Anchoring writing scores with candidates' performances: IELTS and TOEFL perspectives

Li Liu, Barley Mak and Tan Jin

Faculty of Education, The Chinese University of Hong Kong

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

Abstract

Since 2000s, growing attention has been attached to anchor writing scores with candidates' performances in L2 writing tests in order to contribute to validation of score interpretation and use (see Cumming *et al*., 2000; Shaw & Falvey, 2008). In the context of international language tests, discourse analysis has been widely used for linking the scores to candidates' performances. This paper focuses on two international language tests, namely, IELTS and TOEFL, to review studies undertaken for this purpose. In the end, the paper suggests further exploration of distinguishing writing features with support of automated essay scoring systems.

Keywords

writing scores, candidate performances, discourse analysis, IELTS, TOEFL

Introduction

Discourse analytic approach focuses on investigating written performances of candidates at different performance levels. This approach aims to examine the linguistic features of written responses at each level to justify and complement the construct underlying raters' impressionistic scoring. (e.g., Cumming & Mellow, 1996; Tedick & Mathison, 1995; Ishikawa, 1995).

International English Language Testing System (IELTS) and the Test of English as a Foreign Language (TOEFL) are the most well-established and widely practiced large-scale language assessments worldwide. Scores of these two tests are internationally recognized for academic or general purposes. Both TOEFL and IELTS researchers have undertaken a series of empirical studies to anchor scores and band levels with performances of L2 writers, intending to complement the more interpretive analyses previously pursued (e.g., Cumming *et al*., 2001, 2002, 2006), followed the assumptions of the *reader-writer* model (Cumming *et al*., 2000) by describing and evaluating the thinking and perceptions of experienced essay raters as they assigning scores to the written responses of candidates.

Methodologically, these studies usually adopt quantitative (Kennedy & Thorp, 1999; Frase *et al*., 1999; Cumming *et al*., 2006) or mixed-methods design (Mayor *et al*., 1999), thus providing part of the validity evidences for score interpretation and use.

1 IELTS studies

During the processes of IELTS revision, studies of analyses of candidates' writing performances directly provide insights into the continuing validity and standardization of the IELTS test.

Two groups of researchers conducted detailed investigation of written responses of test takers. Descriptive in nature, Kennedy and Thorp (1999) analyzed written responses of 130 candidates to investigate linguistic features of these scripts at three proficiency levels (8—expert user, 6—competent user and 4—limited user). WordSmith Tools program was employed to analyze selected writing features. Results revealed that higher level essays have internal coherence without overt cohesive ties, using appropriate register and show strongly-developed reader awareness. Level 6 essays present a better version of the level 4 essays, but they are substantially different from level 8 essays in terms of content, linguistic and discourse features examined. The study thus proved the distinguishing effect of band levels and established the link between scores and real candidate performances. However, instead of carrying out statistical tests of significance, the authors only reported percentages for all the data.

Moving a step forward, Mayor *et al*. (1999) examined the linguistic features of 186 candidate scripts which were categorized into high-scoring (a score of 7-8) and low-scoring (a score of 5) groups, attempting to establish the extent to which these features are associated with the band scores awarded for the task. Mixed methods design was

adopted to analyze various categories of writing features: linguistic analysis of errors (spelling, punctuation, grammar, preposition, and lexis/idiom), sentence structure (t-unit types) and argument structure at the sentence level (Theme analysis) and discourse level (genres). Correlation coefficients were calculated between error categories while t-test was used for uncorrelated samples. Qualitative analysis was also used for categories employed in the application of Systematic Functional Grammar. Results found that high and low-scoring scripts were differentiated by a constellation of features. Features including word length, low formal error rate, complexity in sentence structure, and occasional use of the impersonal pronoun '*one*' are strong predictors of a high band score. In terms of functional features, Thematic structure, argument genre and some of more subtle ways of expressing the interpersonal tenor of the text are significant distinguishing features between task scores. The study provided support for claims about the discriminating power of the rating scales. The features differentiated effectively between high-scoring and low-scoring scripts and the strong predictors of high scores all appear in IELTS band descriptors at higher performance levels.

Different from previous studies by taking a developmental perspective, Banerjee, Franceschina and Smith (2004) addressed the question of how competence levels operationalized in a rating scale might be related to L2 developmental stages. The authors explored the defining characteristics of 275 test-takers' (Chinese and Spanish) written performances at each 3-8 band in terms of linguistic features, including cohesive devices; vocabulary richness; syntactic complexity and grammatical accuracy. Analysis of Variance (ANOVA), between group designs were used to test the main effects and potential interaction of band level with L1 on identified linguistic features for different tasks. The findings suggested that except the syntactic complexity, all other features contributed to scores of test takers. In addition, analysis of grammatical accuracy proved to be rather informative, which confirmed findings from previous literature on L2 development. The authors pointed out that future studies on features of levels of L2 proficiency should take account of the accuracy of grammatical categories such as Subject-Verb agreement and passives which proved good discriminators of level regardless of L1 and writing task. The study therefore established the link of candidates' performance measured by salient linguistic markers with different band levels.

2 TOEFL studies

Followed the *text characteristics* model (Cumming *et al*., 2000), TOEFL studies on features of writing performance also provide additional and complementary evidences about construct inherent in writing tasks and relevant rating scales. In the early stage of the TOEFL writing test revision, Frase *et al*. (1999) conducted a variety of text analyses of Test of Written English (TWE) essays to summarize and compare linguistic properties of 27 TWE essays at levels 3, 4 and 5 to determine how TWE scores could relate to linguistic text properties. The authors analyzed 106 variables for each essay employing statistical procedures including correlation, analysis of variance, discriminative analysis and factor analysis. Results indicated that major distinguishing features of academic writing include a nominal style, passive constructions and complexity of sentence structure. Two variables that can be measured unambiguously by computer—number of words and the average length of words—were quite predictive of TWE essay scores of nonnative English speakers. The study has two important contributions. First, relationship between essays scores and linguistic properties were established through more detailed statistical analysis. Second, the study developed a measure of essay content by comparing linguistic features of essays with those of essays at highest score level. However, it did not further explore the extent of different features measured contributes to the overall judged scores.

With the aim of prototyping new TOEFL iBT writing tasks, Cumming *et al*. (2006) analyzed the discourse features of 216 compositions written for 6 tasks by 36 ESL students at three levels (Levels 3, 4, and 5). Discourse analysis approach was adopted, including careful identification of writing features and objective measures. Nine indicators for discourse analysis were then identified: text length, lexical and syntactic sophistication, grammatical accuracy, argument structure, orientations to evidence, and verbatim uses of source text. Nonparametric Multivariate Analysis of Variance (MANOVA), following a 3-by-3 (task type by English proficiency level) within-subjects factorial design was used for analyzing the quantified writing features. The results showed that writing proficiency was significantly different in terms of text length, lexical sophistication, syntactic complexity, argument structure, grammatical accuracy, voice and message in source evidence. Results provided strong support for the scoring rubrics and levels for both the independent and integrated tasks for the new TOEFL in that these tasks can consistently distinguish features of examinees' written discourse across different levels.

In addition, statistical analysis adopted in the study can further explain the contribution of each feature to the overall score awarded.

3 Discussion and conclusion

To sum up, studies by IELTS and TOEFL document how discourse features of written responses varied with proficiency levels and tasks. Two important implications can be drawn from these studies. First, scores on writing tasks can be verified and anchored empirically through the analysis of discourse features that differentiate proficiency levels. Second, discourse analysis approach towards L2 writing is usually based upon quantifiable features of written compositions such as accuracy, complexity, coherence and cohesion. However, it limits in the way that it cannot conceptualize such non-linguistic aspects of L2 writing as content quality, originality, or creativity.

Furthermore, the reliable and robust measures assessed in studies investigating writing features based on candidates' performances may be complementary to automated scoring of compositions (see Shermis & Burstein, 2003). Though the lexical features analyzed already feature in many automated programs for scoring writing, but other important features such as the aspects of argumentation, voice in uses of source evidence, or modes of paraphrasing or summarizing source. A promising line of inquiry thus is to expand the coverage and make full use of features in automated essay scoring by judging its theoretical and practical relevance of them to the target construct of writing for moving beyond reliance on rater agreement as sources of evidence for score interpretation and use.

References

Banerjee, J., Franceschina, F., & Smith, A. M. (2004). Documenting features of written language production typical at different IELTS band score levels. *IELTS Research Reports, 7*, 249–309.

Connor, U., & & Carrell, P. L. (1993). The interpretation of the tasks by writers and readers in holistically rated direct assessment of writing. In J. G. Carson & I. Leki (Eds.), *Reading in the composition classroom: Second language perspectives* (pp. 141–160). Boston, MA: Heine and Heine.

Cumming, A. (1998). Theoretical perspectives on writing. In W. Grabe (Ed.), *Annual Review of Applied Linguistics, 18*, 61-79. New York: Cambridge University Press.

Cumming, A. (2001). ESL/EFL instructors' practices for writing assessment: Specific purposes or general purposes? *Language Testing, 18* (2), 207-224.

Cumming, A., & Mellow, D. (1996). An investigation into the validity of written indicators of second language proficiency. In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 72-93). Clevedon, UK: Multilingual Matters.

Cumming, A., Kantor, R., & Powers, D. (2001). *Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: An investigation into raters' decision making and development of a preliminary analytic framework.* (TOEFL Monograph 22). Princeton, NJ: Educational Testing Service.

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

Cumming, A., Kantor, R., Powers, D., Santos, T., & Taylor, C. (2000). *TOEFL 2000 writing framework: A working paper (TOEFL Monograph 18)*. Princeton, NJ: Educational Testing Service.

Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U., & James, M. (2006). *Analysis of discourse features and verification of scoring levels for independent and integrated prototype written tasks for the new TOEFL test (TOEFL Monograph 30)*. Princeton, NJ: ETS.

Frase, L., Faletti, J., Ginther, A., & Grant, L. (1999). *Computer Analysis of the TOEFL Test of Written English (TOEFL Research Report 64)*, Educational Testing Service, Princeton, NJ.

Grabe, W., & Kaplan, R. (1996). *Theory and practice of writing: An applied linguistic perspective*. London: Longman.

Kennedy, C., & Thorp, D. (1999). A Corpus-based Investigation of linguistic responses to an IELTS Academic Writing Task. In L. Taylor & P. Falvey (Eds.), *IELTS Collected Papers: Research in speaking and writing assessment. Studies in language testing (19)* (pp. 316-377). Cambridge, UK: Cambridge University Press.

Hamp-Lyons, L. (2003). Writing teachers as assessors of writing. In B. Kroll (Ed.), *Exploring the dynamics of second language writing*(pp.162-189).Cambridge:Cambridge University Press.

Ishikawa, S. (1995). Objective measurement of low-proficiency EFL narrative writing. *Journal of Second Language Writing, 4*, 51-69.

Mayor, B., Hewings, A., North, S., Swann, J., &

Coffin, C. (1999). A Linguistic analysis of Chinese and Greek L1 scripts for IELTS Academic Writing Task 2. In L. Taylor & P. Falvey (Eds.), *IELTS Collected Papers: Research in speaking and writing assessment. Studies in language testing (19)* (pp. 250–315). Cambridge, UK: 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.

Shaw, S., & Falvey, P. (2008). *The IELTS writing assessment revision project: Towards a revised rating scale*. Retrieved from http://www.cambridgeesol.org/assets/pdf/research_reports_01.pdf.

Shermis, M. D. & Burstein, J. (2003). *Automated Essay Scoring: A cross disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.

Tedick, D., & Mathison, M. A. (1995). Holistic scoring in ESL writing assessment: What does an analysis of rhetorical features reveal? In D. Belcher & G. Braine (Eds.), *Academic writing in a second language: Essays on research and pedagogy* (pp. 205-230). Norwood, NJ: Ablex.

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

Weigle, S. C. (2002). *Assessing writing*. Cambridge, UK: Cambridge University Press.