasi, 2013), academic writing (Hood, 2010), media literacy (White, 2006; Iedema, Feez & White, 1994), and in analyzing public discourse (Martin, 1995), political discourse (Miller 2004, 2002), textbooks (Coffin, 1997), literature (Rothery & Stenglin, 2000), and narratives (Macken-Horarik & Martin, 2003). Appraisal theory is a method of analysis that examines how speakers or writers value the entities within the text they produce (Martin & White, 2005).

Informed by appraisal theory (Martin & White, 2005) this study examines rater variability and scoring criteria, and also includes an analysis of the evaluative nature of rater comments in order to explore the beliefs and values raters bring to the scoring task. The written comments of three raters scoring the same sixteen writing tests are analyzed and correlated to the scores of each test. The focus of this research is on the decision-making criteria raters use when scoring a high-stakes writing exam. The main objective here is to demonstrate the functionality and value of appraisal theory in writing assessment studies. Moreover, the findings may have implications for rater monitoring, rater training, and scoring rubric revision. The following are questions this study sought to answer:

1. What do raters value when scoring L2 writing and do these values manifest themselves in the comments they make?
2. What textual features do raters attend to in the scoring process?
3. What attitudes and values do these comments convey with respect to the construct of writing?
4. Which textural features are commented on most when there is agreement and no agreement on scores?

Employing appraisal theory to analyze raters' written comments, this paper illustrates that textual features external to the scoring rubric influenced raters' scoring decisions. The findings shed light on raters' perception of the construct of "good writing" and show how raters bring their own interpretations to the rating task. The findings also suggest that there may be unidentified shared rater values, as was evidenced when all raters awarded the same score but disagreed on the quality of specific features of a text. These observations may have implications for rater monitoring, rater training, and scoring rubric revision. This study illustrates that appraisal theory may offer a systematic means to analyze rater comments as they relate to the rating process.

Context

The CanTEST, which tests approximately 600-800 candidates per year, is a standardized English proficiency test used for university admission and professional licensure in Canada. The writing section of the test is paper-based and requires candidates to produce an essay of approximately 300 words within a time limit of 45 minutes. The test is intended to be a direct measure of written English expression that provides information about a candidate's ability to present and develop ideas in standard academic/professional language. To increase reliability, the elicited samples are evaluated independently by two trained raters according to criteria outlined in the test's writing rubric. The writing rubric consists of five full (and 10 half) bands based on the following four categories of descriptors: 1. Overall effectiveness in conveying message, 2. Accuracy of vocabulary and structures, 3. Range of vocabulary and structures, and 4. Organization and development of topic. Although there are descriptions for each criterion at each band level, raters score each paper holistically. The manual and rater training material clearly state that the criteria should be considered in light of the writer's ability to convey the message. Therefore, overall effectiveness in conveying the message should take priority in the rating process.

During the individual scoring process, raters also complete a *Rater Comment Sheet* for each paper. On these comment sheets, raters record basic information about the paper including the number of the paper (raters do not have access to candidate names), the date, the testing site and topic, the individual scores the rater and co-rater assigned, the final agreed upon score, and whether the paper needs to be assessed by a third rater. On this document, raters are also required to note the salient features of each paper that influence their scoring of that paper. The Comment Sheet is organized with the following seven headings to guide raters:

- Effort to understand?
- Reads like English?
- Org & Devel & Links
- On topic?
- Accuracy: Grammar: # errors, types of errors, effect of errors
- Accuracy: Vocab: wrong forms, spelling, effective use
- Range: Structures and Vocab

The Comment Sheet serves as a record of the rater's justification for the given score. It functions as a reminder of how a rater arrived at a score and this record is used in the discussion with the co-rater when establishing a consensus score. If raters cannot establish a consensus score, a third rater is asked to rate the paper. The individual and final scores recorded on the Comment

Sheet are reviewed after each testing session as part of the ongoing test validation process. This study examines how comments made on the Comment Sheet reflect raters' values and decision-making processes in scoring writing tests.

This paper will first review relevant literature on rater variability and use of rating criteria. It will then provide a brief explanation of appraisal theory and how it will be applied in this context. Finally, it will discuss the analyses of the data and conclude by highlighting the implications resulting from the analysis.

Literature Review

The most common method of rating direct assessments of second language (L2) writing in large-scale composition tests is through a process of holistic scoring or analytic schemes by at least one rater. It is generally acknowledged that with careful rater training and monitoring, this kind of scoring procedure can produce reliable results (McNamara, 1996; Weigle, 1994, 2002). However, these rating processes have been criticized for oversimplifying the constructs they are supposed to represent. As Cumming, Kantor, and Powers (2002) explained,

> Holistic rating scales can conflate many of the complex traits and variables that human judges of students' written composition perceive (such as fine points of discourse coherence, grammar, lexical usage, or presentation of ideas) into a few simple scale points, rendering the meaning or significance of the judges' assessments in a form that many feel is either superficial or difficult to interpret. (p. 68)

As these types of scoring procedures dominate large scale testing, it is necessary not only to explore the decision-making behaviors and values of raters as they relate to the scoring criteria while rating compositions, but also to examine how these values contribute to the definition of the test construct. As will be shown, appraisal theory can be an effective tool to conduct such examinations.

McNamara (1996) stated, "Performance assessment necessarily involves subjective judgments" (p. 117), and this subjectivity often results in variability in scoring, or problems with inter-rater reliability. Raters may vary in scoring for various reasons, such as: overall rater leniency or severity; different rating styles (Charney, 1984); bias towards a certain group of candidates or task type; differences in interpretation of rating criteria; type of rating criteria and scoring procedures (Barkaoui, 2007; Schoonen, 2005); the existence or absence of rater training and the influences of the type of training, (Harsch and Rupp, 2011; Huot, 1993; Knoch, 2010; McNamara, 1996; Vaughn, 1991; Weigle, 1994, 2002); whether raters provide comments on a particular aspects of the text under evaluation (Huot, 1993); and the social context in which the scoring takes place and the perceived importance the scoring consequences have for all stakeholders (Baker, 2010).

Investigations into the rating process and how raters apply scoring criteria reveal that judging papers is an iterative process (Freedman & Calfee, 1983) involving self-monitoring (Cumming, Kantor & Powers, 2002). Raters also engage in extensive problem-solving when making scoring decisions as opposed to simply matching rating criteria to aspects of test papers (Cumming, 1990; DeRemer, 1998). They generally focus on distinguishing features of texts, consider text audience and task requirements (Freedman & Calfee, 1983), and assign more weight to different textual features (Eckes, 2008; Smith, 2000; Vaughn, 1991). Little attention has been paid to the nature of these decisions and the significance this self-monitoring and self-feedback, which is often manifested in written comments, has in the raters' scoring outcomes.

As Weigle stated (1998), "It is not enough to be able to assign a more accurate number to examinee performances unless we can be sure that the number represents a more accurate definition of the ability being tested" (p. 281). Of central importance in test validity, then, is construct validity or how the construct is defined and how the operationalized definition is assessed. As McNamara (2006) argued,

> both the construct (what we believe 'coping communicatively' in relevant settings to mean) and our view of the individual's standing are matters of belief and opinion, and each must be supported with reasoning and evidence before a defensible decision about the individual can be made. (p. 16).

In other words, it is crucial in the validation process to amass evidence that supports a well-defined construct. Raters contribute to the definition of the construct of a test in that they interpret the rating criteria, which in many tests serves as the most explicit definition of the construct. From their position as judges, raters speak as authorities, and their perceptions of language and their biases serve to reinforce the values represented in a test. As it will be shown in this paper, appraisal theory can be very constructive and informative to identify and map out such values and biases.

Appraisal theory has been chosen as analytical framework to analyze the comments made by the raters because it not only allows for the quantification of data, but also provides a way to interpret the social phenomena that they represent. Appraisal theory stems from work in systemic functional linguistics (SFL), a theory and methodology that views language as a social action and a semiotic system serving to encode and reinforce ideological positions (Eggins, 1994/2004). From this perspective, the function of language is to make meaning, and meanings are always influenced by the social context in which they are exchanged. Language is part of a dialectal process in which it informs and is informed by the values of a particular culture. In this analysis, raters' comments presumably can be interpreted as not only reflecting what an institution values in writing, but also as contributing to and reinforcing those values. Rater comments serve as a window into how the test defines the construct of good writing.

In an SLF model of language analysis, all communication can be examined by mapping the interleaving ideational, interpersonal, and textual meaning in texts (Halliday, as cited in Eggins, 1994/2004). Appraisal is concerned with the interpersonal metafunction of language, which deals with role relationships and attitudes. Appraisal works within the realm of discourse semantics, or meaning beyond the clause, and it views the act of evaluating as the expression of the writer's values as to "what counts as good or bad, what should or should not happen, what counts as true or untrue" (Hunston & Thompson, 2003, p. 8). Specifically, it looks at how texts establish, amplify, target, and source evaluation to enact power and create solidarity (Martin & White, 2005). Analyses using an appraisal theory framework allow researchers to reveal the underlying ideological assumptions of writers and texts through the systematic close reading of texts.

As Hunston and Thompson (2003) explain, evaluation includes both a conceptual and linguistic component. Conceptually, evaluation is comparative, subjective and value-laden. Linguistically, evaluation can be identified through lexis, grammar, and recurring patterns of these features in texts. In this study, the analysis focuses on how raters subjectively compared texts to their own definition of good writing. The analysis identifies those linguistic features of comments that convey approval or disapproval of the texts the raters scored. This study uses Martin and White's (2005) appraisal system network, which sees the act of appraisal as featuring the three components of source, attitude, and amplification. The source of the comments is clearly the raters working within a large-scale testing context. In terms of attitude, this study focuses on appreciation: What the raters value in writing as manifested through the language of their comments, and judgment; institutionalized feelings or "norms about how products,

performances and naturally occurring phenomenon are valued" (Martin, as cited in Harvey, 2004, p. 254). It should be noted, however, that appreciation and judgment at times overlap. Whereas appreciation deals with the appraisal of things, judgment is concerned with the behavior of people. When analyzing the raters' comments about writing, it is sometimes difficult to distinguish between comments that are being strictly addressed to the properties of the text and those that are a reflection of the raters' attitudes toward the writers themselves.

The third component of the appraisal system network, amplification, refers to the force and focus of attitudes of the comments. Force is realized through linguistic choices such as intensifiers and attitudinal lexis that indicate how strongly a rater feels about a judgment he or she is making. For example, a rater may intensify her comment by adding an adjective: "Excellent paragraphing." In the context of this study, attitudes were also identified as intensified by the use of non-linguistic means such as underlining comments (e.g., Reads like English?: "No"), capitalization (e.g., "spelling BIGGEST PROBLEM"), or by the use of symbols (e.g., "good range"). Focus is the sharpening or softening of an experiential category that is not normally gradable.

Method

The theoretical assumption in this study is that rater comments reveal the features of texts that raters value. The most commonly used method of gathering rater comments is through think-aloud protocols (Lumley, 2006). While think-aloud protocols can be a valuable tool in data collection of this type, they can also be time-consuming and expensive to administer, transcribe, and analyze. The data collection technique used in this study may offer a more efficient and accessible method of gathering evidence for test administrators searching for information about the values their raters hold.

Data Collection

The written comments of three raters who scored the same 16 papers were gathered. Raters 1 and 2 scored the papers in an official testing context. Each rater individually blind-scored the papers and made written comments during this phase of the scoring procedure. Raters 1 and 2 then discussed the papers, compared scores, and reached a consensus for a final score. In an actual testing context, a third rater is asked to re-rate papers when there is disagreement; therefore, a third, more senior rater was included in this study to simulate actual rating procedures as well as to obtain more data. Rater 3 also blind-scored the papers, but was not involved in determining a final score. Experience rating CanTEST writing varied among raters, with Rater 1 having two years experience, Rater 2 having seven years experience, and Rater 3 having 10 years experience. Both Raters 1 and 3 had training in rating another international high-stakes second language writing test.

Units of Analysis

Comments were identified at the semantic level; that is, any word, group of words, or symbol (e.g., !,) that represented an evaluation of the text being rated was considered a comment. Three different analyses were conducted. First, a simple sentiment analysis was conducted where all comments for all papers were coded as either generally positive or negative (Appendix). This was done by using Martin and White's (2005) concepts of inscribing or invoking of ideational meanings. Inscribed evaluations are those that are directly stated, such as in the comment "grammar - extremely weak." Invoked evaluations, on the other hand, are indirect or implicit comments about the value of an entity. For example, the comment "absent articles" is not inherently negative; however, within the context of grading papers it is reasonable to assume that it is a quality that the rater considers to be negative. The comments were then organized under the eight predetermined categories outlined in the Comment Sheet. Emerging categories were also identified.

Second, more in-depth analyses of individual papers were conducted by examining the conceptual and linguistic aspects of the raters' comments. Finally, comments unique to certain papers and idiosyncratic comments of raters were analyzed. In the second and third analyses, comments were not only classified as positive or negative, but were also considered in terms of their force. This was achieved by examining the linguistic features of comments that conveyed the intensity of the attitude of the rater.

Findings and Discussion

This section will begin by briefly outlining the general trends revealed in the analysis. It will then go on to analyze raters' comments in relation to scoring on individual papers, first in instances where raters agree on a final score, and then where raters disagree on the final score. Finally, comments and rating behavior unique to individual raters and papers will be examined.

For the 16 papers, the raters made a total of 567 comments in the predetermined categories on the Comment Sheet. Three other categories of comments emerged. First, all three raters made comments on the length of some of the papers; these comments were coded as either positive or negative. Second, two of the raters commented that some writers had taken risks; it is unclear whether this type of comment is negative or positive. Therefore, they were included in only the total number of comments and not categorized as positive or negative. Finally, one rater also included a numbering system for each paper scored. In this system, he numbered each of the four areas of the scoring rubric and wrote a score under each category.

Negative comments outnumbered positive comments by almost three-fold. Rater 2 was the most prolific commentator, making well over twice as many comments as the other two raters. Accuracy in grammar and vocabulary were the most commented on categories, while the category of whether the paper was on topic was the least commented on. Raters also made individual comments unique to specific papers and unrelated to the predetermined categories or emergent categories.

To identify patterns of commenting amongst the three rates, the data were first analyzed by reviewing all the comments written by all the raters. Next, a more in depth analysis of comments for specific papers was conducted to determine how comments related to papers where there was agreement or disagreement among the raters (Individual Paper Analysis). Finally, comments that did not relate to the rating criteria and that were specific to raters or papers were analyzed (Comments and Behaviors Unique to Individual Raters).

Several patterns surfaced as data from all raters and papers were examined. The most salient pattern that emerged in the analysis is that negative comments outnumbered positive comments, which may be explained by a number of factors. First, as Hunston and Thomson (2003) stressed, the very act of evaluation consists of comparing an entity with a norm or standard. This requires that anything not meeting the standard is of lesser quality. Rating papers against a standard aligns with this process; it is integral to the process of rating to look for deviations from the norm. Another reason why negative comments predominate could be because the categories in the Comment Sheet itself are worded negatively:
Accuracy: Grammar: #errors, types of errors, effect of errors
Accuracy: Vocab: wrong forms, spelling, effective use

The repetition of the word "errors" and the inclusion of the word "wrong" invite the rater to comment negatively. In addition, the scoring rubric, the major point of reference for raters, also contains negative language even at higher levels. For example, one of the descriptors under the category of range of structures and vocabulary for a band score that meets the cut score for unconditional university admission reads: "occasional inappropriacies or misuse of idioms."

With respect to scoring, negative comments may have been more influential than positive comments in final score decisions. Of the 16 papers, there were six papers where Raters 1 and 2 disagreed on their individual assessments by half a band level. For five out of six of these papers, the raters settled on the lower score given by one of the raters. Where the raters disagreed by one band level, the final rating was most often an average between the scores; rarely did the score go up to the higher original score given by one of the raters. In the one case where it did go up half a band score, the impact of the score is negligible because the original lower score awarded is the cut score for unconditional admission to university and for most professional licensure.

Although the raters did comment on all areas of the Comment Sheet, by far the highest number of comments was on accuracy in grammar and vocabulary, with negative comments outnumbering positive comments. Again, the wording on the Comment Sheet may have affected the number of comments in each area. For example, the categories of "Effort to Understand?," "Reads like English?," and "On-Topic?" require only a yes or no comment, whereas the wording of the other categories prompt the rater to comment in more detail and on more areas: Accuracy: Grammar: #errors, types of errors, effect of errors; Accuracy: Vocab: wrong forms, spelling, effective use. It may also be that aspects of a text having to do with these areas are more easily identified and therefore more easily commented on. Features of syntax and vocabulary are identifiable at the sentence level, so the rater can easily note these as she or he reads. The other areas of the Comment Sheet less frequently commented on require the rater to take into account larger pieces of text, and thus require more reflection and possibly a second or third read through. Therefore, reading style, as explained by Charney (1984) and Milanovic et al. (1996), may also affect what raters attend to, what gets commented on, and how the paper is final scored.

Negative commenting is related to low scores. For papers below the cut score for university admission, the ratio between negative and positive commenting is between 2:1 and 3:1 or higher. For papers scoring at or above the cut score, the ratio is between 2:1 and 1:3. This suggests some consistency among raters in their general impressions about the quality of the texts they are scoring; they clearly saw more negative textual features in the weaker papers than in the stronger papers.

Individual Paper Analysis

This section presents a more in depth examination of four papers. Specifically, it explores the relationship between raters' comments and papers where there is agreement among raters in scores and papers where there is disagreement in scores.

Agreement on low scoring paper: Paper 31.

All three raters agreed on the final score of Paper 31. Although Rater 1 initially awarded the paper a half band below the other raters' scores, the difference between these scores is very close. More importantly, the consequences for the candidate do not change with the half band difference in scores. Two of the three raters negatively commented on the amount of effort the paper requires to read and all the raters agreed that the paper did not read like English, with one rater adding, "No, not at all because of spelling." Spelling is also mentioned by all three raters under the category of *Accuracy: Vocab: wrong forms, spelling, effective use*.

Rater 1 circles the descriptor "spelling" and gives examples of the spelling errors: "para 1 - technolog, theisemochly?"
Rater 2: circles the descriptor "spelling" and writes the following comments and gives examples of the spelling errors: "extremely weak - phonetic @ best! Finanshly, emochly"
Rater 3: circles the descriptor of spelling and indicates that the spelling errors are so severe that the spell-check function in a word processing program could not fix them: "hinders understanding - even computer couldn't fix. BIGGEST PROB."

Spelling seems to be a determining factor in the score of this paper. This is evidenced in the amplifications (in bold) of "**extremely** weak," "**@ best!**,", and "**even** computer couldn't fix. **BIGGEST PROB**."

Though spelling is not directly addressed in the rubric, *effort to understand* the text is. Therefore, raters could have applied the descriptors related to effort to understand to the weak spelling they identified. As DeRemer (1998) suggested, raters do not simply match descriptors with features of text, but bring their problem-solving and interpretive skills to the task. Spelling is often considered a simple literacy problem which can be easily addressed with the spellcheck function of a word processing program; however, the examples above illustrate that at a certain point the seriousness of spelling errors cross a line where they may impede comprehension to such an extent that the errors can be considered a limitation in proficiency.

All three raters also cite a lack of accuracy in grammar and vocabulary as contributing to the weakness of the paper. In particular, they agree that the number of errors, the type of errors, and the incorrect use of word forms are problematic. Interestingly, all three raters commented positively on the text's organization, development, and links:

Rater 1: underlined Links on the Comment Sheet and comments, "mostly ok"
Rater 2: comments, "excellent 5 p. essay set up in intro, 3 reasons and conclusion, good job"
Rater 3: circled Links on the Comment Sheet and comments, "first, second, third, OK. Development is OK."

Although all three raters commented positively on this area of the text, there is quite a difference in the degree of acceptability. Rater 1's use of "mostly" weakens the comment, and Rater 3's comment of "OK" is a simple acknowledgement of the use of the transitions "first, second, third," which is in contrast to Rater 2's use of the intensifier, "excellent," and attitudinal lexis, "good." This difference in amplification indicates a difference in standards. What is also notable is the fact that the comments in this area are all positive but the paper received a relatively low score from all three raters. In discussions with the test administrator, it was revealed that the raters are trained to value language proficiency over content, thus explaining the low score the paper received despite the positive comments on the organization and the development of essay. In addition, Cumming et al. (2002) speculated that L2 writers may need to attain a certain threshold in their L2 writing abilities before "assessors can attend thoroughly and sincerely to their ideas and rhetorical abilities in written compositions" (p.89). A paper receiving such a low score would likely be below this threshold.

Two other notable comments are made on this paper. Rater 1 commented on the writer as "- a risk taker," and Rater 2 commented that the text is "very short." It is unclear whether being a risk-taker is positive or negative but the fact that raters make repeated comments about taking risks in the 16 papers indicates that it is a feature that they can identify and attend to. The comment, "very short," is clearly negative and was likely a contributing factor to the final score.

Agreement on high scoring paper: Paper 26.

All three raters awarded Paper 26 a score that is normally considered sufficient for conditional admission into many university programs. A candidate whose paper is awarded the cut score will be admitted to a program of study but may also be required to take and succeed in one or two writing courses depending on the scores from the other components of the test and departmental requirements. There were almost an equal number of positive and negative comments made on this paper. All but one of the comments on the four categories of *Effort to understand?*, *Reads like English?*, *Organization*, *Development and Links*, and *On-topic?* were positive. The sole negative comment was by Rater 1, in the area of *Organization*, *Development and Links*: "a bit disjointed." The qualifier, "a bit," mitigates the severity of the comment, however.

The category of accuracy in grammar received the most negative comments, including comments on errors in subject-verb agreement, collocations, and word forms. There are two clear areas of disagreement within these categories, however. First, two raters commented on article use:

Rater 2: "articles: strong"
Rater 3: "absent articles"

Rater 2 noted that the writer's use of articles was strong while Rater 3 noted their absence. What is not evident in these comments is whether this textual feature is important. Many raters realize that articles are often one of the last linguistic features to be acquired in English, even by advanced learners. The fact that this is the only negative comment that Rater 3 wrote may indicate that the writer's accuracy is quite good. Rater 2 also praises the writer's use of prepositions ("prepositions: strong") and indicates that other errors in accuracy, aside from sentence structure, are "minor."

The second area of disagreement was on sentence structure. All three raters commented on sentence structure, but disagreed on the writer's abilities in this area:

Rater 1: "para. 2 - complex sentences confusing. -Detracts from completion of task, some complex sentences, but errors detract from understanding, comma splice- last paragraph, starts sentences with 'conj'"
Rater 2: "not great complex structures. Keeps to short structures, errors within simple sentences"
Rater 3: "good complex sentences"

Regardless of the evaluation of the writer's ability to produce complex sentences, all raters awarded this paper the same score. How the quality of this textual feature influences rating is therefore unclear. The rating criteria outline the degree of accuracy and control over simple and complex structures a paper must demonstrate to be awarded the cut score level. According to the comments by Raters 1 and 2, the text does not meet the rating criteria; according to Rater 3, it exceeds it.

These differences in interpretation may be the result of raters focusing on textual features in different parts of an uneven text. They may also once again be the result of differing standards among the raters. The fact that despite these differences the raters all assigned the same score suggests that this paper displays some overall qualities that all raters recognize as equaling the cut score.

Another area all raters comment on is spelling:
Rater 1: "sp"
Rater 2: "spelling - some errors, but not too distracting: technologie, dayly"
Rater 3: "sp many mistakes but can be understood"

Spelling, therefore, is again a feature that raters attended to, but in contrast to the weak paper, Raters 2 and 3 conceded that the errors in this paper are not too serious. The distinguishing factor between the spelling errors in this paper and the weak paper is that the errors in this text were not as distracting and did not affect the comprehensibility of the text. Again, no descriptor in the rubric addresses spelling, but the descriptor for the cut score includes comments on accuracy and whether the meaning of a text is obscured by errors. The raters appeared to be applying the more general descriptor of accuracy to the issue of spelling.

Only Rater 3 commented on the length of the paper: "SHORT." The intensification of the all-caps indicates that this is a fairly serious issue. It did not, however, seem to affect the candidate's score.

To summarize, the analysis of these two papers where there was initial agreement on the final scores reveals that raters do attend to different features of text, which supports similar findings by Vaughn (1991), Smith (2000), Cumming, (2002), and Sakyi (2001). It also reveals that raters may respond in similar ways to particularly salient textual features, such as the poor spelling and the amount of effort that is required to understand the text in Paper 31, and the organizational quality of Paper 26. In addition, the analysis shows that although the raters chose to comment on the same textual features, some of their interpretations of these features, as illustrated in the comments about articles in Paper 31 and complex sentences in Paper 26, are quite different. What is interesting is that despite these differences in interpretation, the raters awarded the same scores. This may indicate that the raters are rating holistically; that there is some sense of shared values about the overall qualities of these texts identifiable to each rater. One might suggest that raters start with a holistic appraisal and then find features and levels of proficiency that support their initial, non-analytical judgment. Such a possibility can be investigated using appraisal theory as long as another set of complementary data based on interviewing raters prior to carrying out their actual task is obtained.

Also observable through this analysis is the use of criteria not articulated in the scoring rubric, such as spelling and text length. This rater behavior could be seen as a result of a weakness of this specific scoring rubric and scoring rubrics in general. However, both spelling and text length are addressed in the training manual of the test. As noted by Cumming (1990), it is virtually impossible to capture all the complexities of language in scoring rubrics, which are usually fairly short documents in order to make them manageable. In terms of spelling, the raters appear to have related this textual feature to the scoring descriptors. This corresponds to DeRemer's (1998) theory that raters use problem-solving techniques in their interpretation of scoring criteria. It also supports Vaughn's (1991) suggestion that raters use personal reading strategies in scoring decisions when essays do not clearly fit with a rating scale.

Disagreement on paper below and at cut score: Paper 39.

The variation in the scoring of Paper 39 is striking. Rater 1 scored this paper two bands below the cut score, Rater 3 scored it one band below, and Rater 2 scored it at the cut score. A disagreement of two full band levels is problematic, particularly when the variations straddle a cut score for university admission or for licensure. Negative comments far outweighed positive comments for this paper. All raters agreed that effort is required to understand the text and that it did not read like English:

Effort to understand?:
Rater 1: "Hard"
Rater 2: "a bit. Spelling distracting."
Rater 3: "Some"

Reads like English?:
Rater 1: "NO"
Rater 2: "not quite"
Rater 3: "mostly no"
Rater 1's certainty is reflected in the evaluative lexical choice of "hard," underlining the word, and capitalizing the word "no." His fairly harsh criticism was also reflected in the low score he awarded this text. Raters 2 and 3 also made negative comments with regard to the writer's ability to produce text that is easy to understand and reads like English, but their comments are much less harsh, as are their ratings. This disagreement in both comments and scores indicates that there is considerable variability in the interpretation of these two categories.

All three raters made positive comments on this paper as well, with some overlap in categories:

Org & Devel & Links:
Rater 1: "tries to use examples/illustrations"
Rater 2: "good intro. Good para breaks, missing essay P links, good sent. links, good examples"

Range:
Rater 1: "limited"
Rater 2: "modals, good vocab range"
Rater 3: "some good chunks- few and far between"

The important difference in these comments is that Rater 1 qualified the comment about examples with the evaluative, "tries," whereas Rater 2 saw the use of examples as successful or "good." In terms of range, we see a graduation from good to bad evaluations among the raters - again signaling variability in the interpretation of this feature.

All three raters commented negatively on the categories of accuracy in grammar and vocabulary, with Rater 2, the rater awarding the highest score, also making the highest number of negative comments. Clearly then, this is an area where there is some agreement among raters, but it is not reflected in the scores they award. This suggests that different raters may place greater weight on different features of a text (Milanovic et al., 1996). For example, in addition to making the highest number of negative comments about grammatical accuracy, Rater 2 made a relatively high number of positive comments (4) about the organization and development of the text. The high score Rater 2 awarded the paper, therefore, may be a reflection of the greater weight placed on the sophistication of the development and organization of the text than on the linguistic proficiency to express it.

Disagreement on paper below and on cut score: Paper 38.

Paper 28 represents another text for which there is a full band level of variation in scores. Rater 1 scored the essay one band level below the cut score for conditional admission to university, Rater 3 scored it a half band below, and Rater 3 2 scored the paper at the cut score. All raters negatively comment on the effort required to read the text:

Effort to understand:
Rater 1: "some"
Rater 2: "a bit. Spelling is distracting + wrong forms, too"
Rater 3: "some"

Rater 2 gave the most detailed and strongest criticism in this category, but rated the paper the highest. This inconsistency is repeated in the category of *Read like English?* and *Org & Devel & Links*:

Read like English?
Rater 1: "A bit"
Rater 2: "No - too wordy and awkward"
Rater 3: "somewhat uneven - many very fluent sections"

Org & Develop &Links:
Rater 1: "can be followed"
Rater 2: "good para breaks but no essay links"
Rater 3 circled Links and commented: "GOOD, flows well despite spelling and word form"

With regard to the organization of the paper, Rater 2 made the clearest negative comment, "no essay links"; Rater 1 conceded that the paper "can be followed," and Rater 3 offered the most praise for this aspect of the paper. Perhaps the most telling comment that may explain this inconsistency is Rater 3's comment "uneven." The raters may be responding to the possibility that the writer's interlanguage has developed to a degree where he or she is able to convey some thoughts with precision and ease, but not all. Rater 1 also commented that the writer is a risk-taker. Although one cannot discern whether this comment is negative or positive, it indicates that the raters feel the writer may be overextending his or her capabilities.

The unevenness of the paper could be a problematic feature of this text. Raters may not know which parts of the text are representative of the true capabilities of the candidate. One rater may be responding to the strongest sections while the other raters may be responding to the weakest sections. This provides support for and extends Vaughn's (1991) and Smith's (2002) findings that raters attend to different text features to arrive at their scores. In the analysis of rater comments, we can see that raters not only attend to different features, but also evaluate features of texts differently.

The problem of scoring uneven texts may be related to the assessment technique of holistic rating. Although raters are guided to focus on specific textual features by the descriptors in the rubric, they must produce a single score based on an impressionistic reading of the paper. Holistic scoring, however, may be inadequate to deal with papers that exhibit uneven proficiency in textual features (Hamp-Lyons, 1995; Johnson & Hamp-Lyons, 1995; Haswell, 1991). For example, a writer may display a high level of quality in the trait of grammatical range but not in the trait of development and organization of the text. Holistic scoring of such papers "reduces the writer's cognitively and linguistically complex responses to a single score" (Hamp-Lyons, 1991, p. 244), thus providing less useful information to all test stakeholders than other assessment techniques. Hamp-Lyons argues that multi-trait

scoring, which can provide both individual trait scores as well as composite scores, gives a clearer picture of a writer's strengths and weakness, and thus may provide useful feedback for writers and diagnostic information for administrators and teachers. Multi-trait scoring does not, however, address inconsistencies in proficiency levels within a single trait. For example, some papers may exhibit strengths in lexical range in parts of the text, but may break down in other parts. In cases such as these, raters must decide on which part of the paper to score - the weakest, the strongest, or somewhere in between.

The idea of risk-taking is also interesting in that the raters may recognize that the writer could have produced more accurate writing if he or she had kept to simpler language. By trying to write in more complicated language, the candidate may have unintentionally penalized him or herself. The fact that the raters recognize this feature of writing indicates that it is an area that may need clarification in training in terms of how the rubric deals with it and whether or not it should be rewarded or penalized. It is also interesting to note that the two raters who comment on risk-taking have also been trained in another test that uses this term in its rating criteria. It is conceivable, therefore, that some transfer in training and scoring is taking place.

The analysis of the two papers where raters disagreed on their initial scores reveals some similar patterns of rater behavior when there is agreement in scores. First, in both cases, the raters did attend to different textual features to arrive at their scores. In addition, they interpreted some textual features quite differently; however, in these cases, the differences appear to have affected the scores. Compared to the differences in interpretation where there was rater agreement, the comments here vary much more not only in degree (e.g., all raters agree that Paper 39 is difficult to understand, but to varying degrees), but also in absolutes (e.g., one rater sees Paper 39 as having "good range," while another sees it as being "limited"). The subjective nature of the rating process is evidenced in the range of these comments. Clearly, the raters have made different language choices to describe what constitutes good and bad writing and this suggests that a standardization session may be required.

Comments and Behavior Unique to Individual Raters and Papers

Comments unique to specific raters and papers may also shed light on the causes of variability among raters. An analysis of the comments indicates that individual raters may be responding to specific criteria not addressed in the scoring rubric or they may be applying the criteria in a unique way. For example, on each Comment Sheet, Rater 1 included scoring notes to illustrate the procedure used to arrive at a score. He wrote the numbers 1-5 and awarded a score under each number:

Example:

Presumably numbers 1-4 refer to the different categories of the scoring rubric and number 5 is an average of the scores awarded. The writing test, however, uses multi-trait holistic scoring. In other words, although the rubric breaks down the construct into four different features, raters are supposed to arrive at a final score holistically. As this rater has been trained in other tests, some of which use analytic scoring procedures, he may be applying those same procedures here. Weigle's (1994) research indicates that training in rating criteria improves the reliability of interpretation and application of the criteria. How training in another test affects the interpretation of a scoring rubric is unclear, however. Rater 1 is consistently more severe than Raters 2 and 3, and it would be interesting to see if and how training in other tests influenced his scoring on this test.

Rater 3 made two comments on two different papers that reveal the use of rating criteria external to the grid. For Paper 40 she awarded a score much lower than the final score awarded by the other two raters. The majority of Rater 3's comments on this paper did not relate to the scoring rubric. Rather, they reflected suspicions about the candidate simply using memorized chunks of formulaic language; then provide examples of this type of language:

Rater 3:
hard to tell what he can do outside of the template...TOEFL TEMPLATE: "evidence reveals..." "there is no doubt...", memorized test prep chunks in intro: "has sparked much debate..."

In essence, the rater is questioning the true ability of the writer - something she feels this performance has not captured.

Rater 2 is the most prolific commentator. She also uses the most evaluative language. Although Raters 1 and 3 do use terms such as "good and "weak," their comments often consisted of simply documenting examples of the writers' strengths and weaknesses they believe are representative of a particular feature of a text. Rater 2, on the other hand, consistently used dramatic amplifications in her comments. These amplifications include lexical choices, (e.g., "throws around vocab - often wrong forms"), capitalization (e.g., "WEAK"), expressive punctuation (e.g., "!!!!"), symbols (e.g., √), and circling her comments. She even offered advice through modality: "extra long. Should write less, more carefully" (comment is circled). The intensity and the number of comments this rater makes may indicate that she is reading the papers more reflectively than the other raters and that her reading style or strategy may affect what she attends to and how she scores each paper. The fact that Rater 2 consistently awarded higher scores than the other raters may be the result, at least in part, of the approach she takes when reading test papers. As Charney (1994) pointed out, rating style is related to intra-rater reliability. Although intra-rater reliability was not analyzed in this study, it is conceivable that different reading styles among raters may also affect *inter*-rater reliability.

Conclusion

Appraisal theory provides a systematic framework to analyze the use of evaluative language in written and spoken texts. The objective of this study was to explore the merits and limitations of appraisal theory to build on previous research on rater behaviors and values in ESL writing assessment illustrate the usefulness of employing appraisal theory in ESL writing assessment investigations. Using this theory, the paper analyzed rater comments to build on previous research of rater behaviors and values. An analysis of this type using an appraisal theory framework does provide information on the textual features that raters attend to during the scoring process. First, it can identify which features raters notice; one can assume that in order to comment on a textual feature a rater has noticed that feature. Second, it can identify the relative importance of a particular textual feature by the number of comments it receives and how strong the comments are. Third, this type of analysis may be especially useful when there is disagreement among raters. As was evident in the analysis of papers where raters disagreed, raters viewed some of the same aspects of a text differently. Once identified, these differences can be addressed through rater training and scoring rubric revision. Finally, this type of analysis can reveal areas outside of the scoring rubric that raters attend to. The categories that emerged as well as idiosyncratic comments raters made illustrate that textual features external to the Comment Sheet and the scoring rubric influenced raters' scoring decisions.

This type of analysis also sheds light on raters' perception of the construct of "good writing." As was illustrated, both the Comment Sheet and the scoring rubric inform, define, and limit the construct. Both these documents try to capture the complex task of writing in texts that are meaningful and useful to the raters. The raters both work within this framework and reinforce it. An analysis of the

scoring rubric, the Comment Sheet, and the raters' comments reveal that one of the most valued features of academic writing in this context is accuracy; it is clearly rewarded while errors are punished. The analysis also shows that rhetorical features of texts are valued, but not as much as other features, particularly for lower proficiency texts. Features external to the rubric, such as appropriate text length and correct spelling, are also valued by raters. There is also the suggestion that at times, there are unidentified and perhaps unidentifiable shared rater values. This is evident when all raters award the same score but disagree on the quality of specific features of a text.

This type of analysis does not, however, tell us precisely how a textual feature affects scoring decisions. The full context in which these comments were made and how they were used to arrive at scores is unknown. Therefore, appraisal analysis of raters' written comments should be complemented by interviews with raters about their decision making in order to shed more light in this area. Nor does this type of analysis take into account other possible factors that contribute to scoring decisions. For example, Rater 2 often deferred to Rater 1's scoring when they were a half band level apart. It is not clear whether Rater 1's documentation of a paper's quality was the determining factor in Rater 2's acquiescence to his initial score. It is just as likely that personality, perceived authority, or something else led to the final scoring decision. An analysis of the negotiations between raters about final scores may reveal more clearly how this process works.

One purpose of examining the factors that contribute to rater reliability is to improve test fairness. It is desirable for all raters to apply the rating criteria consistently and in the same way. Another purpose is to examine how the test construct is being interpreted by raters, and in doing so, more clearly define construct validity. However, as Constable and Andrich (as cited in Lumley and McNamara, 1995) argue, increased reliability can paradoxically decrease the validity of the construct by limiting the definition of the construct through the use of what Charney (1984) and Cumming (2002) describes as ad hoc criteria meaningful only to the specific discourse community of trained test raters. Identifying areas of disagreement among raters and criteria raters use that are not addressed in the scoring rubric may provide opportunities for test developers to re-examine, refine, and expand the construct as articulated through the rating criteria. From this perspective, variability in rating serves a positive function in the on-going test validation process. In the case of this test, it has led to revisions of the rating rubric, training procedures, and training materials. Changes to the writing rubric have included clarifying the descriptors for the category of "Organization and Development of Topic." As the comments in Papers 31 and 38 suggest, raters vary in their judgments of these descriptors both when they agree and disagree on a final score. Subsequent discussions with raters about the original descriptors of this category revealed that some of the phrasing was vague and difficult to interpret. These descriptors were then changed to be clearer and more distinct between levels. The changes will be field-tested and rater feedback will be collected.

In an effort to maintain the manageable size and length of the rating rubric, many of the other issues arising from this analysis will be addressed through training materials and training sessions. Training materials include slides, a guide booklet on the administration and scoring of the CanTEST writing examination, and an annotated exemplar booklet. The concern about text length, particularly short papers, is already addressed in the guide booklet, but it will also be more explicitly discussed in training and in the annotated exemplar booklet. Spelling and risk-taking will be addressed in training and training materials as accuracy issues. Information on uneven texts and memorized chunks of texts will be included in the section of "Problematic Papers" in the guide booklet, and examples will be added to the annotated exemplar booklet.

Author Note

Carla Hall is a language instructor and test development coordinator at the Official Languages and Bilingualism Institute at the University of Ottawa. Her recent publications include articles in L'Association canadienne des professeurs d'immersion and a co-authored chapter in Immersion Education (2011, Multilingual Matters). Her areas of research include interaction in language testing and computer mediated testing.
Contact: cahall@uottawa.ca
70 Laurier, Room 027, Ottawa, ON. K1N 6N5, Canada

Jaffer Sheyholislami is associate professor at the School of Linguistics and Language Studies of Carleton University where he teaches courses in applied linguistics and discourse studies. His recent publications include Kurdish Identity, Discourse, and New Media (2011, Palgrave Macmillan) and several encyclopedic entries, book chapters, and papers in peer-reviewed journals such as Discourse & Society, International Journal of the Sociology of Language, and Language Policy. His areas of research also include Systemic Functional Linguistics, appraisal theory, critical discourse analysis, and educational linguistics.
Contact: jaffer.sheyholislami@carleton.ca
1125 Colonel By Drive, Room PA 236, Ottawa, ON. K1S5B6, Canada

References

Abasi, A. R. (2013). Evaluating choices and rhetorical impact: American learners of Persian as a foreign language writing to appraise. *International Journal of Applied Linguistics 23*(2). DOI:10.1111/ijal.12024.

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

Baker, B. (2010). Playing with the stakes: A consideration of an aspect of the social context of a gatekeeping writing assessment. *Assessing Writing, 15*, 133-153.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing, 12*(2), 86-107.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English, 18*, 65-81.

Coffin, C. (1997). Constructing and giving value to the past: An investigation into second school history. In F. Christie and J. R. Martin (Eds.), Genre and institutions - *Social processes in the workplace and school* (pp. 196-230). London: Cassell.

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

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

DeRemer, M. (1998). Writing assessment: Raters' elaboration of the rating task. *Assessing Writing, 5*, 7-29.

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

Eggins, S. (1994/2004). *Introduction to systemic functional linguistics*. New York: Continuum International Publishing Group.

Freedman, S., & Calfee, R. (1983). Holistic assessment of writing : Experimental design and cognitive theory. In P. Mosenthal, L. Tamor, & S. Walmsley (Eds.), *Research on Writing* (pp.75-98). New York: Longman.

Hamp-Lyons. L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 241-276). Norwood, NJ: Ablex Publishing Corporation.

Harsch, C., & Rupp, A. (2011). Designing and scaling level-specific writing tasks in alignment with the CEFR: A test-centered approach. *Language Assessment Quarterly, 8*(1), 1-33.

Harvey, A. (2004). Charismatic business leader rhetoric: From transaction to transformation. In L. Young & C. Harrison (Eds.), *Systemic functional linguistics and critical discourse analysis - Studies in social change* (pp. 47- 260). New York: Continuum. Retrieved from: http://books.google.ca/books?hl=en&lr=&id=gh1bng-bakQC&oi=fnd&pg=PA247&dq=Appraisal+theory+martin+systemic+functional+linguistics+A+Harvey&ots=Cr6CEes15y&sig=t2eb9vGK0dNfApKM02ocDzG6Bw8&redir_esc=y#v=o

Haswell, R. H. (1991). *Gaining ground in college writing: Tales of development and interpretation*. Dallas, TX: Southern Methodist University Press.

Hood, S. (2010). *Appraising research: Evaluation in academic writing*. New York: Palgrave Macmillan.

Hunston, S., & Thompson, G. (2003). *Evaluation in text: authorial stance and the construction of discourse*. Oxford: Oxford University Press.

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60*, 237-263.

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

Iedema, R., Feez, S., and White, P.R.R. (1994). *Media literacy, Sydney, disadvantaged schools Program*. Sydney: NSW Department of School Education.

Johnson, D. M., & Hamp Lyons, L. (1995). Research on the rating process: Rating nonnative writing: The trouble with holistic scoring. *TESOL Quarterly, 29*(4), 759-762.

Knoch, U. (2010). Investigating the effectiveness of individualized feedback to rating behavior - a longitudinal study. *Language Testing, 28*(2), 179-200.

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

Lumley, T., & McNamara, T.F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*, 54-75.

Macken-Horarik, M., and Martin. J.R. (Eds.) (2003). Text (special issue) - *Negotiating heteroglossia: Social perspectives on evaluation*, Vol. 23. New York: Mouton de Gruyter.

Macksoud, R. (2010). Using interview data in case studies. In S. Hunston and D. Oakey, (Eds.), *Introducing applied linguistics: Concepts and skills* (pp. 151-159). New York: Routledge.

McNamara, T. (1996). *Measuring second language performance*. New York: Addison Wesley.

Martin, J.R. (1995). Interpersonal meaning, persuasion, and public discourse: Packing semiotic punch. *Australian Journal of Linguistics, 15*, 3-67.

Martin, J.R., & Rose, D. (2007). *Working with discourse: Meaning beyond the clause*. New York: Continuum.

Martin, J.R., & White, P.R. (2005). *The language of evaluation: Appraisal in English*. New York: Palgrave MacMillan.

Milanovic, M., Saville, N., & Shuhong, S. (1996). A study of the decision making behavior of composition markers. In M. Milanovic and N. Saville (Eds.), *Studies in language testing 3: Performance testing, cognition and assessment*. Selected papers from the 15th Language Testing Research Colloquium. Cambridge: Cambridge University Press. Retrieved from: http://books.google.ca/books?hl=en&lr=&id=q82vhSnA3jwC&oi=fnd&pg=PA1&dq=performance+testing+Milanovic&ots=lm5gd7m2wA&sig=wd-EJAiPps6B1cPU_PLaOOe5K1I&redir_esc=y#v=onepage&q=performance%20testing%20Milanovic&f=false

Miller, D. (2004). "...to meet our common challenge": ENGAGEMENT strategies of alignment and alienation in current US international discourse . *Intercultural discourse in domain-specific English, Textus, 17*(1), 39-62.

Miller, D. (2002). Multiple judicial opinions as specialized sites of engagement: Conflicting paradigms of valuation and legitimation in Bush v. Gore 2000. In M. Gotti and C. Dossena (Eds.), *Conflict and negotiation in specialized texts* (pp. 119-141). Berlin: Peter Lang.

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

Rothery, J., & Stenglin, M. (2000). Interpreting literature: The role of APPRAISAL. In L. Unsworth (Ed.), *Researching language in schools and functional linguistic perspectives* (pp. 222-244). London: Cassell.

Sakyi, A.A. (2001). Validation of holistic scoring for ESL writing assessment: How raters evaluate compositions. In J.J. Kunnan, (Ed.), *Fairness and validation in language assessment* (pp. 82-96). Cambridge: Cambridge University Press.

Schoonen, R. (2005). Generalizability of writing scores: an application of structural equation modeling. *Language Testing, 22*(1), 1-30.

Shaw, S.D., & Wier, C.J. (2007). *Examining writing: Research and practice in assessing second language writing.* New York: Cambridge University Press.

Shohamy, E. (1998). Critical language testing and beyond. *Studies in Educational Evaluation, 24*(4), 331-345.

Smith, D. (2000). Rater judgments in the direct assessment of competency-based second language writing ability. In G. Brindley (Ed.), *Studies in immigrant English assessment* (pp. 159-189). Sydney, Australia: National Centre for English Language Teaching and Research.

Titscher, S., Meyer, M., Wodak, R., & Vetter, E. (2002). *Methods of Text and Discourse Analysis.* London: SAGE Publications.

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

Weigle, S. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*, 171-184.

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

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

White, P.R.R. (2006). Evaluative semantics and ideological positioning in journalistic discourse. In Lassen, I. (Ed.), *Image and ideology in the mass media* (pp. 45- 73). Amsterdam: John Benjamins.

Appendix

Table A.1: *Summary of Comments in Predetermined Categories*

A= Effort to understand?
B= Reads like English?
C= Org & Devel & Links
D=On Topic
E= Accuracy: grammar: #errors, types of errors, effect of errors*
F=Accuracy: Vocab: wrong forms, spelling, effective use*
G= Range of Structures and Vocab

*Accuracy refers to the grammatical and lexical accuracy and precision of the writing and how it affects the comprehensibility of the text. Raters are trained to evaluate the type and seriousness of the errors candidates make.

CS= cut score for conditional admission
-# = band level below cut score
+#= band level above cut score
FS= Final score awarded to paper

Paper/ Scores/ Total comments - Positive (+) Negative (-)	Category	R1		R2		R3		Total	Total	Total comments all raters
		+	-	+	-	+	-	+	-	
#26	A	1		1		1		3		3
(CS)R1	B	1		1		1		3		3
(CS)R2	C		1	2		1		3	1	4
(CS)R3	D	1		3				4		4
(CS) FS	E		2	2	3		1	2	6	8
+19	F		1		3	1	3	1	7	8
-20	G	1	4	2	2			3	6	9
Total 39										
#2	A			2			1	2	1	3
(-2) R1	B		1	2	2		2	2	5	7
(-.5) R2	C				2		1		3	3
(-1.5)	D	1						1		1
(-1)FS	E		7		4		1		12	12
+5	F		3		3		3		9	9
-34	G				4				4	4
Total 39										
#28	A				1		1		2	2

Table A.2. *Summary of Comments per Pre-determined Category*

Category	**Positive Comments**	**Negative comments**	**Total**
A. Effort to understand	20	38	58
B. Reads like English	24	36	60
C. Org&Devel&Links	35	35	70
D. On-topic?	23	15	38
E. Accuracy: Grammar: #errors, types of errors, effects of errors	10	115	125
F. Accuracy: Vocab: wrong forms, spelling, effective use	8	90	98
G. Range: Structures and Vocab	18	60	78

Table A.3 *Emerging Categories*

Category	**Positive**	**Negative**	**Total**	**Example**
Length	7	6	13	*'good length'* *'Short'*
Risk-taking			5	*'a bit of a risk-taker'*
Rating technique			16	*1 2 3 4 5* *3 3 3.5 3 3*

Copyright © 2021 - ***The Journal of Writing Assessment*** - All Rights Reserved.