When Raters Talk, Rubrics Fall Silent

MASOUMEH AHMADI SHIRAZI
University of Tehran, Iran

Bio Data:
Masoomeh Ahmadi Shirazi received Ph.D. in Applied Linguistics at the University of Tehran in December 2008. She received her MA in TEFL from the same university in May 2003. She is now an assistant professor at the faculty of foreign languages in University of Tehran. She takes interest in academic writing, writing assessment, second language acquisition, vocabulary learning, research methodology and statistics.

Abstract

The research reported here suggests that raters, when involved in writing assessment, are more concerned with their own criteria to set a basis for their judgment rather than the standards provided by scale descriptors. This study sampled think aloud of eight raters who scored 15 essays in accord with Test of Written English (TWE) holistic scoring guide. Verbal report data indicated that just less than five percent of the statements made by the raters are related to the issues assessed in TWE. These findings background the utility of holistic rating scale descriptors, foregrounding the raters' descriptors-independent judgments.

Keywords: raters, rubrics, descriptors, TWE, holistic scales

Introduction

This study attempts to see how much of raters' judgments is the reflection of scoring rubrics and descriptors. In fact, this article focuses on the degree of raters' compliance with the scoring rubric. It is important to consider the role of rubrics and descriptors in writing assessment for they could contribute to higher reliability (Connor-Linton, 1995; DeRemer, 1998).

The study collected raters' judgments on written essays of a number of participants using Verbal Protocol Analysis (VPA) as the main instrument of the research. Raters' verbalizations, then, were transcribed so that we can analyze these texts, looking for the rubrics and descriptors suggested both by the study and what other writing features are introduced by the raters to scoring procedure.

In fact, we aim to illustrate that in spite of the crucial role that scoring guides play in controlling raters' assessment behavior hence increasing reliability (Connor-Linton, 1995; Pollitt & Murray, 1996; DeRemer, 1998; Marby, 1999), they may get marginalized by the raters' own criteria. The flexibility of descriptors and rubrics, as Norton Pierce (1991) has suggested, leaves some rooms for the rater to award the writing with features that are not part of the scoring guide resulting in lower reliability. This is especially true in holistic scoring where raters experience scoring guides which give them a bonus to include writing features in their assessment that are not specified by the scoring guide. Here we consider the raters' idiosyncratic preferences as prior to holistic scoring guides which have a minor role in writing assessment processes.

Review of Related Literature

We can begin this section with the following question: "What are these rubrics and descriptors, which provide the basis for scoring written essays?"

There are different looks on the definitions of descriptors and rubrics in the literature. Davies, Brown, Elder, Hill, Lumley, and McNamara (1999) view descriptor as "a statement which describes the level of performance required of candidates at each point on a proficiency scale" (p. 43); descriptors typically make reference to the level of language skill required (for example, level of grammatical accuracy, vocabulary range), to production skills (for example, asking questions, giving personal information, filling in forms), or to content of the message (the relevance of information, organization of ideas).

Herman, Aschbacher, and Winters (1992) enumerate the following as the elements of a scoring rubric:

✓ One or more traits or dimensions that serve as the basis for judging the student response

✓ Definitions and examples to clarify the meaning of each trait or dimension

✓ A scale of values on which to rate each dimension

✓ Standards of excellence for specified performance levels accompanied by models or examples of each level

Generally, descriptors are statements showing essential qualities of the written performance of learners. Scale descriptors are regarded as necessary since they serve as guides for the raters. Jacobs et al. (1981) view descriptors as important due to the fact that they focus readers' attention on significant aspects of the composition hence more reliable composition evaluation. These descriptors provide raters with a series of descriptions which help determine the level of learners' performance. According to DeRemer (1998), these scoring guides realized in scoring rubrics should be sufficiently specific to enable consistency across raters in categorizing aspects of a piece of writing. There has been so much emphasis upon the significance of rating scales and descriptors in assigning scores to the test takers (Pollitt and Hutchinson, 1087; Underhill, 1987; Upshur and Turner, 1995; North and Schneider, 1998; Turner and Upshur, 2002; Weaver, 2006).

However, rating scales and their accompanying descriptors are still under question. Rating scales are founded on either theory or are empirically driven rating scales. There have been controversy on the basis of rating scales and some scholars have cast doubt on the theory-based rating scales (Chalhoub-Deville, 1997; Snow, Cancino, de Temple, and Schley, 1991); they contend that any rating scale the basis of which is a general theory would not be appropriate for performance assessment on specified tasks. In order to develop an empirically based rating scales, different essays should be read by multiple raters; text features which attract the readers' attention can form the basis of the rating scales, these text features which we call rubrics are scoring guides that help raters to score the essays taking these features into account. Lumley (2002) argues that "the role of scale wording [i.e., rubrics] seems to be providing justifications on which the raters can hang their scoring decisions rather than offering descriptions of the texts" (p. 266-7). When papers are read by multiple raters, "convergence on the same linguistic variables would have implications for developing scoring guides, (Cumming, Kantor, Powers, Santos, and Taylor, 2000), and perhaps, for anchoring the meaning of test performances" (p. 21).

The extent of agreement between raters depends on the common interpretations of the scale contents. Each scale is defined by statements, wordings, or rubrics which can be a cause for different interpretation of raters, less consensus and lower reliability.

Scoring descriptors and rubrics help raters to focus on certain features of written essays which are considered as necessary in decision making processes. In fact, rubrics are directing guides for raters in assessing examinees' performance. According to Valdes, Haro, and Arriarza (1992), rubrics are directing guides for raters in assessing scores; in fact, they present an "implicit theory about the nature of writing as well as implicit assumptions about the development of L2 writing skills" (pp. 334-335).

Upshur and turner (1995) assert that "published scales, and scales used as examples in books about testing, are often too broad" (p. 5). If this is the case, then can we conclude that we leave raters to assess according to their own criteria? If the answer is positive, then what is the role of rubrics in assessment?

Consistency of scores among raters can be achieved if we try to expose them to the same scoring guides. Training is one way of bringing more consistency into scoring processes. In fact, training is directed at enabling raters to make the most of scoring guides to obtain consistent results. Although training has been shown to enhance raters' reliability, its effect cannot be taken with certainty. The reasons for this dubious state are the background of raters who are trained (Cooper, 1977; Torrance, 1998), their language background (Brown, 1991; O'Loughlin, 1994; Cumming, Kantor, Powers, 2001), professional and lay raters (Shohamy, Gordon, & Kraemer, 1992), individual differences (Pula and Huot, 1993; Wolfe &Feltovich, 1994), rater idiosyncrasies (Engelhard, 1994), the effect of rater variable in the development of an occupation-specific language performance test (Brown, 1995), expert and non-expert raters (Wolfe and Ranney, 1996), rater's ethnicity, culture, mother tongue, academic and assessment experience, teaching and learning experiences (Erdosy, 2000, 2004).

White (1998) pictures the training session as the one in which the chief reader not only brings order, but also makes some room for discussing the standards with readers so that they internalize and come to own the scoring guide. In fact, he reiterates what he previously (1984) suggested for the goal of training sessions:

> The goal of rater training is to help raters internalize the scoring rubric by combining description (the rubric) with example (the anchor texts). Well-trained raters score accurately and quickly and need only occasional reference to the rubric or anchor texts. (p. 404)

The important question is whether the assortment of training and rubrics can reduce rater variability or not. Jacobs et al. (1981) state that the use of scoring rubrics in the course of training plays down inconsistency among raters which, in fact, gives rise to more consensus even if raters do not enjoy similar backgrounds.

McNamara (1996) asserts that "raters display certain characteristics in their participation in the rating process, and these characteristics are a source of potentially considerable variability in the ratings of a performance" (p. 127). He adds that traditional methods attempted to reduce or eliminate unwanted rater characteristics through training and accreditation. McNamara, however, is not sure of the success of this process of training and its effect in the course of time.

Eckes (2008) cites a number of studies showing rater training to be not as effective in reducing rater variability as expected (Lumley & McNamara, 1995; Weigle, 1998; Hoyt and Kerns, 1999; Barrett, 2001).

Eckes (2008) contends that "scoring criteria play a crucial role in rater-mediated performance assessments" (p. 156). He asserts that this is true with multi-trait or analytic scoring methods where assessments are based on a number of criteria specified to feature required traits of a written performance.

O' Sullivan (as cited in Shaw & Weir, 2007) names different raters' characteristics: physical/physiological, psychological, and experimental. Taking all these characteristics in mind, an important question is how to nullify the effect of all these characteristics upon scoring through rater training. Stahl and Lunz (1991) assert that rater training cannot eliminate differences among raters in terms of their severity. If we believe that training is directed towards, inter alia, exposing raters to a set of rubrics the reference to which can somehow secure inter- and intra-rater reliability, then we may surmise that there will be little or no inconsistency among raters. Now the question is: What if raters shared the same professional training background and teaching and assessment experience, but still we observe variability and as a result, inconsistency in their rating? The reason can be rooted in raters basing their judgments on their own criteria rather than on the traits stated in the scoring guide especially when the rubrics are holistic, vague, or wide in terms of traits introduced. Based on these ideas, then two research questions were developed:

1. Are raters' judgments limited to scoring rubrics suggested by TWE scoring guide?

2. How much of raters' scoring criteria comply with rubrics of TWE scoring guide?

Method

Participants

The writing performances of 15 students were rated by eight raters, four English native speakers and four Persian speakers who had majored in TEFL. These raters had high education with at least Master of Arts and at most Doctor of Philosophy. Their education and age were controlled. As is shown in Table 1, this group was matched in a way that each rater had a double on another group as far as age and educational degree were concerned.

Table 1
Raters' Characteristics

	Raters	Education	Major	Gender	Age	Teaching Experience	Assessment Experience
Native	A	MA	Applied Linguistics	Male	>50	28	18
	B	PhD	F/SL Education	Female	31-40	17	7-10
	C	MA	TESOL	Female	31-40	15	15
	D	MA	TEFL	Female	<30	7	5
Non-native	E	MA	TEFL	Female	31-40	5	2
	F	MA	TEFL	Female	<30	4	3
	G	MA	TEFL	Female	>50	15	10
	H	PhD	TEFL	Female	31-40	7	4

Instrument and Procedure

The rating scale used in writing assessment of TOEFL comprised the assessment instrument of this study. Its choice lies in the fact that the way it looks at writing assessment is holistic. With its six bands, TOEFL writing scale can be used to put examinees in the appropriate writing proficiency band. (see Appendix A for TOEFL Writing Band Scale)

Verbal Protocol Analysis (VPA), according to Falvey and Shaw (2006) is "a methodology based on the assertion that an individual's verbalizations may be perceived as an accurate account of information that is (or has been) attended to as a particular task is or has been undertaken" (pp. 9-10).

The type of VPA used for data collection of this study was non-mediated concurrent Think-Aloud (TA) where individuals were asked to verbalize their thoughts as they carried out the task. Introspection was preferable as Ericsson and Simon (1993) support its use due to the fact that the contents of the short term memory remain available for a very short time after they are experienced. The VPA helped to get to know the processes which led to specific marking according to the TOEFL holistic scales available to the raters of this study. TAs were collected just when raters assessed the scripts holistically.

Raters verbalized individually. After the raters' verbalizations were recorded, they were transcribed. Table 2 shows a sample of these transcripts.

Table 2

Sample Rater A's transcripts

TS	R1-250-T1
1	(well), paragraphing is present
2	[the rater looks at the number of pages]
3	just one side
4	and only two paragraphs
5	inappropriate vocabulary in the first line
6	then it starts with a definition
7	and the wrong definition *effective teaching is called eclecticism* it is not
8	overly the argument follows fair enough
9	the writer is polemic about traditional teaching
10	now we go to the second paragraph
11	just in the middle of the first paragraph he talks about effective teaching
12	and then questions without answers
13	they don't really hang together
14	and the language does not upset
15	not well organized, I am looking at
16	I am looking at the band of three
17	it is not that much
18	the description of the language
19	it is not appropriate
20	errors and insufficient details
21	the paragraphs are just two
22	there is no frequent errors of language, I think
23	the problem with focus
24	so it is three,
25	250 is three

To check for their accuracy, especially in case the researcher was not sure of the accuracy of the transcripts, the recordings besides the written records were sent back to the raters to be re-checked for their correctness.

The raters generated 48,244 words within almost 17 hours of speech. The raters produced at least 3,141 words each, and maximally the number of words came to 10,749. Table 3 illustrates the number of words uttered by the raters of the study.

Table 3
Approximate Number of Words Counted through Think Aloud

	Raters	Approximate number of words
Non-native	RA	3,377
	RB	3,141
	RC	6,105
	RD	3,234
Native	RE	10,042
	RF	6,055
	RG	5,541
	RH	10,749

Before attempting to segment the protocol, some conventions should be noted. Table 4 shows that raters' non-verbal behavior goes into brackets; italics indicate quotations read from the scripts; exact wording of the scale descriptors are underlined; pauses and unusually long silences between two words, phrases, or sentences are noted down by dots and finally parentheses separate interjections used by the raters form other TA components.

Table 4
Conventions in the TAs

Text specifications	Transcript	Samples	Sources
Conventions		(Rater, Script, Task)	
Rater nonverbal behavior	[ ]	[the rater flips over the page]	R1, 205, T1
Quotations read from scripts	Italics	*the long history of teaching has seen many fluctuations*	R5, 276, T1
Repetition of scale descriptors	Underlined	noticeably inappropriate choice of word/words	R3, 276, T1
Incomplete or undecided statements	…	umm, one thing that I do have to mention students need to take full advantage of …	R2, 284, T2
Interjections used by raters	( )	(uhuh)	R1, 258, T1

Segmentation is a controversial issue which some researchers define as arbitrary and intuitive. Ericsson and Simon (1993) assert that in speech the boundaries of phrases are usually marked by pauses. Linguistic boundaries cannot always solve segmentation problem.

Paltridge (1994) suggests that textual boundaries are made on the basis of the content of the texts instead of the way in which the content is expressed linguistically.

vanSomeren, Bernard, and Sandberg (1994) put forward another method for segmentation; they suggest that "the combination of these pauses and the linguistic structure provide a natural and general method to segment a think aloud protocol" (p. 120). They underscore that a high level of agreement between people exists while they listen to think aloud protocol. They believe that what makes segmentation more difficult and less reliable is to segment the protocol on the basis of the written text only.

The combination of linguistic structure and pauses met the segmentation of the present study. Besides, some rules were followed to complement its segmentation:

- Identification of each script by mentioning the code
  e.g., (Ok), this is two hundred and five, 205
- Traits considered as important in scoring
  e.g., there is no organization
- Personal reaction to the text which led to a specific marking system
  e.g., To give three to this paper could be a bit harsh
- Non-verbal act
  e.g., [clears the throat]
- Justifications made for a specific score
  e.g., and it is not for the argument, but for the language
- Interpretations made on the basis of the band scale
  e.g., definitely organized
- First impression scoring strategy
  e.g., quite flashy language
- Finalizing scoring processes
  e.g., so I think I would go for three for that
- Confirming given scores
  e.g., I think this is a clear four whereas others might be questioned but this one is a clear four
- Reporting the end of the scoring
  e.g., finished
- Fluctuations in scoring
  or three point seventy five

Then, a coding scheme was used for the study. The major coding categories which appear in this study are of the following types:

1. Referencing to main writing features, i.e., content, organization, grammar, vocabulary, and mechanics (for example: the code 0.0.0 for nominating the category of content)
2. Referencing to the subcategories of the main writing features (for example: the code 1.1.3 for introduction as a subcategory for organization)

3. Paper identification (for example: 5.0.0. for labeling or identifying the paper)
4. Rating behavior (for example: 6.1.0 for raters' rereading to score)
5. Referencing to the band scale (for example: 7.0.0 for reference to the band scale after reading the whole essay)
6. Interpretation of the band scale (for example: 7.1.1 for the interpretation of the band scale)
7. Scoring decisions (for example: 8.0.0 for first impression scoring)
8. Overall judgments on quality and errors (for example: 9.0.0 for commenting on the overall quality of the essay)
9. The researcher coded the whole protocol. Table 5 indicates one sample of these coded protocols.

Table 5
Sample coded protocol

TS	R1-250-T1	Codes
1	(well), paragraphing is present	1.1.2
2	[the rater looks at the number of pages]	6.0.0
3	just one side	1.0.0
4	and only two paragraphs	1.1.2
5	inappropriate vocabulary in the first line	3.1.0
6	then it starts with a definition	1.1.3
7	and the wrong definition *effective teaching is called eclecticism* it is not	0.3.1
8	overly the argument follows fair enough	6.1.1
9	the writer is polemic about traditional teaching	6.1.3
10	now we go to the second paragraph	1.1.2
11	just in the middle of the first paragraph he talks about effective teaching	0.3.1
12	and then questions without answers	7.1.6
13	they don't really hang together	0.0.4
14	and the language does not upset	7.1.1
15	not well organized, I am looking at	7.0.0
16	I am looking at the band of three	7.0.0
17	it is not that much	7.1.1
18	the description of the language	7.1.1

19	it is not appropriate	7.1.3
20	<u>errors and insufficient details</u>	7.0.0
21	the paragraphs are just two	1.1.2
22	there is no frequent errors of language, I think	7.1.1
23	<u>the problem with focus</u>	7.0.0
24	so it is three,	9.1.1
25	250 is three	9.1.2

To check for the reliability of the coded protocol, a second coder coded several parts of the transcripts. Agreement reached on the basis of 533 coding decisions within 8 episodes was 0.84, which suggests satisfactory reliability of the applied coding system.

Data Analysis and Results

The result of protocol analysis demonstrates that the raters made comments on five macro-categories: content, organization, grammar, vocabulary, and mechanics; also the number of micro-categories (e.g., content as the macro-category and clarity of ideas as its micro-category) reached 126. In all, raters produced 5657 statements. Table 6 indicates the number of statements each rater uttered.

Table 6
Total Number of Statements Uttered by Each Rater

Raters	R1	R2	R3	R4	R5	R6	R7	R8
Total number of words uttered	437	245	638	373	1164	814	733	1253

The 5657 statements were coded as was previously mentioned. The number of codes which were assigned to illustrate different categories came to 166. The difference between the number of codes and the number of categories that were mentioned before (i.e., 166 vs. 126) is because a number of codes are related, for example to: nominating categories, labeling, rating behavior, overall comments some of which are shown in Table 7.

Table 7
Some Coded Categories

Labeling	Identifying the Paper
Rating Behavior	Reading the Essay
Rating Behavior	Rereading to Score
Rating Behavior	General Impression
Rating Behavior	Rereading out loud
Rating Behavior	Personal Reaction to the Text
Rating Behavior	Nonverbal Act
Rating Behavior	Justifying the Given Score
Rating Behavior	Comments on Ideas
Rating Behavior	Self Criticism for Harshness/Lenience
Rating Behavior	Eliciting information from the researcher
Rating Behavior	Hesitation
Rating Behavior	Separating the Scoring of the Categories
Rating Behavior	Rater's Use of Fillers
Rating Behavior	Raters' rating strategies
Overall	Comment on Quality
Overall	Comment on Errors
Overall	Existence of the Title
Overall	Referring to the Topic of the Essay
Overall	The End of the Scoring
Overall	Confirming the End of the Scoring
Any	Suggestions for Mistakes

Of these 166 codes, those that were related to categories appearing in TWE were separated and counted. These categories include: reference to the band scale after reading the whole essay, reference to the band scale in the course of rating the essay, interpretation of the band scale, and criticizing the band scale. Of *5657* statements made totally, just *182* were related to TWE; as is obvious, their frequency of occurrence in the TAs came to less than *5* percent of all statements which cannot be taken as significant as it was expected to be. Regarding these analyses, our first question cannot be supported. Raters made references to TWE but they did not limit themselves to those just appearing in TWE scales; besides just five percent of all statements complied with TWE which again shows that raters go beyond what they are exposed to in TWE scales and related rubrics.

Discussion

As was obvious, the result of this study showed that raters do not refer to the scoring guide as much as it was thought. A very small proportion of raters' TAs explicitly refers to the scale which suggests that the raters' preferences for certain traits interfere with what has been stated in the rubrics. The reason can stem from the scales being either too general or too vague. Shaw (2004) sees it necessary to understand how a rating scale is going to be used by examiners. He enumerates the following rating scale criteria that are worthy of consideration:

✓ Does the scale capture the essential qualities of the written performance?

✓ Do the abilities the scale describes progress in the ways it suggests?

✓ Can raters agree on their understanding of the descriptions that define the levels?

✓ Can raters distinguish all the band levels clearly and interpret them consistently?

✓ Can raters interpret effectively any 'relative' language terms for example 'limited', 'reasonable', and 'adequate'?

✓ Do raters always confine themselves exclusively to the context of the scale?

✓ What is the role of re-training examinees in the use of the new rating scale in the rating process?

Although rubrics and rating scales are developed to assist raters to assign an accurate account of examinee's performance, they get marginalized because either raters find them in contrast with their pre-established criteria for assessment or the rubrics themselves cannot be a perfect guide for raters. Lumley (2005) states that raters find the scale descriptors "*essentially inadequate* for their purpose at times, but they continue to articulate their rating decisions in terms of the scale, because that is what is required of them" (p. 259). The only alternative, he continues, would be to refrain from further reading.

Each rater brings with him/herself a set of criteria for assessment which has been developed within years of teaching and testing experience. Not only does experience have a crucial role in assessment, but raters' preferences and viewpoints can change the criteria for assessment. Raters' characteristics cannot be neutralized through using a specific scoring scale and training. Especially raters utilize their own criteria when rubrics do not come up with robust theoretical or empirical standards.

Weigle (2002) contends that rubrics are "the most concrete statement of the construct being measured" (p. 72). Does this mean that rubrics are tangible enough for the raters to rate as accurately as it is expected? Matthews (1990) is in doubt about this claim saying:

> … they are not clearly defined; they are not always appropriate for the particular task assigned; or they straddle too obviously the linguistic/non-linguistic divide. The same descriptions make reference to abilities which were not tapped by the task set. Bare statements such as 'may pause to prepare next utterance' are of little assistance to the assessor, because they describe behavior which is ambiguous. (p. 119)

Norton-Pierce (1991) thinks that the flexibility of descriptors and rubrics leaves some rooms for the rater to award the writing with features that are not part of the scoring guide. TWE offers such flexibility to raters. Norton-Pierce (1991) states that this flexibility, which is evident in TWE, is a bonus to the writers:

> The TWE scoring guide encourages readers to focus on what the examinee does well and the descriptors and rubrics are sufficiently flexible to allow readers to use their discretion in making judgments (p. 160).

As was previously mentioned, it is believed that in the light of scoring rubric and ample training, raters can show intra- and inter-consistency in scoring. However, as Eckes (2008) states, "rater training has been shown to be much less effective at reducing rater variability than expected (p. 156).

Apart from shortcomings of rubrics, is training time sufficient for directing raters' attention to specified features on the one hand, and making them consistent with other raters through the use of a single scoring guide, on the other hand? Jacobs, Zinkgraf, Wormuth, Hartfiel, and Hughey (1981) suggest that training time for both inexperienced and experienced readers; the former, they estimate, need thirty minutes to an hour of training, whereas the latter requires fifteen or thirty minutes of training due to their backgrounds as either teachers or evaluators of composition. What can be inferred is the fact that scoring rubrics and training are the means to an end, that is to say, taking care of the traditional psychometric characteristics of tests, namely reliability. When rubrics are found inadequate by the raters, no choice is left but to employ one's own criteria and preferences for assessment.

If we take a look at TWE scoring guide, the band ranges from 0 to 6. The areas of concern are:

1. To effectively address the writing task
2. To organize well
3. To use clearly appropriate details to support a thesis
4. To display consistent facility in the use of language
5. And to demonstrate syntactic variety and appropriate word choice

Perhaps we can find traces of the macro-categories of writing in the above-mentioned factors: content, organization, and language control. As can be seen, they are too general concepts for the raters and at the same time vague terminologies to be used for accurate assessment. How can one differentiate between *well-organized, adequately organized, inappropriately organized, and inadequately organized* and much more relative concepts? Especially this applies to novice raters who commented on the essays of this study twice as much as expert raters. One reason for this difference between novice and expert raters can have its genesis in insufficient information provided in the scoring guide. The novice raters lose track of the main job by going into details that are not at times relevant to the assessment.

Vaughan (1987) found great variation in rating strategy among the nine raters in the study, even to the point that the main reason for passing an essay was not included in the detailed CUNY (City University of New York) rubrics used by the raters. It can be concluded that no matter the rubrics are too general or limited, raters develop their own criteria for the assessment and even training effect last for a while, hence the need for re-training. This is in contrast with what White (1998) believes about the role of training. He envisions that training helps raters to internalize the scoring guide, however, if raters really internalize the scoring criteria, why do we need to re-train them from time to time?

Huot (1988) names rater expectation as a determining factor in assessment procedure. He relates rater expectation to rater experience. He quotes Smith saying that "information available in the brain is more important in reading than information available to the eyes from the print on the page" (p. 77). He underscores that "reader expectation shaped by personal and professional experience will always be a strong yet hard-to-define influence on holistic raters" (p. 80). We can infer that raters exceed the limit imposed by given rubrics which is due to their experience and previous expectations.

Conclusions

On the whole, it seems logical to closely examine what raters bring to the assessment field. The objective of this qualitatively-based study was to bring into limelight the importance of raters' variables in assessment prior to developing any kind of scoring rubrics. Rater expectation (Huot, 1988), rater type (Eckes, 2008), rater characteristics (O'Sullivan, 2000), or other rater-related concepts tap on the fact that rubrics should be in tune with what raters think of assessment criteria. As Eckes (2008) states that we can solve many problems associated with the lack of consensus on scoring criteria by incorporating into scoring guide those criteria which was not a part of raters' profile for assessment, therefore, "redirecting the attention of particular rater types to criteria not captured within their respective scoring profile, thus contributing to a more balanced use of the criteria deemed relevant in the assessment" (p. 179).

It is worth to duplicate this study taking into account the analytic scoring guide instead of holistic one. Besides, introspection, if supplemented with retrospection, can be a stronger basis for any study. The number of raters and the number of papers can also end up in different results. And above all, the more we know about raters, the more information we can obtain and more robust conclusion we can make. It is of value to inspect rater variables more closely to understand how they can affect the result of assessment. There have been a few studies which made use of verbal protocol. Perhaps verbalized thoughts can give us much of insight that we cannot obtain through other quantitatively-based methodologies.

References

Cumming, A., Kantor, R., Powers, D., Santos, T., & Taylor, C. (2000). *TOEFL 2000 writing framework: A working paper (TOEFL Monograph Series No. 18).* Educational Testing Service: Princeton, NJ.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing*. Cambridge University Press: Cambridge.

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

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (2nd ed.). The MIT Press: Cambridge, MA.

Falvey, P., & Shaw, D. S. (2006). IELTS writing: Revising assessment criteria and scales (phase5). *Cambridge ESOL Research Notes, 23*, 7-13.

Greenberg, K. L., Weiner, H. S., & Donovan, R. A. (1986).*Writing assessment: Issues and strategies*. Longman: New York.

Herman, J., Aschbacher, P., & Winters, L. (1992). *A practical guide to alternative assessment.* Association for Supervision and Curriculum Development: Alexandria, VA.

Huot, B. A. (1988). *The validity of holistic scoring: A comparison of the talk-aloud protocols of expert and novice holistic raters* (Unpublished doctoral dissertation). Indiana University of Pennsylvania, US.

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

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

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

Matthews, M. (1990). The measurement of productive skills: Doubts concerning the assessment criteria of certain public examination. *ELT Journal, 44*(2), 117-121.

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

Norton-Pierce, B. (1991). Review of the TOEFL Test of Written English (TWE) scoring guide. *TESOL Quarterly, 25*(1), 159-163.

O'Sullivan, B. (2000). *Towards a model of performance in oral language testing* (Unpublished doctoral dissertation). University of Reading, UK.

Paltridge, B. (1994). Genre analysis and the identification of textual boundaries. *Applied Linguistics, 15*(3), 288-299.

Shaw, S. (2004). *IELTS writing: Revising assessment criteria and scales (phase 3). Cambridge ESOL Research Notes, 16*, pp. 3-7.

Stahl, J. A., & Lunz, M. E. (1991). *Judge performance reports: Media and message.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal, 49*(1), 3-12.

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

vanSomeren, M. W., Bernard, Y. F., & Sandberg, J. A. C. (1994). *The think aloud method: A practical guide to modeling cognitive process*. Academic Press: London.

Vaughan, C. (1987). *What affects raters' judgments?* Paper presented at the meeting of the Conference on College Composition and Communication, Atlanta, GA.

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

White, E. M. (1998). *Teaching and assessing writing: Recent advances in understanding, evaluating, and improving student performance* (2nd ed.). Calendar Islands Publishers: Maine.

Appendix A

TOEFL Writing Band Scales[1]

6 An essay at this level

- Effectively addresses the writing task
- Is well organized and well developed
- Uses clearly appropriate details to support a thesis or illustrate ideas
- Displays consistent facility in use of language
- Demonstrates syntactic variety and appropriate word choice though it may have occasional errors

5 An essay at this level

- may address some parts of the task more effectively than others
- is generally well organized and developed
- uses details to support a thesis or illustrates an idea
- displays facility in the use of language
- demonstrates some syntactic variety and range of vocabulary, though it will probably have occasional errors

4 An essay at this level

- addresses the writing topic adequately but may slight parts of the task
- is adequately organized and developed
- uses some details to support a thesis or illustrate an idea
- demonstrates adequate but possibly inconsistent facility with syntax and usage
- may contain some errors that occasionally obscure meaning

3 An essay at this level may reveal one or more of the following weaknesses:

- inadequate organization or development
- inappropriate or insufficient details to support or illustrate generalizations
- a noticeably inappropriate choice of words or word forms
- an accumulation of errors in sentence structure and/or usage

2 An essay at this level is seriously flawed by one or more of the following weaknesses:

- serious disorganization or underdevelopment
- little or no detail, or irrelevant specifics
- serious and frequent errors in sentence structure or usage
- serious problems with focus

1 An essay at this level

- may be incoherent

[1] Educational Testing Service (1986). Test of Written English Guide. Princeton, NJ: ETS

- may be undeveloped
- may contain severe and persistent writing errors

0 A paper is rated 0 if it contains no response, merely copies the topic, is off-topic, is written in a foreign language, or consists of only keystroke characters.