World Applied Sciences Journal 11 (6): 695-699, 2010
ISSN 1818-4952

Raters' Conceptions of a Good Writing and Effect of Training on Their Conceptions

Sehnaz Sahinkarakas

English Language Teaching Department,
Faculty of Arts and Humanities, Cag University, Yenice, Mersin, Turkey

Abstract: In the field of assessing writing, various studies have been conducted on raters' judgments and effects of training on these judgments. This study investigated how raters conceived a good writing and whether rater training affected their conceptions. The conceptions of five raters were identified through a repertory grid technique which was conducted before and after a rater training session. The constructs they produced in these repertory grids revealed that raters put more emphasis on some of the text features in a good writing. It was also found that training might affect raters' understanding of a good writing to some extent.

Key words: Personal construct theory · Repertory grid · Rater beliefs · Rater training

INTRODUCTION

Assessment in direct testing of writing is a complex process. It relies on human judgment and is mostly subject to rater behavior. Literature in the field of assessing writing has revealed that in the process of rating, raters focus on various themes and their decisions are affected by many factors. Huot [1] and Pula and Huot [2], for example, looked at the influence of raters' backgrounds on scoring. Their think-aloud protocol studies and interviews revealed that raters' personal backgrounds, professional trainings and work experiences might affect the way they score writing. Similarly, Wolfe [3] investigated rater differences and found that less proficient raters make evaluative decisions earlier than more proficient ones. Weigle [4] looked at rater variables and examined the effects of training on raters. She pointed out that the norming process, that is, the scoring criteria, modifying raters' expectations and raising awareness of the need for rater agreement would result in bringing the raters in line with each other. Lumley [5] carried out think-aloud protocol study with four experienced teachers and stated that raters' understanding of the scale was one of the justifications raters made while deciding the score.

Freedman and Calfe [6] described a three-staged model suggesting that the raters read the text to construct a text image, evaluate the image and articulate the evaluation. In this model, while building the text image, the raters interpret the text based on their own world knowledge, beliefs and values and knowledge of the writing process. It is necessary to refer to Kelly's Personal Construct Theory here. According to Kelly [7], each of us tries to make sense of the world as we experience it by creatively forming and testing 'constructs' about the world around us. These constructs are used for predicting things to come and as we have new experiences, we realize these predictions either as correct or misleading which might provide the basis for revision of the constructs. Thus, it is possible to assume that in the rating process, raters' own constructs formed by their earlier experiences might affect their decisions in rating and these constructs might change when they receive rater training. During the training, raters are expected to come to an agreement by discussing the criteria on which the scores are based. These discussions might lead the raters to test their previously developed constructs and to modify them or change them with new ones.

With these considerations in mind, this study mainly deals with the personal constructs of raters related to 'good writer' and seeks to answer these questions:

- What do the raters mainly focus on while identifying a good writer?
- Is there a role of rater training in raters' constructs of a good writer?

Corresponding Author: Sehnaz Sahinkarakas, English Language Teaching Department, Faculty of Arts and Humanities, Cag University, Yenice, Mersin, Turkey, Tel: +90-532-472-1011, E-mail: sehnazsahin@gmail.com.

Method: The study's participants were five English teachers of English all of whom were pursuing doctoral degrees in Teaching English as a Second Language. All the raters had taught English for at least five years. Two of these teachers were experienced in rating writing papers and familiar with the TOEFL, Test of Written English (TWE) scoring criteria. The other three were inexperienced in such a rating process.

Initially, Repertory Grid Technique, based on Kelly's Personal Construct Theory, was used to collect data from all the participants to observe how they perceived 'a good writer'. For this, each participant was asked to fill in a grid in which they compared and contrasted the writing abilities of nine people whose writing they were familiar with. Using this data collection technique, each of which lasted approximately 45 minutes, the constructs of each rater were identified. This was followed with an interview about the constructs they developed in the grid to understand what they really meant with the constructs. After that, all the raters were invited for a training session. This session, which lasted about an hour, was tape-recorded and the raters scored 10 essays using the TOEFL, TWE scoring rubric. After the training session, each rater again was asked to fill in the Repertory Grid, which was again followed with an interview.

The text features to which ESL/EFL raters attend while rating TOEFL essays identified by Cumming, Kantor and Powers [8] were used to categorize the constructs that the raters produced. In their study, Cumming, *et al.* identified 12 categories, each of which was expanded by possible text characteristics that a rater might use. To reach consensus, five experienced researchers in the language teaching field matched the constructs with these 12 categories. Accordingly, some of the categories were combined and the constructs that most of the researchers agreed on were grouped under the same category. Thus, in this study, the text features were accepted to be grouped under eight categories: Layout, Vocabulary, Rhetorical Organization, Personal Situation of Writer, Style, Topic Development, Sentence/Grammar and Mechanical Competence.

RESULTS AND DISCUSSION

The data were first analyzed determining the sum of the constructs produced by the five participants in each data collection. The constructs, a total of 79 in each data collection, were grouped according to the text features to which they refer. As seen in Table 1, raters apply the biggest weighting to *style*: a total of 42 out of 158 constructs were produced for this text feature only. Some of the constructs that were mostly referred to for style were fluency, clarity, originality, intelligibility, interestingness and having a clear message. Other specific features referred to for style were *topic development* (30 constructs) and *personal situation of writer* (28 constructs). The constructs that could be linked to topic development were mainly related to whether the ideas were focused and rich and whether the text addressed the task. As for personal situation of writer, the constructs produced were generally related to the writers' motivation to write, their experience in writing and their readiness to learn writing.

Aside from the text features mentioned above, quite a number of constructs were produced related to *sentence/grammar* and *rhetorical organization*. The ones under the sentence/grammar category were on the topic of complexity and/or variety of the structures used and control of the language. Only a few constructs referred to specific grammar items such as tenses and run-on sentences. Constructs on rhetorical organization were mostly those regarding organization of the text.

Table 1: Total number of constructs produced before and after training

	Before training	After training
Layout	4	3
Vocabulary	3	6
Rhetorical Organization	9	8
Personal Situation of Writer	17	11
Style	22	20
Topic Development	12	18
Sentence/Grammar	9	10
Mechanical Competence (punctuation, spelling)	3	3
TOTAL	79	79

Table 2: Distribution of constructs for each rater

	Rater 1		Rater 2		Rater 3		Rater 4		Rater 5	
	before	after	before	after	before	after	before	after	before	after
Style	8	7	1	1	9	7	2	4	2	1
Topic Development	2	2	2	3	5	7	--	2	3	4
Personal Situation of Writer	2	--	--	--	3	--	9	8	3	3
Sentence/Grammar	2	4	4	2	--	1	1	2	2	1
Rhetorical Organization	2	2	3	1	1	2	1	1	2	2
Vocabulary	--	1	1	1	--	2	--	1	2	1
Layout	--	--	--	--	--	--	3	2	1	1
Mechanical Competence	--	--	3	3	--	--	--	--	--	--
TOTAL	16	16	14	11	18	19	16	20	15	13

As seen in Table 1, *vocabulary*, *layout* and *mechanical competence* were not often applied by the participants. Constructs on vocabulary were generally about the variety and richness of the word choice, constructs on layout were related to editing and proofreading the writing and mechanical competence included constructs about punctuation and spelling.

It is possible to imply from these findings that raters generally give importance to style, topic development and personal situation of the writer in a written text. However, this might not be the case if each rater is examined separately. As seen in Table 2, it is still possible to assume that style is an important feature on which each rater focuses. All the raters developed at least one construct about style in each of the data collections. It seems that this feature was important to all the raters while rating, but for Rater 1 and 3 it was the priority among the text features as they developed most of the constructs under this category.

Topic development was another feature that the raters attended to while rating. All the raters except Rater 4 produced constructs under this category in both data collections. Although Rater 4 did not produce a construct on this category in the data collection before training, he developed two constructs after that, which might show that topic development was a feature for him to consider while rating.

Although having quite a big number of constructs in total, the picture of personal situation of the writer is not the same when raters are evaluated individually. Seventeen out of 28 constructs related to this feature were produced only by Rater 4, which might indicate that he considers personal situation of a writer the most important feature in a written text. This feature was not that important to other raters, especially to Rater 2 as he did not generate a construct on this feature at all. In fact, if the distribution of the constructs among the raters rather than the total number of them is considered, it is possible to state that rhetorical organization and sentence/grammar are more important than personal situation. Constructs related to these features were produced by all of the raters at least once. The least important features were vocabulary, layout and mechanical competence.

The study's second aim was to determine whether rater training session would play a role in the constructs the raters produced related to a good writer. Analysis of the data revealed that the training session had some effect on vocabulary. Before training only two raters produced constructs related to vocabulary. After training, however, all of the raters produced at least one construct associated with vocabulary. This might be due to the scoring criteria and the discussions done during the training session. The TOEFL, TWE which was used as the scoring criteria in training has explicit statements of items related to vocabulary. Thus, while discussing scores the raters gave for the writing papers, vocabulary was referred to many times. For example, Rater 3 mentioned vocabulary a few times while talking about the rationale of the scores she assigned for the papers:

I can see a range of vocabulary; not enough to get a 6 but still… (Rater 3)

It made it difficult to follow his argument. It was difficult because of his word choice. It impedes the communication, so I gave 3. (Rater 3)

The problem wasn't at the word level; I find it difficult in terms of the organization of the ideas. (Rater 3)

Similarly, other raters referred to the vocabulary in the papers from time to time:

Also the words were not really appropriate. (Rater 1)

Word choice, well you have to read it twice. (Rater 4)

Band 3 says a noticeably inappropriate choice of words or word forms. Definitely more than 60% of all are this last line but not all; that's why 3.5. (Rater 5)

Here, raters' experience is worth mentioning when considering the effect of training on raters. Raters 1, 3 and 4 were not experienced in assessing writing papers in high-stake tests. This might be why they did not produce any constructs related to vocabulary before training. The scoring criteria and the discussions in the training session might have affected their understanding of a good writer and thus they explicitly referred to vocabulary in the data collection after training.

The training session probably affected the raters' focus on the topic development as well. As seen in Table 1, there is a tendency to produce more constructs on this category after training: the total number of constructs increased from 12 to 18. In the training session, this issue was discussed for all the papers many times. As a result of these discussions, the raters might have placed more emphasis on topic development. Following are some examples related to topic development that the raters mentioned while discussing.

It does address the topic; if you read the whole topic, it really does address the whole thing. It does use some details to support and illustrate ideas. (Rater 5)

The paper gives specific examples. (Rater 4)

*She answers it at the end (*the task*)….Compare the situation in your country to the USA. She did that in the body paragraph with the example and the example was appropriate and supportive so basically she knows what to answer.* (Rater 5)

I gave 5 because I thought that the examples are good. (Rater 3)

The other effect of the training might be seen on the category of personal situation of the writer. The picture for this category was different from those on vocabulary and topic development. One experienced and three inexperienced raters produced constructs about this feature before the training session. However, after training, two of the inexperienced raters (Raters 1 and 3) dropped constructs related to personal situation of writer. The constructs they produced in this category before training were generally related to the cultural elements present in a written text. Probably they were affected by the fact that the TWE writing criteria did not include statements about this issue and thus, they did not refer to such text features after training. The third inexperienced rater (Rater 4), however, still had many constructs under this category. In fact, he was the one who produced most of the constructs related to personal situation of the writer before and after the training. This might imply that for Rater 4, personal situation is the priority of being a good writer.

The analysis showed that aside from the vocabulary, topic development and personal situation of the writer, there was not an important effect of the training on the other text features. The minor changes in some of the constructs could be the natural outcome of such a data collection.

CONCLUSION

Having a small number of participants in this study, it is not possible to generalize the results. However, it still reveals some implications related to raters' understanding of a good writing which might affect their scoring decisions. The first implication is that style of a writing piece is the chief text feature that a rater perceives in a good writing. Whether raters are experienced or inexperienced, trained or not trained, they judge a writing piece effective if it has fluency, original ideas, creativity and messages, all of which refer to style. The next text feature that they give importance is topic development, which refers to whether the written text is focused and supported with rich and relevant examples. Undoubtedly, raters sense a good writing piece from other points of views as well; however, they generally tend to consider them after evaluating style and topic development.

Many studies have been conducted on the effect of training on raters' judgments [4], [6], [9]. This study, however, can only reveal that rater training has some effect on raters' perceptions of a good writing. Training sessions can make raters add or drop new concepts in their belief systems. It can only be assumed that this change in the perception might affect raters' scoring decisions. Thus, it might be worth conducting a further research that studies to what extent raters' perceptions are reflected in their scoring decisions.

REFERENCES

1. Huot, B.A., 1993. The influence of holistic scoring procedures on reading and rating student essays. In M.A. Williamson and B.A. Huot (Eds.), Validating Holistic Scoring for Writing Assessment, pp: 206-236. Cresskill, NJ: Hampton Press.

2. Pula, J.J. and B.A. Huot, 1993. A model of background influences on holistic raters. In M.A. Williamson and B.A. Huot (Eds.), Validating Holistic Scoring for Writing Assessment, pp: 237-265. Cresskill, N.J.: Hampton Press.
3. Wolfe, E.W., 1997. The relationship between essay reading style and scoring proficiency in a psychometric scoring system. Assessing Writing, 4(1): 83-106.
4. Weigle, S.C., 1994. Effects of training on raters of ESL compositions. Language Testing, 11: 197-223.
5. 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.
6. Freedman, S.W. and R.C. Calfee, 1983. Holistic assessment of writing: experimental design and cognitive theory. In P. Mosenthal, L. Tamor and S. Walmsley, (Eds), Research on writing: principles and methods, pp: 75-98. New York: Longman.
7. Kelly, G., 1955. Principles of Personal Construct Psychology. New York: Norton.
8. Cumming, A., R. Kantor and D.E. Powers, 2001. Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: an investigation into raters' decision making and development of a preliminary analytic framework. Princeton, New Jersey: Educational Testing Service.
9. Huot, B., 1990. Reliability, validity and holistic scoring: what we know and what we need to know. College Composition and Communication, 41(2): 201-213.