TOEFL-Related Publications (2000-Present)

Attali, Y., Bridgeman, B., & Trapani, C. (2010). Performance of a generic approach in automated essay scoring. *The Journal of Technology, Learning, and Assessment, 10* (3*).*

Barkaoui, K. (2014). Examining the impact of L2 proficiency and keyboarding skills on scores on TOEFL-iBT writing tasks. *Language Testing, 31*(2), 241-259.

Barkaoui, K., Brooks, L., Swain, M., & Lapkin, S. (2013). Test-takers' strategic behaviors in independent and integrated speaking tasks. *Applied Linguistics, 34*(3), 304-324.

Biber, D. (2003). Variation among university spoken and written registers: A new multi-dimensional analysis. In Meyer, C., & Leistyna, P. (eds.), *Corpus analysis: Language structure and language use*, 47-70. Amsterdam: Rodopi.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers.* Amsterdam: John Benjamins.

Biber, D. (2006). Stance in spoken and written university registers. *Journal of English for Academic Purposes, 5*(2), 97-116.

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes, 26*(3), 263-286.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), *371-405.*

Biber, D., Conrad, S., Reppen, R., Byrd, P., & Helt, M. (2002). Speaking and writing in the university: A multi-dimensional comparison. *TESOL Quarterly, 36*(1), 9-48.

Biber, D., Csomay, E., Jones, J. K., & Keck, C. (2004). A corpus linguistic investigation of vocabulary-based discourse units in university registers. In Connor, U., & Upton, T. A. (eds.), *Applied Corpus Linguistics: A Multi-Dimensional Perspective*, 53-72. Amsterdam: Rodopi.

Biber, D., Csomay, E., Jones, J. K., & Keck, C. (2004). Vocabulary-based discourse units in university registers. In Partington, A., Morley, J., & Haarman, L. (eds.), *Corpora and Discourse*, 23-40. Bern: Peter Lang.

Bridgeman, B., Powers, D., Stone, E., & Mollaun, P. (2012). TOEFL iBT Speaking Test scores as indicators of oral communicative language proficiency. *Language Testing, 29* (1), 91-108.

Bridgeman, B., Trapani, C, & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education, 25* (1), 27-40.

Brooks, L., & Swain, M. (2015). Students' voices: The challenge of measuring speaking for academic contexts. In B. Spolsky, O. Inbar, & M. Tannenbaum (Eds.), *Challenges for language education and policy: Making space for people* (pp. 65-80). New York: Routledge.

Carrell, P. L., Dunkel, P. A., & Mollaun, P. (2004). The effects of notetaking, lecture length, and topic on a computer-based test of ESL listening comprehension. *Applied Language Learning, 14,* 83-105.

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

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice, 29 (3),* 3-13.

Cho, Y., & Bridgeman, B. (2012). Relationship of TOEFL IBT™ Scores to Academic Performance: Some Evidence from American Universities, *Language Testing, 29*(3), 421-442.

Cho, Y., Rijmen, F., & Novák, J. (2013). Investigating the effects of prompt characteristics on the comparability of TOEFL iBT integrated writing tasks. *Language Testing, 30*(4), 513-534.

Chodorow, M., Gamon, M., & Tetreault, J. (2010). The Utility of Article and Preposition Error Correction Systems for English Language Learners: Feedback and Assessment [Special issue]. *Language Testing, 27*(3), 419-436 [Special issue].

Cohen, A. D., & Upton, T. A. (2004). Strategies in responding to the next generation TOEFL reading tasks. *Language Testing Update, 35,* 53-55.

Cohen, A. D., & Upton, T. A. (2007). I want to go back to the text: Response strategies on the reading subtest of the New TOEFL. *Language Testing, 24*(2), 209-250.

Cumming, A., Grant, L., Mulcahy-Ernt, P., & Powers, D. (2004). A teacher-verification study of speaking and writing prototype tasks for a new TOEFL. *Language Testing, 21* (2), 159-197.

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for the next generation TOEFL. *Assessing Writing, 10*(1), 5-43.

Cumming, A., Kantor, R., & Powers, D. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal, 86*(1), 67-96.

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

Futagi, Y., Deane, P., Chodorow, M., & Tetreault, J. (2008). A computational approach to detecting collocation errors in the writing of non-native speakers of English. *Computer Assisted Language Learning*, *21*(4), 353-367.

Gallagher, A., Bridgeman, B., & Cahalan, C. (2002). The Effect of Computer-Based Tests on Racial-Ethnic and Gender Groups. *Journal of Educational Measurement, 39,* 133-147.

Gu, L. (2014). At the interface between language testing and second language acquisition: Language ability and context of learning. *Language Testing, 31(1), 111-133.*

Hansen, E. G., Mislevy, R. J., Steinberg, L. S., Lee, M. J., & Forer, D. C. (2005). Accessibility of tests for individuals with disabilities within a validity framework. *System: An International Journal of Educational Technology and Applied Linguistics, 33*(1), 107-133.

Heilman, M., Cahill, A., & Tetreault, J. (2012, June) *Precision Isn't Everything: A Hybrid Approach to Grammatical Error Detection.* In Proceedings of the 7th Workshop on Innovative Use of Natural Language Processing for Building Educational Applications, Montreal, Canada.

Higgins, D., Burstein, J., & Attali, Y. (2006). Identifying off-topic student essays without topic-specific training data. *Natural Language Engineering*, *12*(2), 145-159.

Higgins, D., Xi, X., Zechner, K., & Williamson, D. (2010). A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech and Language.*

Israel, R., Tetreault, J., & Chodorow, M. (2012, June) *Correcting Comma Errors in Learner Essays, and Restoring Commas in Newswire Text.* In Proceedings of the 2012 Meeting of the North American Association for Computational Linguistics: Human Language Technologies, Montreal, Canada.

Kang, O., & Rubin, D. (2009). Reverse linguistic stereotyping: Measuring the effect of listener expectations on speech evaluation. *Journal of Language and Social Psychology, 28(4),* 441-456.

Kang, O., & Rubin, D.L. (2012). Intra-rater reliability of oral proficiency ratings. *International Journal of Educational and Psychological Assessment, 12(1),* 43-61.

Kang, O., Rubin, D., & Pickering, L. (2010). Suprasegmental Measures of Accentedness and Judgments of Language Learner Proficiency in Oral English. *Modern Language Journal, 94(4),* 554-566.

Keck, C. M., & Biber, D. (2004). Modal use in spoken and written university registers. A corpus-based study. In Facchinetti, R., & Palmer, F. (eds.), *English modality in perspective: Genre analysis and contrastive studies* (pp. 3-25). Frankfurt am Main, Germany: Peter Lang.

Leacock, C., Chodorow, M., Gamon, M., & Tetreault, J. (2010). Automated Grammatical Error Detection for Language Learners. *Synthesis Lectures on Human Language Technologies, 3*(1), 1-134.

Lee, Y.-W., & Sawaki, Y. (2009). Application of three cognitive diagnosis models to ESL reading and listening assessments. *Language Assessment Quarterly, 6*(3), 239-263.

Liu, O. L. (2011). Does major field of study and cultural familiarity affect TOEFL iBT readiness performance? A confirmatory approach to differential item functioning. *Applied Measurement in Education, 24*(3), 235-255.

Madnani, N., Tetreault, J., & Chodorow, M. (2012, June) *Exploring Grammatical Error Correction with Not-So-Crummy Machine Translation.* In Proceedings of the 7th Workshop on Innovative Use of Natural Language Processing for Building Educational Applications, Montreal, Canada.

Plakans, L., & Gebril, A. (2013). Using multiple texts in an integrated writing assessment: Source text use as a predictor of score. *Journal of Second Language Writing, 22*(3), 217-230.

Sawaki, Y., Kim, H.-J., & Gentile, C. (2009). Q-matrix construction: Defining the link between constructs and test items in cognitive diagnosis. *Language Assessment Quarterly, 6*(3), 190-209.

Sawaki, Y., Quinlan, T., & Lee, Y.-W. (2013). Understanding learner strengths and weaknesses: Assessing performance on an integrated writing task. *Language Assessment Quarterly, 10*(1), 73-95. [Special issue].

Sawaki, Y., Stricker, L., & Oranje, A. (2009). Factor structure of the TOEFL Internet-based Test (TOEFL iBT). *Language Testing, 26*(1), 5-30.

Tannenbaum, R. J., & Wylie, E. C. (2008). *Linking English language test scores onto the Common European Framework of Reference: An application of standard-setting methodology* (TOEFL iBT Research Report RR-08-34). Princeton, NJ: Educational Testing Service.

Wall, D., & Horák, T. (2007). Using Baseline Studies in the Investigation of Test Impact. *Assessment in Education, 14*(1), 99-116.

Weigle, S. (2010). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability [Special issue]. *Language Testing, 27*(3), 335-353.

Williamson, D., Xi, X., & Breyer, J. (2012). A Framework for evaluation and use of automated scoring. *Educational Measurement, Issues and Practice, 31*(1), 2-13.

Winke, P. & Gass, S. (2013). The influence of second language experience and accent familiarity on oral proficiency rating: A qualitative investigation. *TESOL Quarterly, 47*(4), 762-789.

Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing, 30*(2), 231-252.

Xi, X. (2007). Validating TOEFL® iBT Speaking and setting score requirements for ITA screening. *Language Assessment Quarterly, 4*(4), 318-351.

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

Xi, X. (2008). What and how much evidence do we need? Critical considerations for using automated speech scoring systems. In C. Chapelle, Y-R. Chung, & J. Xu (Eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment* (pp. 102-114). Ames, IA: Iowa State University.

Xi, X., Bridgeman, B., & Wendler, C. (2013). Tests of English for academic purposes (EAP) in university admissions. In A. Kunnan (Ed.), *The Companion to Language Assessment* (pp. 318-337). Malden, MA: Wiley-Blackwell.

Xi, X., Higgins, D., Zechner, K., & Williamson, D. (2012). A comparison of two scoring methods for an automated speech scoring system. *Language Testing, 29*(3), 371-394.

Xi, X., & Mollaun, P. (2011). Using raters from India to score a large-scale speaking test. *Language Learning, 61*(4), 1222-1255.

Zechner, K., Higgins, D., & Xi, X. (2007). SpeechRater™: A construct-driven approach to scoring spontaneous non-native speech. *In proceedings of the SLaTE Workshop on Speech and Language Technology in Education.* Farmington, PA.

Zechner, K., Higgins, D., Xi, X., & Williamson. D. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. Special issue on Spoken Language Technology for Education. *Speech Communication, 51*(10), 883-895.