The Principles of Constructing a Rating Scale for Diagnostic Writing Assessment

Jie Zhang

School of Foreign Studies, Xi'an University, Xi'an 710065, China.

zhangoutstanding@qq.com

Abstract. Diagnostic writing assessment is significant in learning and teaching writing as it can provide meaningful feedback information to help teachers recognize their students' strengths and weaknesses. But the popular rating scales are mostly confined to the realization of the performance assessment purpose only. Thus, it is necessary to study the related fields concerning a scale for the diagnostic purpose. By reviewing the popular rating scales nowadays, this paper explores the principles of constructing a rating scale for diagnostic writing assessment. Four principles, including reliability, validity, feasibility, and amenability, are put forward and are discussed in detail. This study is meaningful in probing into a relatively untouched field and providing useful guidance to the development of a diagnostic rating scale.

Keywords: Principles, rating scale, diagnostic writing assessment.

1. Introduction

Diagnostic writing assessment has been proved meaningful in learning and teaching writing as it measures students current knowledge, skills, and difficulties of writing during the learning process and it provides a foundation for the purpose of deciding a suitable teaching project for students' further learning. Thus, it is of great interest for researchers and teachers alike to have a framework to carry out diagnostic writing assessment. In practice, there is no established instrument for such kind of practice [1], which leads to a problematic choice that the rating scales for writing in exams like CET (College English Test) in China and TOEFL (Test of English as a Foreign Language) are often adopted for the purpose of diagnosing students' writing performance. But few of them have fulfilled the role of diagnosis.

Indeed, little work has been done to provide a kind of valid instrument for diagnostic concern. Some important issues, including the principles of constructing a rating scale for diagnostic writing assessment, the process of developing such a scale, and so on, have got little attention. Therefore, it is of necessity to explore this field. From that understanding, this paper tends to discuss one of the most important subjects: the principles of constructing a rating scale for diagnostic writing assessment.

2. Existing Rating Scales for Writing

To get a clear picture about the present situation of rating scales for writing, the rating scales for writing which are widely used nowadays are discussed. There are all kinds of assessments with writing section nowadays. Among them, CET in China and TOEFL are chosen for this study.

The Rating Criteria for writing of CET is employed for this research and the English version of it (translated by the author himself) is present in Table 1.

The Rating Criteria follow a holistic rating method which requires raters to provide a score by the impression from reading the paper. The total score is 15 and has been classified into 5 levels: 14 (13-15), 11 (10-12), 8 (7-9), 5(4-6), 2(1-3). Although the scale is widely used in China today, there are some shortcomings. One problem is that the descriptors are with less detailed information. It focus mostly on the grammatical aspect rather than the discourse aspect; and the weight of each factor is not clearly stated. Another problem is that the description is a little vague, which causes problems while raters are evaluate students' papers. [3]

The writing section in TOEFL is composed of two sections: the independent writing and the integrated writing. As this research limits its scope in writing only, the Independent Writing Rubrics is cited in Table 2.

Table 1. Rating criteria for writing of CET [2]

Rank	Rating Criteria
14	An essay addresses the topic. Ideas are expressed clearly and the paper runs smoothly and coherently with few and small language errors.
11	An essay addresses the topic. Ideas are expressed clearly and the paper runs smoothly and coherently with a few language errors.
8	An essay addresses the topic to a large extent. Ideas are not expressed clearly enough and the paper does not run so smoothly and coherently with a few language errors and some are serious .
5	An essay addresses the topic to some extent. Ideas are not expressed clearly and the paper does not run smoothly and coherently with lots of serious language errors.
2	Ideas are ill expressed poorly and the paper runs terribly. Language is fragmented or most sentences are erroneous, and most of them are serious mistakes.

Table 2. Independent writing rubrics [4]

Score	Task Description
5	An essay at this level largely accomplishes all of the following: Effectively addresses the topic and task Is well organized and well developed, using clearly appropriate explanations, exemplifications and/or details Displays unity, progression and coherence Displays consistent facility in the use of language, demonstrating syntactic variety, appropriate word choice and idiomaticity, though it may have minor lexical or grammatical errors
4	An essay at this level largely accomplishes all of the following: Addresses the topic and task well, though some points may not be fully elaborated Is generally well organized and well developed, using appropriate and sufficient explanations, exemplifications and/or details Displays unity, progression and coherence, though it may contain occasional redundancy, digression, or unclear connections Displays facility in the use of language, demonstrating syntactic variety and range of vocabulary, though it will probably have occasional noticeable minor errors in structure, word form or use of idiomatic language that do not interfere with meaning
3	An essay at this level is marked by one or more of the following: Addresses the topic and task using somewhat developed explanations, exemplifications and/or details Displays unity, progression and coherence, though connection of ideas may be occasionally obscured May demonstrate inconsistent facility in sentence formation and word choice that may result in lack of clarity and occasionally obscure meaning May display accurate but limited range of syntactic structures and vocabulary
2	An essay at this level may reveal one or more of the following weaknesses: Limited development in response to the topic and task Inadequate organization or connection of ideas Inappropriate or insufficient exemplifications, explanations or details to support or illustrate generalizations in response to the task A noticeably inappropriate choice of words or word forms An accumulation of errors in sentence structure and/or usage
1	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, or questionable responsiveness to the task Serious and frequent errors in sentence structure or usage
0	An essay at this level merely copies words from the topic, rejects the topic, or is otherwise not connected to the topic, is written in a foreign language, consists of keystroke characters, or is blank.

Compared with the rating scale of CET, TOEFL's WRITING Rubrics are more complicated. But it is criticized "for being imprecise, for using impressionistic terminology...and for often resulting in holistic assessment. [It implies that popular rating scales] might not be effective when used for

diagnosis, as they might not be able to identify strengths and weaknesses, and thus would result in feedback which is not informative for students."[5]

Nowadays, most researchers, teachers, and students use the existing scoring criteria of writing for diagnostic purposes. But owing to the limitations stated above, those criteria cannot be a sound and reasonable base for the diagnostic purpose of teaching and learning writing. Thus, it is easy to reach the conclusion that a valid scoring criterion for diagnostic writing assessment is needed and the study on this subject should be addressed.

3. The Principles in Building A Rating Scale for Diagnostic Writing Assessment

As it has been discussed in previous sections, the field of the rating scales for diagnostic writing assessment is relatively untouched and various aspects of that field need to be done. Among them, the proposal of a set of sound principles should be firstly explored as they are beacons that guide the following studies.

Based on such assumption, this paper focus on the construction of the principles of the rating scales for diagnostic writing assessment. Based on a review on the merits and demerits of the previous rating scales for writing, four principles (reliability, validity, feasibility, and revisability) are put forward and discussed.

3.1 Reliability

First, as rating scales nowadays are not considered to be adequate and sound measures of the multifaceted nature of writing[6], a more sound rating scale should trace the changing understanding of language and advanced linguistic theories. In doing this, the rating scales chosen can reflect the current understanding of language, especially writing. The theories that provide a reasonable basis for the development of rating scales should be tested and validated intending to provide more insights of the adoption of scoring criteria.

Second, the rating scale should be made up from empirical studies. There are two ways to develop a rating scale: one is intuition-based and the other is data-based.[7] But the intuition-based scales are criticized for the descriptors of the rating scales are based on very little empirical evidence.[8] Ideally, the descriptors should be empirically-developed with reference to learners.

In summary, the scale should be developed with a solid theoretical foundation with sound empirical evidence from field studies.

3.2 Validity

First, as one person develops any scales, he must bear in mind what he intends to acquire from those instruments. A scale for diagnosis is different from a scale for placement or proficiency and it is designed to testify the strengths and weaknesses of a leaner. Thus, in composing a rating scale for diagnostic writing assessment, the objective that the scale should facilitate a comprehensive analysis and present a report with details to provide meaningful feedback must be clearly illustrated.

Second, as a diagnostic assessment is to provide feedback on which the future learning is taking place, the coverage of the rating scale must be extended as widely as possible. Dimensions, like content, coherence, mechanics, and so on, should be considered and put into an overall framework. Only in this way can the feedback be meaningful and comprehensive to provide as much information as possible, and can students benefit more.

In summary, the validity of rating scales for diagnostic writing assessment can be established only if the purpose for which the scales were designed for can be realized and the comprehensive and meaningful feedback can be provided.

3.3 Feasibility

One thing that must be born in mind is that a rating scale developed for any purpose, including for diagnostic purpose, is intended to be used by raters, researchers, teachers, and even students sometimes. Thus, it implies that the rating scale should be user-friendly and practical in the rating

process. Raters can be guided through the rating process and realize the goal of diagnosis efficiently and effectively.

In the past years, different dimensions of writing has been illustrated and different writing scales has covered various factors due to different focus. Some scale are rather simple. The new rating scale for TEM 8 (Test for English Majors-Band 8) covers only three dimensions: Ideas and Arguments, Language Use, and Mechanics.[9] But a more complicated rating scale covering nine dimensions is developed by Knoch in 2009 covers traits including Accuracy, Lexical complexity, Data description, Data interpretation, Data-Part 3, Hedging, Paragraphing, Coherence, and Cohesion.[10] As a broader scale can cover more fields in writing and provide more information for diagnosis, a diagnostic writing scale with more different dimensions can provide much better diagnostic results.

In facilitating raters' scoring process, each dimension should be ranked. Clear ranks, such as the ranks in TOEFL's writing rubrics ranging from 0 to 5 clearly, should be employed. The descriptors should be clearly stated that subjective interpretations may take place if the impressionistic terminology is used in the criteria.[11] To avoid ambiguity and to cause less subjective personal judgement, vague descriptors, like 'better' , 'good', or "bad", should be eliminated from those scales. To realize the diagnostic purpose, all descriptors in a rating scale should offer detailed and accurate descriptions to reveal the nature of writing features at each rank.

In summary, to realize the purpose of diagnosis, the scale should cover more dimensions with sound theoretical background and each dimension should be ranked in different levels clearly with detailed and precise descriptors.

3.4 Revisability

Once a writing scoring criterion for diagnostic purpose has been set up, it will face different contexts as students are changing, teachers are different, and linguistics and education theories are being pushed forward. Because this is a dynamic world that everything changes, the scoring criterion for diagnostic writing assessment must be adapted to fulfill specific tasks in specific contexts with specific subjects.

As situations may change, a set of criteria cannot work to its idealization if things changes. A scale works in the United Stats where English is the first language may not work in China where English is a foreign language. So, it should be born in mind that the criterion for diagnostic purpose of writing should be modifiable and necessary amendments should be made according to authentic situations with different subjects.

4. Summary

To construct more valid and reliable ratings for a diagnostic purpose, rating scale for writing assessment arranging in clear levels with detailed and clear-cut descriptors should be developed with a reasonable theoretical foundation based on convincing empirical studies.

4.1 Results

With a probe into the rating scales presently, this paper has figured out four principles for constructing rating scales for diagnostic writing assessment. They are reliability, validity, feasibility, and revisability.

4.2 Significance

This study is significant theoretically and practically. This study does an important contribution to the writing assessment literature and fuels the development of theory in the field of diagnostic writing assessment. And four principles are interpreted and put forward, which are meaningful and practical in composing a rating scale for the purpose of evaluate learners' writing performance as these principles can provide a clear guidance.

4.3 Suggestions for Future Study

As this study focus itself solely on the principles of constructing a scoring criterion for diagnostic writing assessment, several related topics — the dimensions of the criteria, the weights of each dimension, the construction process, and the similarity and difference of of the principles in this paper with the principles of other kinds of writing scoring criteria for different purpose — have not been discussed. It is strongly suggested that those topics should be studied in the future.

Acknowledgements

References

[1]. Feng Lei. On the Development Principles of Rating Criterion of Diagnostic Writing Assessment for College English Learners: A Case Study of Rating Criterion of EFL Writing Assessment. Journal of Beijing Jiao tong University (Social Sciences Edition). Vol. 12 (2013) No. 1, p. 116-122.

[2]. National College English Test Committee. Syllabus of the National College Test Band 4 and 6. http://www.cet.edu.cn/file_2016_1.pdf.

[3]. Xi Fee, Yun Zhao. An Analysis of Problems of the Scoring Criteria of CET-4. Foreign Language Learning Theory and Practice. Vol. 4 (2008) No. 4, p. 45-52.

[4]. Education Testing service (ETS). TOEFL Writing Rubrics. Https: // www. Ets. Org/ s /toefl/ pdf/ toefl- writing- rubrics. Pdf.

[5]. Koch, U: Diagnostic writing assessment: the development and validation of a rating scale. Doctor of philosophy. The University of Auckland, New Zealand, 2007, p. 1-64.

[6]. Wiggle, S. C. Assessing writing. Cambridge: Cambridge University Press, 2002, p. 138.

[7]. Filcher, G. & Davidson, F. Language testing and assessment. New York: Routledge, 2007, p. 98.

[8]. Alderson, C Bands and scores. In C. Alderson & B. North (Eds.), Language testing in the 1990s: The communicative legacy. London: Modern English Publications/ British Council/ Macmillan, 1991, p. 71-86.

[9]. Jingling Yan. An analysis on the feedback of the new writing scoring scale for TEM 8. Foreign Language Testing and Teaching. Vol. 4 (2012), p. 6-14.

[10].   Koch, U. Diagnostic assessment of writing: A comparison of two rating scales. Language Testing. Vol. 26 (2009) No. 2, p. 275-304.

[11].   Watson Todd, R., Thienpermpool, P., & Keyuravong, S. Measuring the coherence of writing using topic-based analysis. Assessing Writing. Vol. 9 (2004), p. 85-104.