3rd International Conference on Culture, Education and Economic Development of Modern Society (ICCESE 2019)

Linking China's Standard of English Language Ability with the Automated Essay Assessment*

Wei Ye
School of Foreign Languages and International Business
Guangdong Mechanical & Electrical Polytechnic
Guangzhou, China

Dandan Huang
School of Foreign Languages and International Business
Guangdong Mechanical & Electrical Polytechnic
Guangzhou, China

***Abstract*—The purpose of this two-stage study is to examine the potential of China's Standard of English Language Ability (CSE) for Automated Essay Scoring (AES) development. In the first stage of the study, 657 descriptors from 24 writing-related CSE scales were inspected to generate categories representative of EFL learners' academic writing ability. In the second stage, the findings were compared to e-rater construct composites. Similarities between the CSE and e-rater categories were evident in terms of the overall emphases on the grammatical and textual dimensions as well as the coverage from the basic concepts like mechanics to the more abstract terms such as style. The comparison also pointed to great difference. The CSE categories presented positive descriptions of language use for learning/teaching purpose, while e-rater focused on language errors out of scoring concern. E-rater offered more explicit and operational interpretations of surface linguistic features. This paper not only supplies validation evidence of the CSE scales, but also demonstrates their usefulness for developing a customized AES system for EFL writers in China.**

Keywords—China's Standard of English Language Ability (CSE); Automated Essay Scoring (AES) system; e-rater

I. INTRODUCTION

Language proficiency scales offer useful guidelines for language teaching and assessment practice. A great number of language proficiency scales were developed in the last century, such as the Interagency Language Roundtable (ILR), the International Second Language Proficiency Ratings (ISLPR), the Euro centers Scale of Language Proficiency, the Canadian Language Benchmarks (CLB). It is the Common European Framework of Reference (CEFR), published in 2001 by the Council of Europe, which comes to be recognized as a key reference point for language education practice around the world. Opinions diverged on whether the CEFR could also be useful for guiding and supporting language education beyond the European continent (Yoneoka, 2011; Zhao et al., 2017). The China's Standard of English Language Ability (CSE) released in the 2018 is a tailor-made general framework for English teaching, learning and testing in China. As many as 86 scales are presented to describe various dimensions of learners' language use in different EFL scenarios. Most of these scales follow a descriptive pattern of nine levels, supplying detailed information regarding how EFL learners' proficiency evolved over the elementary, intermediate and advanced stages. The CSE is expected to contribute to the innovation of the Automated Essay Scoring (AES) system, as features characterizing EFL learner's contextualized language use have been specified. To take full advantage of the CSE, it becomes imperative for AES developers to examine scales that are of potential importance to writing assessment.

II. CHINA'S STANDARD OF ENGLISH LANGUAGE ABILITY

English is considered a crucial ladder leading to higher education. The problem is China is a large country with imbalanced economic developments in various regions, and students learn and use English under different educational and social conditions. For learners at the developing stage, that is high school graduates and college students, they are expected to make use of the English they have learned in the elementary stage to handle the cognitive-demanding academic writing assignments in the college and beyond. The CSE could offer a starting point for assessing their academic writing for two reasons. On the one hand, the scales and descriptors were grounded upon empirical studies to provide a universal reference framework for the English teaching, learning and testing practices at different educational stages and in various teaching institutes (Liu & Peng, 2018), thus efficient for assessing the writing abilities of all EFL learners in China. On the other hand, it draws upon theoretical models(Bachman & Palmer, 1996, 2010; Brown & Yule, 1983) and scientific calibration and validation process well-tested in the CEFR building, hence offering a set of prescriptive standards in line with the global standards and practice(Liu & Han, 2018), which is of special importance for academic writing. Therefore the potential of the CSE, with its 86 scales and over 5000 descriptors, in offering a suitable and precise evaluation of EFL learners' academic writing ability is worth investigating.

III. AUTOMATED ESSAY SCORING AND E-RATER

AES provides a reliable, accurate, and efficient feedback for the writing performance(Bennett, 2006). While human

*This research was supported by the Philosophy and Social Sciences of Guangzhou City Planning Project (2018GZGJ110) as well as the Innovative Research Team Project and the Start-up Fund for Distinguished Scholars of Guangdong Mechanical & Electrical Polytechnic

raters rely on their understandings of the rating scale and repeated practices, AES draws on machine learning techniques for computing essay scores on a number of language variables. A set of clearly-defined linguistic features reflecting the focal writers' proficiency features is of crucial importance for realizing its full potential in providing quality assessment for writing.

Among the five major working AES systems, including PEG (Project Essay Grader), IEA (Intelligent Essay Assessor), e-rater, Intellimetric, and Besty (Bayesian Essay Test Scoring System), e-rater is chosen as the subject of comparison for its global impacts and explicit framework of a set of hierarchical language features.

E-rater, developed by the Educational Testing Service (ETS) in the 1990s, aims to support the high-stakes tests like the Graduate Management Admission Test (GMAT) and Test of English as Foreign Language (TOEFL). Each year, millions of English learners around the world take part in these English proficiency tests to demonstrate their English proficiency for application to the English-medium higher-education institutes. Over the years, researchers managed to aggregate a large set of measures, namely, linguistic indices useful for discriminating writing performance across different proficiency levels, into a small set of eight more recognizable categories (Quinlan et al., 2009). At present, a total of eight construct composites are used to evaluate academic writing of EFL learners' academic writing ability, including grammar, usage, mechanics, style, organization, development, lexical complexity and content, each comprising its own subordinate variables for measurement. This categorization offers a basis for comparison with the categories emerged from the CSE.

IV. METHODS

To explore the CSE and its implications for automated essay scoring of EFL students' academic writing, this paper presents a two-stage study that first identified the categories characterizing EFL learners' writing performance and then examined its validity as AES measures against the working construct composites used in the e-rater. The first study is a grounded analysis of CSE scales. Grounded theory is useful for building up a theory based on a thick description of observed and collected data (Corbin & Strauss, 2014), and thus considered as the most influential qualitative data analytic approaches developed in the twentieth century (Somekh & Lewin, 2005). 24 scales related to writing were selected from the CSE to generate a pool of data consisting of 657 descriptors. In the second part of the study, finding of the above grounded analysis was compared to linguistic index used in the e-rater via thematic analysis.

V. RESULTS AND DISCUSSION

In the first stage of the study, 24 scales, the title of which indicated connections with writing, were first inspected, and a total of 657 descriptors were examined to obtain statements directly connected to the textual quality of academic writing. That is to say, descriptors related to non-academic writing activities (e.g. narration) and cognitive process that could hardly be traced from print (e.g. planning) were excluded. The remaining 404descriptors generated 87 open coding categories, which were then grouped into 29 axial categories and summarized into seven selective coding categories, namely, lexical knowledge, syntactic knowledge, productions skills, composing skills, organization, connection and style, as demonstrated in "Table I".

TABLE I. THE AXIAL CODING SCHEME

Axial coding	Selective coding
Word formation Lexical accuracy (e.g. use of pronouns, connectives, terminology) Lexical diversity Lexical complexity	Lexical knowledge
Syntactic accuracy (e.g. structural sequence, subject-verb agreement, tense, voice, third-person singular form) Syntactic diversity Syntactic accuracy	Syntactic knowledge
Punctuation Spelling Capital letters	Production skills
Furnishing details and examples Reader awareness reinforcing ideas	Composing skills
Emphasizing/highlighting the main idea Global coherence Connecting sentences Connecting paragraphs Presenting ideas in a logical way Forming a consistency stance Powerful start/ending	Organization
reporting others' ideas refuting/supporting others' ideas commenting others' ideas summarizing others' ideas	Connection
Rhetorical devices discourse pattern Genre knowledge Collocation Nominalization	Style

Findings indicated that the CSE covers a range of dimensions significant for language assessment, suggesting that it is a suitable starting point for evaluating EFL writing performance. Grammatical features like lexical and syntactic features essential for evaluating proficiency levels of the EFL learners (Weigle, 2002) were included. Categories associated with writing quality also emerged, covering significant textual/discoursal aspects (Grabe & Kaplan, 1996), such as production skills, composing skills, organization, style, as well as connection, a dimension key to successful source-based writing (Plakans & Gebril, 2017). One of the characterizing features of EFL learners is their imbalanced development of language proficiency. The CSE scales and descriptors presented a comprehensive reference framework for examining those specific and minute language use features, and some might be unique for Chinese students due to the great difference between their mother tongue and the target language (Lei, 2016).

In the second stage, the above findings were compared to the eight e-rater construct components, including grammar, usage, mechanics, style, organization, development, lexical complexity and topic-specific vocabulary usage.

Similarities between the CSE and e-rater categories were evident in terms of the focus on both the grammatical and textual dimensions. These two categories also offer a full coverage from basic concepts like the mechanics (as in the e-rater) and production skills (as in the CSE) to the more abstract one such as the style.

The comparison also pointed to great difference. The CSE categories presented positive descriptions of language use for learning/teaching purpose, while e-rater tended to present language error out of scoring concern, for instance, the category of usage comprising statements like article error, faulty comparisons. E-rater, as a working AES model, offered more explicit and operational interpretations of surface linguistic features. For instance, it defined style in terms of faulty performances like repetition of words, too many long sentences, and so on. Meanwhile, some subordinate categories for style, for instance, genre knowledge, in the CSE remained conceptual. Admittedly, the CSE furnished a variety of specific linguistic features confusing to the EFL learners, such as the subject-verb agreement and the use of connectives. It also pointed to directions for AES development, as the category of connection demonstrated the importance of assessing the quality of source use in academic writing.

This part of the research provides further evidence for the validity of the CSE as a guideline document for language testing and its unique value for highlighting language dimensions significant for assessing EFL writers.

VI. Conclusion

This paper demonstrates that the CSE scales supply a solid foundation for describing Chinese EFL learners writing proficiency, as the key dimensions for writing assessment are fully covered. The comparison with e-rater construct components for AES purpose further illustrates its potential for establishing a customized AES system for Chinese students' academic writing.

This paper presents a preliminary attempt for linking the CSE scales and descriptors with AES development. Researchers are encouraged to take into consideration of the features characterizing EFL learners' language proficiency presented in the CSE in developing a customized automated scoring system for EFL writers in China in the future. Limitations in the research, however, have to be acknowledged. Since this is a purely qualitative study, any conclusions must be viewed as tentative. Further research involving students and quantative research design would serve to provide stronger evidence for the validity of the CSE in language measurement. Furthermore, attempts should be made to operationalize the useful CSE concepts for practical application, so that it could truly serve as the starting point and guideline for language teaching, learning and testing.

References

[1] Bachman, L. F., Palmer, A. S. Language testing in practice: Designing and developing useful language tests. Oxford University Press, Oxford. 1996.

[2] Bachman, L. F., Palmer, A. S. Language assessment in practice. Oxford University Press, Oxford. 2010.

[3] Bennett, R. E. Moving the field forward: Some thoughts on validity and automated scoring. Automated scoring of complex tasks in computer-based testing, 403-412. 2006.

[4] Brown, G., Yule, G. Discourse Analysis Cambridge University Press, Cambridge. 1983.

[5] Corbin, J., Strauss, A. Basics of qualitative research: Techniques and procedures for developing grounded theory. Sage Publications, Inc., Thousand Oaks, CA. 2014.

[6] Grabe, W., Kaplan, R. B. Theory and practice of writing: An applied linguistic perspective. Longman, New York. 1996.

[7] Lei, X. Understanding writing strategy use from a sociocultural perspective: The case of skilled and less skilled writers. System, (60), 105-116. 2016.

[8] Liu, J., Han, B. Theoretical considerations for developing use-oriented China's Standards of English. Modern Foreign Languages, 41(1), 78-90. 2018.

[9] Liu, J., Peng, C. Constructing a scientific China's Standard of English. Foreign Language World, 2-9. 2018.

[10] Plakans, L., Gebril, A. Exploring the relationship of organization and connection with scores in integrated writing assessment Assessing Writing, 31, 98-112. 2017.

[11] Quinlan, T., Higgins, D., Wolff, S. Evaluating the Construct Coverage of the e-rater® Scoring Engine, ETS Research Report. 2009.

[12] Somekh, B., Lewin, C. Introduction to part II: Listening, exploring the case and theorizing. Research Methods in the Social Sciences. London: Sage Publications. 2005, p. 15.

[13] Weigle, S. C. Assessing Writing. Cambridge University Press, Cambridge. 2002.

[14] Yoneoka, J. From CEFR to CAFR: Place for a Common Framework of Reference for Languages in the East Asian Business World? Asian Englishes, 14 (2), 86-91. 2011.

[15] Zhao, W., Wang, B., Coniam, D., Xie, B. Calibrating the CEFR against the China Standards of English for College English vocabulary education in China. Language Testing in Asia, 7(1), 5. 2017.