RESEARCH NOTES

Raters' social considerations in the essay rating process

The case of Chinese assessors of high-stake exam essays written in English*

Jianlin Chen and Lies Sercu
Lanzhou University / KU Leuven

The social cognitive view on essay rating process argues that human essay rating is constrained by a series of the measures issued by the test institution. However, studies of the institutional constraints all focused on text quality relevant factors interfering into the rating process. Text quality irrelevant social factors remain uninvestigated. Taking the TEM8 (a test for English majors in China) as an example, the present study explores into those factors. Raters' think aloud protocols when rating essays and the follow up interviews identified a number of text quality irrelevant social factors raters bring into consideration when scoring essays, such as institutional awareness, test knowledge, test taker expectation, knowledge of rating system, ethical consideration and physical condition. The way those factors influence the rating process is further discussed. The results are meaningful both for the understanding of the essay rating construct and human essay rating practice.

Keywords: essay rating, raters, social factors

1. Introduction

It has often been wondered how international language testing services can guarantee that student essays written in a foreign language in very diverse learning contexts and rated and scored by many different locally based raters are rated and scored in the same way? Do all the raters assign scores with the same degree of

* Supported by the Fundamental Research Funds for the Central Universities in China, with the item number 14LZUGBWZY001.

Dutch Journal of Applied Linguistics 3:1 (2014), **101–108**. DOI 10.1075/dujal.3.1.06che
ISSN 2211–7245 / E-ISSN 2211–7253

lenience or strictness? The key to answering those questions will have to be the understanding of the essay rating process and the factors they bring into consideration when scoring essays. Many studies have investigated human essay rating process (e.g. Freedman & Calfee, 1983; Cumming, Kantor, & Powers, 2002; Lim, 2011), which, however, mainly focus on raters cognitive operations. According to those studies, scores assigned to an essay are only the results of text quality related evaluation and no text irrelevant elements are taken into consideration. As a social practice, however, raters' rating behaviors will be constrained by social factors and, therefore, the scores they assign on essays will be affected by their social considerations. Few studies, however, are found to investigate into raters' social considerations in scoring essays. The present research will focus on the social factors raters consider in the essay rating process and attempts to answer the following two questions: (1) what are the social factors raters consider in the essay rating process and (2) how do they interfere into the rating process.

2. Literature review

Human essay rating process has been the focus of a great number of studies. Some of them investigated essay rating sequence (e.g. Freedman & Calfee, 1983; O'Sullivan & Rignall, 2007); many other investigated into rater behaviors (e.g. DeRemer, 1998; Cumming et al., 2002; Eckes, 2008) and rating styles (e.g. Vaughan, 1991; Lumley, 2005; Lim, 2011). Those studies have revealed that views of the nature of this process have undergone a process of evolution.

The earliest view might be defined as the behavioral view. This view defines essay rating as a process of recognizing features representing text quality. Raters' role in the behavioral rating process is to find out those text features defined in the rating scale and assign a score according to the descriptions of the scale levels. The underlying hypothesis is that there is an explicit relationship between text quality and various quantifiable measures of text features, such as grammatical control and use of vocabulary. For example, Veal (1974) found that there was a high correlation between t-unit[1] of the writings and scores assigned to them. This result was confirmed by other studies (Stewart & Grobe, 1979). Other quantifiable features were also found to have a high correlation with text quality, such as lack of error (Stewart & Grobe, 1979) and word choice (e.g., Neilsen & Piche, 1981).

A significant shift from the conceptualization of human essay rating process probably occurred in the frequently cited studies by Freedman (1981) and Freedman and Calfee (1983). In their studies exploring factors influencing rating

1. T-unit is "one main clause with all subordinate clauses attached to it" (Hunt, 1965, 20).

process, they identified elements other than quantifiable text features, such as rater individual expertise, time of assessment, type of text, the kind of training and supervision provided. These studies implied a cognitive view on human essay rating process in that they regarded raters not simply as an interpreter of the text, rather, they assumed the rating process as raters' cognitive interaction between raters and various other factors, such as task type, task topic, and rater training. An ever-growing body of work on rater cognition have been available in the literature of performance writing assessment and most of them have explored the different scoring styles used by raters (e.g. Vaughan, 1991; Sakyi, 2000; Eckes, 2008) or illustrated a range of rating strategies employed during the rating process (e.g. Cumming et al., 2002; Knoch, 2009; Lumley, 2002).

As the social aspect of language assessment have been recognized widely among language testers and researchers (e.g. Davies, 1997; Shohamy, 2001), the evaluation of writing was regarded not only as rater's cognitive activity, it was also regarded as a social practice. A comprehensive study that took the social cognitive view was conducted by Lumley (2005). Lumley recognized the institutional nature of a test and argued that if the rating process is observed from the social perspective, it involves major socially motivated components. His research regarded the rating process as a social cognitive process where raters' cognition operates under constrains assigned by the institution. Specifically, when candidates' performance samples are elicited and presented before the raters, a range of institutional constraints are brought into operation to regulate the rating process with the purpose for rating consistency, or rating reliability. Those constrains include: (1) the rating scale; (2) rater training; (3) reorientation; (4) the choice of raters, and (5) the requirement of professionalism from the raters.

Lumley's study is meaningful in that it has brought the study of rating activity from the cognitive stage into the larger social field, where social factors are considered to influence the essay rating process. However, his study only focuses on the institutional factors, which is limited in that apart from the institution constraints, raters will also bring into consideration the larger social factors such as raters' knowledge of the test, the stake-holders, their considerations of test fairness, ethical issues etc. More importantly, the institutional constraints in Lumley's study are mainly factors employed to train raters to reach rating consistency in evaluating text quality relevant aspects. The social factors that are text quality irrelevant remain unexamined. The present study, taking the TEM8 essay rating as an example, will focus not only on the institutional factors, but also on the text quality irrelevant social factors. Specifically, the study aims to answer the questions of what are the social factors raters bring into consideration when scoring essays and how do those factors interfere into essay rating process?

3. Research methodology

3.1 The TEM8 and the essay rating

Test for English Majors band8 (TEM8) is a criterion-referenced English language test specifically targeted at university undergraduates majoring in English Language and Literature in China (Jin & Fan, 2011). The writing section of the TEM8 requires test takers to write an essay of 400 words in 45 minutes on the topic prescribed in the prompt. The TEM8 essay rating uses an analytical rating scale that is composed of three broad dimensions (idea/content, language and mechanics). The TEM8 writing test takes a series of quality control measures, including centralized rating, careful choice of raters, the computer-assisted online scoring, the scientific, comprehensive and feasible rating scale; carefully chosen benchmark essays; rater training and rating supervision (Zou, 2011).

The raters were trained before the actual rating. The TEM8 employs a computer-assisted online scoring which not only reduces the rating cost, but also makes rating quality supervision much more convenient and efficient because every rater's rating performance could be checked by the supervisors on the computer.

3.2 Research design

Two main data collection methods were employed in the research: think aloud verbal protocols (TAPs) conducted by six raters when rating ten essays and follow up interviews to the six raters. While the purpose of the TAPs data was to identify what are the social factors raters consider when scoring the essays, the follow up interviews were conducted in order to investigate into how those factors interfere into the rating process and why raters take them into consideration. Both TAPs and interviews were recorded and transcribed and coded in the software Nvivo. A coding system (Table 1), initially developed after literature reviewing and was then revised according the specific rating context, was used to code the TAPs data. The interviews employed a semi structured approach by which the main focuses of the questions were on how do raters think the social factors interfere into their rating process and why.[2] The six raters (three males and three females) all took part in the year 2012's TEM8 essay rating. They were all university English teachers majoring in English language and literature and all experienced raters of large scaled English tests in China. The ten essays were chosen by the testing experts as representing the whole range of quality levels of the students' performance on the

2. Since raters were not likely to have such meta-cognition as identifying social factors, they were prompted to reckon on those hypothesized factors by the researcher.

Table 1. Coding system of TAPs data

Social factors	Definition	Name of nodes
Institutional awareness	Awareness of institutional requirements	IA
Test knowledge	Knowledge of the TEM8	TK
Test taker expectation	Expectations of advanced English learners	TT
Knowledge of rating system	Knowledge of on-line rating, statistical knowledge, etc.	KR
Ethical consideration	Consideration of fairness, morality and other ethical issues	EC
Physical condition	Noise, colleague pressure, fatigue, mood, etc.	PC

year 2012's TEM8 essay test. The six raters did TAPs rating immediately after they finished rating the 2012 TEM8 essays and follow up interviews were immediately conducted after their TAPs rating.

4. Results

Table 2. The social factors and their frequencies

Category	Institutional awareness	Test knowledge	Test taker expectation	Knowledge of the rating system	Ethical consideration	Physical condition
Nodes	IA	TK	TT	KR	EC	PC
Source*	6	6	4	6	4	5
Frequency	28	10	7	9	6	15
Total Freq.	75					

* source refers to the number of the raters' reports in which the nodes were identified

As the table shows, there are six social factors considered by raters in rating the TEM8 essays: institutional awareness, test knowledge, test taker expectation, knowledge of rating system, ethical consideration and physical condition. Institutional awareness (IA) is the most frequently mentioned factor by all the six raters (28 times which is more than one third of the total frequency). Ethical consideration (EC) is, in spite of the low frequency 6, still mentioned by 4 of the six raters.

Based on the above findings, the present research proposes a framework of the social factors interfering into human essay rating process as shown in Figure 1, which is composed of two sub structures: the situational and the external. When

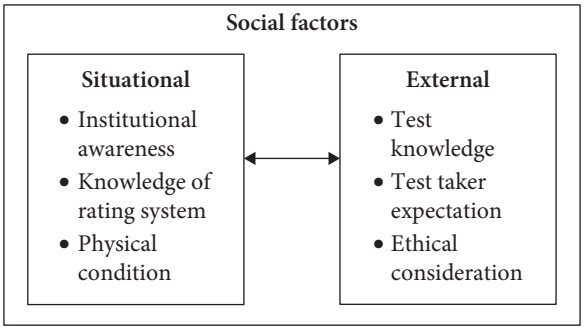

Figure 1. The framework of the social factors

rating an essay, raters will adopt the scores they assigned in order to comply with the constraints issued by the test institution, the rating system and the rating condition. Raters may also adopt the originally assigned scores as a result of the fact that they will refer to their knowledge with respect to the test, test takers, ethical issues and other social factors.

5. Conclusion

Several general conclusions can be drawn from this study. Firstly, human essay rating is not, as the previous studies implied, only a cognitive operation and test scores are not only the results of the evaluation of text quality according to the rating criteria. They are the results of both cognitive and social considerations. Raters may adjust the initially assigned text quality-related scores when social factors are brought into consideration. In addition, there are actually three scores operating in the rating process: the text quality score, the text quality-irrelevant score and the reporting score. The text quality score is the evaluation of an essay when only text features are considered; the text quality-irrelevant score refers to the modulating score when social considerations interfere into raters' decision making. The relationship among these three scores may be expressed in the equation: The reporting score = the quality score ± the quality irrelevant score. Thirdly, knowing the influence of social factors on the rating process is also meaningful for rater training, rating supervision, rating physical condition improvement and etc. For example, the content of rater training will need to include social factors as well; the supervision needs to be promptly and the rating physical condition needs to be more comfortable and etc.

References

Cumming, A., Kantor, R., & Powers, D. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal*, 86, 67–96. DOI: 10.1111/1540-4781.00137

Davies, A. (1997). Introduction: The limits of ethics in language testing. *Language Testing*, 14, 235–241. DOI: 10.1177/026553229701400301

DeRemer, M. (1998). Writing assessment: Raters' elaboration of the rating task. *Assessing Writing*, 5, 7–29. DOI: 10.1016/S1075-2935(99)80003-8

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, 25(2), 155–185. DOI: 10.1177/0265532207086780

Freedman, S.W. (1981). Influences of evaluation of expository essays: Beyond the text. *Research in the Teaching of English*, 15, 245–255.

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

Hunt, K. (1965). *Grammatical structures written at three grade levels. (NCTE Research report No. 3)*. Champaign, IL, USA: NCTE.

Jin, Y., & Fan, J. (2011). Test for English Majors (TEM) in China. *Language Testing*, 28(4), 589–596. DOI: 10.1177/0265532211414852

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*, 26(2), 275–304. DOI: 10.1177/0265532208101008

Lim, G.S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*, 28, 543–560. DOI: 10.1177/0265532211406422

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. DOI: 10.1191/0265532202lt230oa

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

Neilsen, L., & Piche, G. (1981). The influence of headed nominal complexity and lexical choice on teachers' evaluation of writing. *Research in the Teaching of English*, 15, 65–74.

O'Sullivan, B., & Rignall, M. (2007). Assessing the value of bias analysis feedback to raters for the IELTS writing module. In L. Taylor, & P. Falvey (Eds.), *IELTS collected papers* (pp. 446–476). Cambridge: Cambridge University Press.

Sakyi, A. (2000). Validation of holistic scoring for ESL writing assessment: A study of how raters evaluate ESL compositions on a holistic scale. In A.J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 130–153). Cambridge: Cambridge University Press.

Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. London: Longman.

Stewart, M., & Grobe, C. (1979). Syntactic maturity, mechanics of writing, and teachers' quality ratings. *Research in the Teaching of English*, 13, 207–215.

Vaughan, 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.

Veal, L.R. (1974). *Syntactic measures and rated quality in the writing of young children. (Studies in Language Education, Report No. 8)*. Athens: University of Georgia. (ERIC Document Reproduction Service No. 090 55).

Zou, S. (2011). On enhancing test fairness: The case of the TEM4 and TEM8. *Foreign Language Testing and Teaching*, (1), 42–50.