Scoring Second Language Speaking Performance: Exactness or Fuzziness?

Tan Jin, Barley Mak and Li Liu

Faculty of Education, The Chinese University of Hong Kong

tjin@cuhk.edu.hk, barleymak@cuhk.edu.hk, refilane@gmail.com

Abstract

In the development of testing second language speaking, scoring speaking performance has reflected an attempt at "exactness" during different stages (see Jin, Mak, & Zhou, 2011), with automated scoring of speaking performance being greatly developed (see Xi, 2010). The paper first reviews research and practice of automated scoring, represented by *Versant* (Bernstein, Van Moere, & Cheng, 2010) and *SpeechRater* (Xi, Higgins, Zechner, & Williamson, 2008). Then, the paper discusses the relation between "exactness" featured by the automated scoring and "fuzziness" revealed from the human scoring. Finally, the paper suggests an integration between "exactness" and "fuzziness" in scoring speaking performance.

Keywords

speaking tests, automated scoring, human scoring, exactness, fuzziness, integration

1 Introduction

Testing second language speaking first attracted focus since the Second World War (Fulcher, 1997) and attained development afterwards, both from theoretical and practical perspectives (*see* Fulcher, 2003; Luoma, 2004). Motivations to push for testing speaking possibly originate from two factors as McNamara (1996) argues in the development of performance assessments: (1) need to evaluate with specific purposes, and (2) growing focus in learning, teaching and using language to communicate. Since its very early development during the Second World War, performance tests have been "by definition criterion referenced" (Fulcher, 2008: 157), and as such, candidates' performances are linked to quantitative scores for test interpretation and use. Scoring language performance is therefore essential for all second language speaking tests. During different historical stages, scoring speaking performance has reflected an attempt at "exactness" (Jin, Mak, & Zhou, 2011). As a result, automated scoring of speaking performance has been greatly developed accordingly (see Xi, 2010).

2 Automated scoring

Automated scoring of constructed response tasks is a new focus of research in language testing and assessment, being promoted by advances in natural language processing (NLP) and automatic speech recognition and processing technologies (Chapelle & Chung, 2010). Computer technologies has had a great impact on language assessments (see Chapelle, 2008; Chapelle & Douglas, 2006); however, until in recent years, language learning and language testing and assessment theories have began to gradually influence the research and practice of automatic scoring and feedback systems (Xi, 2010).

Automatic scoring of writing performance has reported encouraging results (see Enright & Quinlan, 2010; Lee, Gentile, & Kantor, 2010; Weigle, 2010). The E-rater has been adopted to score writing tasks in TOEFL iBT together with human raters (see ETS website, retrieved on 22 June 2011), although the scoring software rests on analyzing low-level linguistic features (Quinlan, Higgins, & Wolf, 2009). Subsequently, validation studies have primarily been based on correlations with human scores being the "gold standard" (Fulcher, 2010). Such kind of *complementary rating* design (Enright & Quinlan, 2010) is summarized as follows:

(1) Human raters use discrete integers but the E-rater applies a continuous scale, with the final score being the average of the two scores if the difference between the human score and the E-rater score is less than 1.5;

(2) A secondary human rating takes place if the difference is not less than 1.5, and the final score is the mean of the three ratings; if one score is 1.5 points or more from the other two scores, and the final score is the mean of the two closer scores;

(3) An adjudicated human rating will be used if each rating is 1.5 or more from its nearest neighbor, or if an essay is identified as off-topic by the human rater (such essays will not be sent for automatic scoring);

(4) Anomalous essays with too few words or sentences, excessively long essays or an excessive number of errors in grammar, usage or mechanics are scored by at least two human raters and not by the E-rater.

The development of automated scoring of speaking performance, however, is still in infancy when compared with the automated scoring of writing. It is mainly represented by *Versant* (Bernstein, Van Moere, & Cheng, 2010) and *SpeechRater* (Xi, Higgins, Zechner, & Williamson, 2008; Zechner, Higgins, Xi, & Williamson, 2009).

Versant. The *Versant* speaking tests applied constrained tasks to measure "facility" with a spoken language so as to predict speaking ability, reporting four sub-scores of sentence mastery, vocabulary, fluency and pronunciation with respective weights of 30, 20, 30 and 20 to an overall scores; concurrent validity studies showed a strong correlation (r=0.77 to 0.92) between scores from automated tests and scores from oral proficiency interviews (Bernstein, at al., 2010). The logic of *Versant* may be summarized into three points: first, "facility" with a spoken language including core skills of speaking ability is theorized; second, "facility" is assessed through using constrained tasks and measurable criteria; third, the scores of "facility" are used to predict the speaking ability with substantial correlations validating such predictions (see Bernstein, at al., 2010). It can thus be seen that "facility" serves as a bridge between the automated scoring and the speaking ability. However, using such logic to score speaking performance warrants attention. First, the interpretation of "facility" is hardly verified through theoretical or empirical evidence. The construct of facility refers to core skills completing all speaking tasks in every speaking context. It is not uncommon that while a student performs very well in some contexts (he or she obtains the facility using such logic), he or she is probably performing worse in another contexts. Second, the use of scores faces many difficulties. The greatest is possibly the query that "*can the scores be used for all purposes, for example, business and education*"? Third, the "correlation fallacy" (Kaulfers, 1944; also see Fulcher, 2010) must be dealt with caution. Consider the relationship between written tests and spoken tests as an example. Multiple choice items of written tests cannot be used to assess speaking proficiency, even though some written tests are highly correlated with spoken tests.

SpeechRater. The other attempt of automated scoring has been made by *SpeechRater,* to use extended speaking tasks, and it consists of three major components: (1) speech recognizer and feature generation programmes, (2) scoring model, and (3) user interface (Xi, et al., 2008; Zechner, at al., 2009). Technological details will not be mentioned here; our concern is how construct-relevant features contribute to the automated scoring. Based on features from the TOEFL iBT speaking rating scales (or scoring rubrics), which may also be realized computationally, a set of 29 initial features were originally computed, from which 11 features were selected for scoring model training (see Zechner, at al., 2009: 890). Two scoring models, *multiple regression* and *classification and regression trees* were considered. Five features of amscore (delivery, pronunciation), wpsec (delivery, fluency), tpsecutt (delivery & language use, vocabulary & fluency), wdpchk (delivery, fluency) and lmscore (language use, grammar) were eventually analyzed by *multiple regression* (the weights are 4, 2, 2, 1 and 1 respectively). Then, five features of amscore, wpsec, wdpchk, silmean (delivery, fluency) and lmscore were present in the optimal tree of *classification and regression trees.* Finally, *multiple regression* model was chosen for an operational system because of its simplicity and perspicuity. The correlation between machine scores and human scores on TOEFL practice online was 0.57 while on TOEFL iBT field study the correlation was 0.68. The *SpeechRater* was just applied in a low-stakes practice environment (Xi, et al., 2008; Zechner, et al., 2009).

SpeechRater is rooted in assumptions that the speaking proficiency construct is *multidimensional* and that these dimensions are highly *correlated* with each other (Sawaki, 2007; Xi & Mollaun, 2006; also see Xi, 2007). The cognitive studies of experts on scoring speaking (e.g. researchers, teachers and raters) demystify the composition of the speaking construct (e.g. Brown, 2006a; Brown, Iwashita, & McNamara, 2005). The development of discourse analysis of candidate' speaking performance has also greatly enhanced the feasibility of quantifications of features from actual discourse (e.g. Iwashita, Brown, McNamara, & O'Hagan, 2008). As such, certain computationally measurable features in meaningful dimensions of the speaking proficiency construct supported by

theoretical and empirical evidence are measured "to predict the score of a human rater" (Xi, et al., 2008: 77). The rationale for *SpeechRater* is twofold: first, it explores the meaningful dimensions of the construct to find measurable features (*multidimensional*) by using extended speaking tasks and analyzing human rating scales; second, quantifications of ***some*** measurable features are combined to the speaking scores (highly *correlated*).

3 Human scoring and automated scoring

Last decades have been witness to the rise of automated scoring in language performance tests, but the mainstream practice of scoring speaking performance still resorts to human scoring—trained and certified raters' judgements over the quality of a candidate performance based on prescribed rating scales (e.g. IELTS, see IELTS website, retrieved on 22 June 2011, also see Taylor & Falvey, 2007; TOEFL iBT, see Educational Testing Service, 2005, also see Chapelle, Enright, & Jamieson, 2008). As for human scoring, both cognitive studies of rater perception and discourse analyses of candidate performance have revealed that the "fuzziness" exists when assigning an exact score to a candidate's spoken performance (Brown, 2006a, 2006b; Brown, et al., 2005; Iwashita, et al., 2008). With respect to automated scoring, it is featured with exact scores, which are calculated and produced in terms of certain measured features of candidate performance.

In recent years, attempts have been made to complement human scoring with automated scoring. When scoring writing tasks of TOEFL iBT, human scoring and automated scoring are combined together, employing human scoring for "content and meaning" and automated scoring for "linguistic features" (ETS website, retrieved on 22 June 2011). However, there is, to date, little empirical evidence for justifying the use of scores produced by averaging the two scores from human raters and E-rater, for example, Level 3 (human rater) and 3.16 (E-rater). It is proposed that, in future studies, investigations of the relationship between the "exactness" of automated scoring and the "fuzziness" of human scoring should be conducted; attempts to integrate human scoring with automated score should also be made. First, exact values of computationally measurable features with relevant interpretation should be provided with human raters before making judgements. For example, *word tokens* of Candidate A are *n* (feature values), and a possible range of this value is *Level a* to *Level* b (interpretation). Second, human raters learn feature values with relevant interpretation and make judgment based on their scoring confidence to deal with the fuzziness of speaking performance (see Jin, et al., 2011). In the end, exact values of computationally measurable features can also be used to monitor the scoring quality of human raters.

4 Conclusion

Automated scoring of speaking performance is featured with producing exact values of computationally measurable features in meaningful dimensions of the speaking proficiency construct, for example, *Versant* (Bernstein, et al., 2010) and *SpeechRater* (Xi, et al., 2008; Zechner, et al., 2009). However, human scoring is still the mainstream practice of assessing speaking proficiency although the fuzziness exists in scoring speaking performance (Brown, 2006a, 2006b; Brown, et al., 2005; Iwashita, et al., 2008). An integration is suggested for consideration in future studies. Exact values produced by automated scoring with relevant interpretation can facilitate human raters' judgments regarding candidates' speaking proficiency—human raters may resort to their confidence for dealing with the fuzziness of spoken performances during this process (see Jin, et al., 2011). Finally, exact values can also be used to monitor the scoring quality of human raters.

References

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing*, 27(3), 355–377.

Brown, A. (2006a). An examination of the rating process in the revised IELTS Speaking Test. In P. McGovern, & S. Walsh (Eds.), *IELTS research reports 2006* (pp. 41-70), Canberra & Manchester: IELTS Australia and British Council.

Brown, A. (2006b). Candidate discourse in the revised IELTS Speaking Test. In P. McGovern, & S. Walsh (Eds.), *IELTS research reports 2006* (pp. 71-89), Canberra & Manchester: IELTS Australia and British Council.

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test taker performance on English-for-Academic-Purposes speaking tasks* (TOEFL Monograph No. 29). Princeton, NJ: Educational Testing Service.

Chapelle, C. A. (2008). Utilizing technology in language assessment. In E. Shohamy, & N. H. Hornberger (Eds.), *Encyclopedia of language and education*, 2nd Edition,

Volume 7: *Language testing and assessment* (pp. 123–134). New York: Springer Publishers.

Chapelle, C. A., & Chung Y. R. (2010). The Promise of NLP and speech processing technologies in language assessment. *Language Testing*, 27(3), 301–315.

Chapelle, C. A., & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge: Cambridge University Press.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York: Routledge.

Educational Testing Service. (2005). *TOEFL iBT tips: How to prepare for the next generation TOEFL test and communicate with confidence*. Princeton, NJ: Educational Testing Service.

Enright, M. K., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with E-rater® scoring. *Language Testing*, 27(3), 317-334.

ETS website. (22 June 2011). Understanding Your *TOEFL iBT*® Test Scores. Retrieved on 22 June 2011 from http://www.ets.org/toefl/ibt/scores/understand/

Fulcher, G. (1997). The testing of speaking in a second language. In C. Clapham, & D. Corson (Eds.), *Encyclopedia of language and education*, Vol.7: *Language testing and assessment* (pp. 75 – 85). Dordrecht: Kluwer Academic Publishers.

Fulcher, G. (2003). *Testing second language speaking*. London: Pearson Education Limited.

Fulcher, G. (2008). Criteria for evaluating language quality. In E. Shohamy, & N. H. Hornberger (Eds.), *Encyclopedia of language and education*, 2nd Edition, Volume 7: *Language testing and assessment* (pp. 157-176). New York: Springer Publishers.

Fulcher, G. (2010). *Practical language testing*. London: Hodder Education.

IELTS website. (22 June 2011). Score processing and reporting. Retrieved on 22 June 2011 from http://www.ielts.org/researchers/score_processing_and_reporting.aspx

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, 29 (1), 24-49.

Jin, T., Mak, B., & Zhou, P. (2011). Confidence scoring of speaking performance: How does fuzziness become exact? *Language Testing*. Advance online publication (15 June 2011). doi: 10.1177/0265532211404383

Kaulfers, W. V. (1944). Wartime development in modern-language achievement testing. *Modern Language Journal*, 28(2): 136-150.

Lee, Y.-W., Gentile, C., & Kantor, R. (2010). Toward automated multi-trait scoring of essays: Investigating links among holistic, analytic, and text feature scores. *Applied Linguistics*, 31(3): 391-417.

Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.

McNamara, T. F. (1996). *Measuring second language performance*. London: Longman.

Quinlan, T., Higgins, D., & Wolf, S. (2009). *Evaluating the construct coverage of the E-Rater scoring engine* (ETS RR-09-01). Princeton NJ: Educational Testing Service.

Sawaki, Y. (2007). Construct validation of analytic rating scales in a speaking assessment: Reporting a score profile and a composite. *Language Testing*. 24 (3): 355-390.

Taylor, L., & Falvey, P. (Eds.). (2007). *IELTS collected papers: Research in speaking and writing assessment*. Cambridge: Cambridge University Press.

Weigle, S. C. (2010). Validation of automated scoring of TOEFL iBT tasks against non-test indicators of writing ability. *Language Testing*, 27(3), 335–353.

Xi, X. (2007). Evaluating analytic scoring for the TOEFL® Academic Speaking Test (TAST) for operational use. *Language Testing,* 24(2), 251-286.

Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing*, 27(3), 291-300.

Xi, X., Higgins, D., Zechner, K., & Williamson, D. M. (2008). *Automated scoring of spontaneous speech using SpeechRater*^{SM} *v1.0* (ETS RR-08-62). Princeton, NJ: Educational Testing Service.

Xi, X., & Mollaun, P. (2006). *Investigating the utility of analytic scoring for the TOEFL Academic Speaking Test (TAST)* (TOEFL iBT Research Report No. TOEFLiBT-01). Princeton, NJ: Educational Testing Service.

Zechner, K., Higgins, D., Xi, X., & Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, 51, 883-895.