A Study on How to Improve Raters' Expertise in Rating Cet-4 Compositions

Junhong Lu

Foreign Language Dept., Xi'an University of Technology, Xi'an, 710048, China

Keywords: Raters, Etc-4, Online essay rating, Rating process, Quality of rating, Rating reliability

Abstract: Online rating of English compositions is often used in all kinds of English tests, and raters have become one of the important factors affecting the rating quality. This paper introduces the rating process of CET-4 English composition, and analyzes the assessment methods of several common types of compositions. In addition, the statistics provided by the system influences raters' composition rating. This study finally investigates the relevant statistical parameters of the CET Online Marking System to provide reference for raters and examination institutions, thus improving the quality of English essay rating to ensure the fairness, accuracy and the reliability of the examination.

1. Introduction

At present, in all kinds of English tests in China, the assessment of English compositions is mostly implemented online, and the holistic scoring method is adopted to rate English compositions. The online scoring rating is defined as a process by which the essay scripts are scanned and the images transmitted electronically to an image server at the test control center. These images are then distributed electronically and marked on screen by raters. In CET-4 tests, test scores are reached subjectively using rating scale descriptors to guide the rater towards a score. Descriptions of a test takers ability can then be produced by relating the score to the corresponding scale descriptors and the construct of language ability on which the rating scale is based.[1]

For a long span of time, the rating of ESF writing has been dependent on human raters, which is inescapably accompanied by subjectivity on the part of raters. There are some factors that affect the rating result including rater training, raters' interaction with the rating criteria, raters' physical and emotional conditions, raters' attitudes towards the rating work, and raters' English proficiency level.[2] Because raters hold different scoring standards, different severity, and are influenced by subjective factors such as language ability, appreciation habit and mood, the grading results are different. Sometimes, the scores given by different raters for the same composition are quite different. The stress on the different component of writing proficiency might lead to the formation of different judgment criteria, which results consequently in different scores for the same essay.[3]

Some raters are too subjective and arbitrary, some raters are lack of experience in marking essays, or their attitude is not serious, which will lead to low accuracy and poor quality of the assessment. How to Improve the quality of the English composition assessment is crucial to ensure the fairness, accuracy and the reliability of the test. The author has participated in the assessment of English composition of national exams for many times. This article analyzes the assessment methods of several common types of compositions through personal experience and investigates the relevant statistical parameters of the CET Online Marking System to provide reference for raters and examination institutions.

2. Rating Process of Cet-4

The College English Test Band Four (CET4) is a large-scale standardized test administered nationwide by the Higher Education Department of the Ministry of Education in China. The purpose of CET is to examine the English proficiency of non-English major undergraduate students in China and ensure that these students reach the required English levels specified in the National College English Teaching Syllabuses.

  DOI: 10.25236/ieesasm.2020.034

It is taken by millions of students every year. The writing section of the CET4 requires test takers to write an essay of 120 words in 30 minutes on a topic prescribed in prompt. The CET4 writing test rating takes a series of quality control measures, including centralized rating, careful choice of raters, the computer-assisted online scoring, scientific, comprehensive and feasible rating scale, rater training and rating supervision.[4]

First, raters, who are mostly experienced EFL teachers teaching college-level courses to non-English majors, will be recruited by a certain marking center. The rater first learns the rating scale, analyzes the benchmark essays. They are trained in advance and supervised during the rating process. The training process includes the choice and description of the quality of benchmark essays, rating instruction, raters' practicing rating, feedback and negotiation. The rater scores the sample essays according to the scoring criteria. The rating is based on global scoring method. That is, the scorer reads the composition and then grades it according to the overall quality. This method is also called the "impression scoring method". That is, the rater gives reward scores based on their impressions of the essays, rather than the number of language errors committed.

3. Rating Method

3.1 Rating Scale

The marking scheme used in CET 4 essay marking was a 15-point holistic scale with 5 bands (1-3, 4-6, 7-9, 10-12, 13-15). The key to rating a composition is to categorize each essay into one of the five bands. The raters are asked to decide whether essay matches the descriptors (anchor scripts) for that band exactly. Should the rater find the essay to be slightly better or worse than the descriptors (anchor scripts), he or she may either add or subtract one point.[5]

It is necessary to seriously study the anchor scripts (Range-finders) to identify the basic characteristics of each band of composition. For example, the highest level requires a clear expression of ideas, coherent text and basically no language errors. The second-highest level of the composition is a little poor, with a small amount of linguistic errors. The composition of the mid-range essay is not clear enough to express ideas, the text is barely coherent, but overall it is passable. The composition of the five-point band is somewhat "difficult" and "unclear", with serious language errors. There is hardly a complete sentence in the lowest level of essays, and language is fragmented and it is simply a "mess".

The expression of the rating scale is principled. In the actual rating process, raters encounter different kinds of essays that vary widely. It is necessary to judge the true level of an essay in a short period of time, which often requires a lot of practice. In general, due to lack of experience in scoring, new raters tend to look at an essay one-sidedly, and the scoring is not accurate enough, resulting in "out of band".

When raters rate an essay, they usually consider the following aspects: (1) The richness of content: whether it has relevance to the topic; (2) degree of fluency; (3) sentence patterns; (4) correctness of grammar; (5) correctness of vocabulary use. Raters should make a comprehensive judgment from these aspects and give an overall impression score. Experienced raters often use more than one standard as the basis for scoring, while new raters tend to attach great importance to one of them and make one-sided judgment on the level of the writing, resulting in scoring deviation.

3.2 Analysis of Several Common Types of Composition

There are many kinds of composition. In the following we will only analyze some common types of composition.

(1) High score composition

The sentence patterns of this level of essays are flexible and diverse, which read naturally and fluently. The use of conjunctions as well as substitutions, omissions, and anaphora can enhance the cohesion between sentences and the coherence of full texts. The examinee can use some advanced vocabulary correctly. There are sophistication of word choice and variety of sentence patterns.[6]

(2) Some essays are fluent and grammatical with few grammatical errors, but the use of sentence patterns and words are simple

This kind of composition should belong to the middle class, which shows that the examinees have no ability to master complex sentence patterns and control over advanced vocabulary. Some examinees try to use some complex sentence structures, but they are not organized properly and seemed a bit confusing.

(3) Some essays have many grammatical mistakes and poor coherence, but they use some advanced vocabulary

This shows that such students have a large vocabulary base and their reading ability may be better, so they can be given a middle level score. Some students try to use some advanced vocabulary, but the words are either misused or misspelled. This kind of essays cannot be classified as high-level composition.

(4) Essays with serious grammatical or spelling errors

Some teachers, when they see that there are serious grammatical mistakes or misspellings in the composition, cannot tolerate and give them very low scores. At this time, raters should check whether there are complete sentences in the text, whether they really express some ideas, and whether the score can be higher.

(5) Handwriting

Some students' handwriting is too scrawl to recognize. In the case of such examination papers, don't give low marks at will, but have patience to carefully identify its real level. Maybe it is a good paper. On the contrary, the handwriting is well done, and don’t be confused by this and give high scores by mistake.

In the actual marking process, the rater must judge and weigh a composition from various angles, master the characteristics of each level of composition, constantly sum up experiences and become an excellent rater.

4. Analysis of Related Date

The online marking system provides real-lime statistical analysis chart and related data (including score frequency curve) of the scores evaluated by the raters for each individual rater about his or her rating performance. These statistics are readily available to the directors and supervisors, who upon detecting the aberrant behaviors from the raters, will discuss with them the problems in their ratings.[7] The main statistical indicators are: related coefficient (R), average (Avg), standard deviation (Std), subject-object ratio (P), rating speed (Speed), integration (Integration) and scores distribution. These data not only provide support for the quality control of the examination organization, but also facilitate the self-test and correction of the rater. Raters can timely correct the deviation according to the feedback data and strive to improve their own scoring indicators. These statistical parameters are analyzed as follows:

4.1 Related Coefficient (r)

The related coefficient reflects the correlation between the true score of the writing and the true level of the examinee, that is, the quality of the marking. The related coefficient in CET4 online marking system mainly reflects the collaborative relationship between the objective average score and the subjective average score. Since there is a certain correlation between the various language abilities of examinees, the related coefficient between the score of the composition and the score of objective questions can be used as a reference index of evaluation reliability. For example, the related coefficient of a certain rater is low, but the average score is the highest in the whole team, it is very likely that the scores he/she gives are too high. The leader of the marking team should check the compositions he / she marks in time; if the score is high, the leader of the marking team should remain him / her and put forward suggestions.

Although the related coefficient R cannot directly reflect the quality of marking, it reflects the changes of the marking scales of the raters in large-scale tests, that is, the consistency of the scores the rater gives. The greater the absolute value of R, the higher the degree of correlation. Since the

objective scores have been marked by the machine, in general, if students' objective scores are high, the score of the composition should also be corresponding to them. There is a positive correlation between the two. Therefore, the rater with higher related coefficient is more accurate. The R value is the primary indicator to measure the quality of marking essays.[8]

4.2 Standard Deviation (Std)

Standard deviation (Std) reflects the degree of dispersion of the scores given by the rater which is the second important indicator for evaluating the quality of marking essays. If Std is too low (e.g. lower than 2), it indicates that the scores the rater gives is in the middle and do not disperse. The one who should be given a high score is not given, while the one who should be given a low score is given a higher score. In all kinds of examinations, the composition scores tend to be in the middle. Sometimes, because of the tight time and heavy task, in order to speed up, the scorer will give a score of an essay near the average score, so-called “secure score”, which will result in too low dispersion. If the Std value is low, the key to increase it is: raters should dare to give high score compositions, and those who should be given 0, 1, 2 points should also be resolutely given. Be sure to find the high score composition astutely. Imagine that a composition can get 13 points, but one rater gives it 10 points, it is unfair to students. If the value of the Std value is too high, the opposite is true, because essays that are not good enough have been given high scores. There are too many scores in high and low bands, and the method of reducing it is just the opposite.

4.3 Average (Avg)

The average score of each individual rater should be close to that of a team or marking center. If it is too high, it indicates that the rater is too loose; if it is too low, it indicates that the rater is too severe. The average score is very important to improve the related coefficient of the whole team, because every rater moves in a consistent manner and the order of the whole compositions will not be disordered.

4.4 Speed

Generally, the marking center has time requirements for the rater, such as the average marking time of each test paper does not exceed 50 seconds. Some raters are too slow in marking essays who are hesitant and vacillating in giving marks. They should know that rating slowly is not necessarily accurate. Only by mastering the basic characteristics of each level of composition can they make a quick and accurate assessment. Of course, it is not advisable to rate essays too quickly and irresponsibly.

At present, there are few studies on the speed of marking. Charney believes that monitoring and faster review speed will help to improve the reliability of the review, and holistic scoring should be fast and accurate.[4] Deep thinking of the text will reduce the reliability of evaluation.

4.5 Integration

The Integration index is the synthesis of the above indicators, of which the related coefficient accounts for the largest proportion (50%), followed by the Standard deviation (Std) accounting for 30%, and the rest accounting for 20%. The Integration index reflects the overall marking quality of a rater. To improve the Integration index, the quality of each sub-item must be improved.

4.6 Distribution Graph

The regrading software will also provide a personal score distribution graph, which reflects the distribution of the scores the rater gives. The scores of all the compositions should obey Linacre distribution, and the graph is preferably a smooth bell-shaped curve. The distribution curve of some raters is zigzag, such as 10 points more than 9 points. Some are rocket-shaped, such as giving too many 6 points, prominent upward. These are all caused by bad scoring habits. The distribution graph reflects the scoring habits of a rater. The team leader should help the team members to analyze timely, so that the team members can understand and correct their bad grading habits.

5. Conclusion

Raters should be responsible for each examinee, read every composition carefully and strive to be objective and fair. At the same time, they should constantly sum up the experiences of marking essays so as to improve the accuracy and credibility of the marking. The team leader or marking center should also supervise and guide the raters according to the statistical data and help them correct the deviation in the marking timely. The examination institutions should also strengthen the training and management of raters, cultivate a group of excellent raters, and establish a stable, experienced and professional rater team. In this way, the quality of marking essays can be fully guaranteed, and the examinee can get an accurate and fair score.

References

[1] H. Li, Effects of Rater-scale Interaction on EFL Essay Rating Outcomes and Processes, Zhejiang University, 2012.

[2] J.L. Chen, Assessment Construct in Foreign Language Teaching: the Case of Chinese Assessors of High-stake Exam Essays Writing in English, Shanghai International Studies University, 2013.

[3] H.Z.Wang, Rater perceptions of factors that affect the rating of TEM-4 oral test, CELEA Journal, vol.30, no.2, 15-21, 2007.

[4] D. Charney, The validity of using holistic scoring to evaluate writing: A critical overview. Research in the Teaching of English, vol. 18, pp. 65-81, 1984.

[5] L.F. Bachman, Language Testing-SLA Research Interfaces, Cambridge: Cambridge University Press, 1998, pp.19-22.

[6] Y. Lu, Rater Bias Studies in Online TEM4 Essay Marking, Shanghai International Studies University, 2010.

[7] J.L. Chen, Verification of validity of essay scoring standards for large-scale English exams, China Exam, vol.1, pp.21-25, 2016.

[8] Y. Lu, Research on the fairness of writing test-a review of the biased scorer, Foreign Language Testing and Teaching, vol. 4, pp.9-11, 2011.