**TEST OF ENGLISH AS A FOREGIN LANGUAGE: SELECTED REFERENCES**

**(Last updated 17 December 2016)**

Alderson, J. C. & Hamp-Lyons, L. (1996). TOEFL preparation courses: A case study, *Language Testing,* *13*, 280-297.

Attali, Y. (2007). *Construct validity of e-rater in scoring TOEFL essays* (Research Report RR-07-21). Princeton, NJ: Educational Testing Service.

Attali, Y. (2011). *Automated subscores for TOEFL iBT independent essays* (Research Report RR-11-39). Princeton, NJ: Educational Testing Service.

Attali, Y., & Sinharay, S. (2015). *Automated trait scores for TOEFL writing tasks* (Research Report RR-15-14). Princeton, NJ: Educational Testing Service.

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. (2015). *Test-takers’ writing activities during the TOEFL iBT writing tasks: A stimulated recall study* (TOEFL iBT Research Report No. 25). Princeton, NJ: Educational Testing Service.

Barkaoui, K. (2016). What and when second-language learners revise when writing on the computer: The roles of task type, second-language proficiency, and keyboarding skills. *Modern Language Journal, 100*(1),320-340.

Bejar, I., Douglas, D., Jamieson, J., Nissan, S., & Turner, J. (2000). *TOEFL 2000 listening framework: A working paper* (TOEFL Monograph Series No. 19). Princeton, NJ: Educational Testing Service.

Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., Cortes, V., Csomay, E., & Urzua, A. (2004). *Representing language use in the university: Analysis of the TOEFL 2000 spoken and written academic language corpus* (TOEFL Monograph Series No. 25). Princeton, NJ: Educational Testing Service.

Biber, D., & Gray, B. (2013). *Discourse characteristics of writing and speaking task types on the TOEFL iBT Test: A lexico-grammatical analysis* (TOEFL iBT Research Report No. 19). Princeton, NJ: Educational Testing Service.

Breland, H., Lee, Y.-W., & Muraki, E. (2004). Comparability of TOEFL CBT essay prompts: Response-mode analyses. *Educational and Psychological Measurement, 65*(4), 577-595.

Breland, H., Lee, Y.-W., & Muraki, E. (2004). *Comparability of TOEFL CBT writing prompts: Response mode analyses* (TOEFL Research Report No. 75). Princeton, NJ: Educational Testing Service.

Breland, H., Lee, Y.-W., Najaran, M., & Muraki, E. (2004). *An analysis of TOEFL CBT writing prompt difficulty and comparability for different gender groups* (TOEFL Research Report No. 76). Princeton, NJ: Educational Testing Service.

Bridgeman, B., Cho, Y., & DiPietro, S. (2016). Predicting grades from an English language assessment: The importance of peeling the onion. *Language Testing, 33*(3)*,* 307-318.

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

Brooks, L., & Swain, M. (2014). Contextualizing performances: Comparing performances during TOEFL iBT and real-life academic speaking activities. *Language Assessment Quarterly, 11*(4), 353-373.

Brown, J. D., & Ross, J. A. (1996). Decision dependability of item types, sections, tests, and the overall TOEFL test battery. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment* (pp. 231-265). Cambridge, UK: Cambridge University Press.

Butler, F., Eignor, D., Jones, S., McNamara, T., & Suomi, B. (2000). *TOEFL 2000 speaking framework: A working paper* (TOEFL Monograph Series No. 20). Princeton, NJ: Educational Testing Service.

Carrell, P. L., Dunkel, P. A., & Mollaun, P. (2002). *The effects of notetaking, lecture length and topic on the listening component of TOEFL 2000* (TOEFL Monograph Series No. 23). Princeton, NJ: Educational Testing Service.

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*(1), 83-105.

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, NY: Routledge.

Chen, J., & Sheehan, K. M. (2015). *Analyzing and comparing reading stimulus materials across the TOEFL family of assessments* (TOEFL iBT Research Report No. 26). Princeton, NJ: Educational Testing Service.

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., & Burstein, J. (2004). *Beyond essay length: Evaluating e-rater's performance on TOEFL essays* (TOEFL Research Report No. 73). Princeton, NJ: Educational Testing Service.

Choi, I., & Papageorgiou, S. (2014). *Monitoring students’ progress in English language skills using the TOEFL ITP assessment series* (Research Memorandum RM-14-11). Princeton, NJ: Educational Testing Service.

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. (2006). *Strategies in responding to the new TOEFL reading tasks* (TOEFL Monograph Series No. 33). Princeton, NJ: Educational Testing Service.

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. E. (2004). *A teacher-verification study of speaking and writing prototype tasks for a new TOEFL* (TOEFL Monograph Series No. 26). Princeton, NJ: Educational Testing Service.

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

Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, M. U., & James, M. (2005). *Analysis of discourse features and verification of scoring levels for independent and integrated prototype written tasks for the new TOEFL* (TOEFL Monograph Series No. 30). Princeton, NJ: Educational Testing Service.

Cumming, A., Kantor, R., Baba, K., Erdosy, M. 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. E. (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 Series No. 22). Princeton, NJ: Educational Testing Service.

Cumming, A., Kantor, R., Powers, D. E., Santos, T., & Taylor, C. (2000). *TOEFL 2000 writing framework: A working paper* (TOEFL Monograph Series No. 18). Princeton, NJ: Educational Testing Service.

Deane, P., & Gurevich, O. (2008). *Applying content similarity metrics to corpus data: Differences between native and non-native speaker responses to a TOEFL integrated writing prompt* (Research Report RR-08-51). Princeton, NJ: Educational Testing Service.

Douglas, D. (1997). Testing speaking ability in academic contexts: Theoretical considerations. *TOEFL Monograph Series* (No. 8). Princeton: Educational Testing Service.

Douglas, D., (1999). Computer-based TOEFL: What test-takers can expect. Audio-Visual Education 22, 44-65.

Douglas, D., & Smith, J., (1997). Theoretical underpinnings of the Test of Spoken English Revision Project. *TOEFL Monograph Series* (No. 9). Princeton: Educational Testing Service.

Educational Testing Service. (2010). *Linking TOEFL iBT scores to IELTS scores - A research report*. Princeton, NJ: Author.

Educational Testing Service. (2010). *TOEFL iBT test framework and test development* (TOEFL iBT Research Insight Series, Vol. 1). Princeton, NJ: Author.

Educational Testing Service. (2010). *TOEFL research* (TOEFL iBT Research Insight Series, Vol. 2). Princeton, NJ: Author.

Educational Testing Service. (2011). *Information for score users, teachers and learners* (TOEFL iBT Research Insight Series, Vol. 5). Princeton, NJ: Author.

Educational Testing Service. (2011). *Reliability and comparability of TOEFL iBT scores* (TOEFL iBT Research Insight Series, Vol. 3). Princeton, NJ: Author.

Educational Testing Service. (2011). *TOEFL program history* (TOEFL iBT Research Insight Series, Vol. 6). Princeton, NJ: Author.

Educational Testing Service. (2011). *Validity evidence supporting the interpretation and use of TOEFL iBT scores* (TOEFL iBT Research Insight Series, Vol. 4). Princeton, NJ: Author.

Enright, M. K. (2004). Research issues in high-stakes communicative language testing: Reflections on TOEFL's new directions. *TESOL Quarterly, 38*(1), 147-151.

Enright, M. K., Grabe, W., Koda, K., Mosenthal, P., Mulcahy-Ernt, P., & Schedl, M. (2000). *TOEFL 2000 reading framework: A working paper* (TOEFL Monograph Series No. 17). 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.

Erdosy, M. U. (2004). *Exploring variability in judging writing ability in a second language: A study of four experienced raters of ESL compositions* (TOEFL Research Report No. 70). Princeton, NJ: Educational Testing Service.

Educational Testing Service. (2010). *Test and score data summary for TOEFL internet-based and paper-based tests*. Princeton, NJ: Educational Testing Service.

Farnsworth, T. L. (2013). An investigation into the validity of the TOEFL iBT speaking test for international teaching assistant certification. *Language Assessment Quarterly, 10*(3), 274-291.

Ginther, A. (2001). *Effects of the presence and absence of visuals on subjects’ performance on TOEFL CBT listening comprehension stimuli* (TOEFL Research Report No. 66). Princeton, NJ: Educational Testing Service.

Ginther, A., & Elder, C. A. (2014). *A comparative investigation into understandings and uses of the TOEFL iBT test, the International English Language Testing Service (Academic) test, and the Pearson Test of English for graduate admissions in the United States and Australia: A case study of two university contexts* (TOEFL iBT Research Report No. 24). Princeton, NJ: Educational Testing Service.

García Gómez, P., Noah, A., Schedl, M., Wright, C., & Yolkut, A. (2007). Proficiency descriptors based on a scale-anchoring study of the new TOEFL iBT reading test. *Language Testing, 24*(3), 417-444.

Gu, L., & Xi, X. (2015). *Examining performance differences on tests of academic English proficiency used for high-stakes vs. practice purposes* (Research Memorandum RM-15-09). Princeton, NJ: Educational Testing Service.

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing, 18*(3), 218-238.

Haberman, S. J. (2011). *Use of e-rater in scoring of the TOEFL iBT writing test* (Research Memorandum RM-11-25). Princeton, NJ: Educational Testing Service.

Hansen, E. G., Forer, D. C., & Lee, M. J. (2004). *Towards accessible computer-based tests: Prototypes for visual and other disabilities* (TOEFL Research Report No. 78). Princeton, NJ: Educational Testing Service.

Hill, Y. Z., & Liu, O. L. (2012). *Is there any interaction between background knowledge and language proficiency that affects TOEFL iBT reading performance?* (TOEFL iBT Research Report No. 18). Princeton, NJ: Educational Testing Service.

Jamieson, J., Eignor, D., Grabe, W., & Kunnan, A. J. (2008). Frameworks for a new TOEFL. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 55-95). New York, NY: Routledge.

Jamieson, J., Jones, S., Kirsch, I., Mosenthal, P., & Taylor, C. (2000). *TOEFL 2000 framework: A working paper* (TOEFL Monograph Series No. 16). Princeton, NJ: Educational Testing Service.

Jamieson, J., & Poonpon, K. (2013). *Developing analytic rating guides for TOEFL iBT integrated speaking tasks* (TOEFL iBT Research Report No. 20). Princeton, NJ: Educational Testing Service.

Johnson, K.E., Jordan, S.R., & Poehner, M. (2005). The TOEFL trump card: An investigation of test impact in an ESL classroom. *Critical Inquiry in Language Studies, 2*(2), 71-94.

Katz, I. R., Xi, X., Kim, H.-J., & Cheng, P. C.-H. (2004). *Elicited speech from graph items on the Test of Spoken English* (TOEFL Research Report No. 74). Princeton, NJ: Educational Testing Service.

Knoch, U., Macqueen, S., & O’Hagan, S. (2014). *An investigation of the effect of task type on the discourse produced by students at various score levels in the TOEFL iBT writing test* (TOEFL iBT Research Report No. 23). Princeton, NJ: Educational Testing Service.

Kostin, I. (2004). *Exploring item characteristics that are related to the difficulty of TOEFL dialogue items* (TOEFL Research Report No. 79). Princeton, NJ: Educational Testing Service.

Kyle, K., & Crossley, S. A. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing, 34,* 12-24.

Kyle, K., Crossley, S. A., & McNamara, D. S. (2016). Construct validity in TOEFL iBT speaking tasks: Insights from Natural Language Processing. *Language Testing, 33*(3), 319-340.

Lee, Y.-W. (2004). *Dependability of scores for a new ESL speaking test: Evaluating prototype tasks* (TOEFL Monograph Series No. 28). Princeton, NJ: Educational Testing Service.

Lee, Y.-W., Breland, H., & Muraki, E. (2004). *Comparability of TOEFL CBT writing prompts for different native language groups* (TOEFL Research Report No. 77). Princeton, NJ: Educational Testing Service.

Lee, Y.-W., Deane, P., Breland, H. M., & Muraki, E. (2005). Comparability of TOEFL CBT writing prompts for different native language groups. *International Journal of Testing, 5*(2), 131-158.

Lee, Y.-W., Gentile, C., & Kantor, R. (2008). *Analytic scoring of TOEFL CBT essays: Scoring by humans and e-rater* (TOEFL Research Report No. 81). Princeton, NJ: Educational Testing Service.

Lee, Y.-W., & Kantor, R. (2005). *Dependability of new ESL writing test scores: Evaluating prototype tasks and alternative rating schemes* (TOEFL Monograph Series No. 31). Princeton, NJ: Educational Testing Service.

Li, Y., & Brown, T. (2013). *A trend-scoring study for the TOEFL iBT speaking and writing sections* (Research Memorandum RM-13-05). Princeton, NJ: Educational Testing Service.

Ling, G., Powers, D. E., & Adler, R. M. (2014). *Do TOEFL iBT scores reflect improvement in English-language proficiency? Extending the TOEFL validity argument* (Research Report RR-14-09). Princeton, NJ: Educational Testing Service.

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.

Liu, O. L. (2014). *Investigating the relationship between test preparation and TOEFL iBT performance* (Research Report RR-14-15). Princeton, NJ: Educational Testing Service.

Liu, O. L., Schedl, M., Malloy, J., & Kong, N. (2009). *Does content knowledge affect TOEFL iBT reading performance? A confirmatory approach to differential item functioning* (TOEFL iBT Research Report No. 09). Princeton, NJ: Educational Testing Service.

Madnani, N., Tetreault, J., & Chodorow, M. (2012). Exploring grammatical error correction with not-so-crummy machine translation. *Proceedings of the 7th Workshop on Innovative Use of Natural Language Processing for Building Educational Applications,* 44-53. <http://www.academia.edu/26513537/Exploring_grammatical_error_correction_with_not-so-crummy_machine_translation>

Major, R. C., Fitzmaurice, S. F., Bunta, F., & Balastubramanian, C. (2002). The effects of nonnative accents on listening comprehension: Implications for ESL assessment. *TESOL Quarterly, 36*(2), 173-190.

Malone, M. E., & Montee, M. (2014). *Stakeholders’ beliefs about the TOEFL iBT test as a measure of academic language ability* (TOEFL iBT Research Report No. 22). Princeton, NJ: Educational Testing Service.

Manna, V. F., & Yoo, H. (2015). *Investigating the relationship between test-taker background characteristics and test performance in a heterogeneous English-as-a-Second-Language (ESL) test population: A factor analytic approach* (Research Report RR-15-25). Princeton, NJ: Educational Testing Service.

Moglen, D. (2015). The re-placement test: Using TOEFL for purposes of placement. *The CATESOL Journal, 27*(1), 1-26.

Myford, C. M., & Wolfe, E. W. (2000). *Monitoring sources of variability within the Test of Spoken English assessment system* (TOEFL Research Report No. 65). Princeton, NJ: Educational Testing Service.

Myford, C. M., & Wolfe, E. W. (2000). *Strengthening the ties that bind: Improving the linking network in sparsely connected rating designs* (TOEFL Technical Report No. 15). Princeton, NJ: Educational Testing Service.

Ockey, G. J., Koyama, D., Setoguchi, E., & Sun, A. (2015). The extent to which TOEFL iBT Speaking scores are associated with performance on oral language tasks and oral ability components for Japanese university students. *Language Testing, 32*(1), 39-62.

Papageorgiou, S., Tannenbaum, R. J., Bridgeman, B., & Cho, Y. (2015). *The association between TOEFL iBT test scores and the Common European Framework of Reference (CEFR) levels* (Research Memorandum RM-15-06). Princeton, NJ: Educational Testing Service.

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.

Powers, D. E. (2011). *Scoring the TOEFL independent essay automatically: Reactions of test takers and test score users* (Research Memorandum RM-11-34). Princeton, NJ: Educational Testing Service.

Powers, D. E., Albertson, W., Florek, T., Johnson, K., Malak, J., Nemceff, B., & Zelazny, A. (2002). *Influence of irrelevant speech on standardized test performance* (TOEFL Research Report No. 68). Princeton, NJ: Educational Testing Service.

Powers, D. E., & Lall, V. F. (2012). *Supporting an expiration policy for TOEFL scores* (Research Memorandum RM-12-03). Princeton, NJ: Educational Testing Service.

Powers, D. E., Roever, C., Huff, K. L., & Trapani, C. S. (2003). *Validating LanguEdge Courseware against faculty ratings and student self-assessments* (Research Report RR-03-11). Princeton, NJ: Educational Testing Service.

Powers, D., Schedl, M., & Papageorgiou, S. (2016). Facilitating the interpretation of English language proficiency scores: Combining scale anchoring and test score mapping methodologies. *Language Testing*, 0265532215623582, 1-21.

Ramineni, C., Trapani, C. S., Williamson, D. M., Davey, T., & Bridgeman, B. (2012). *Evaluation of the e-rater scoring engine for the TOEFL independent and integrated prompts* (Research Report RR-12-06). Princeton, NJ: Educational Testing Service.

Riazi, M. (2016). Comparing writing performance in TOEFL-iBT and academic assignments: An exploration of textual features. *Assessing Writing, 28,* 15-27.

Roever, C., & Powers, D. E. (2005). *Effects of language of administration on a self-assessment of language skills* (TOEFL Monograph Series No. 27). Princeton, NJ: Educational Testing Service.

Rosenfeld, M., Leung, S., & Oltman, P. K. (2001). *The reading, writing, speaking, and listening tasks important for academic success at the undergraduate and graduate levels* (TOEFL Monograph Series No. 21). Princeton, NJ: Educational Testing Service.

Rosenfeld, M., Oltman, P. K., & Sheppard, K. (2004). *Investigating the validity of TOEFL: A feasibility study using content and criterion-related strategies* (TOEFL Research Report No. 71). Princeton, NJ: Educational Testing Service.

Sawaki, Y., & Nissan, S. (2009). *Criterion-related validity of the TOEFL iBT listening section* (TOEFL iBT Research Report No. 08). Princeton, NJ: Educational Testing Service.

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.

Sawaki, Y., & Sinharay, S. (2013). *Investigating the value of section scores for the TOEFL iBT test* (TOEFL iBT Research Report No. 21). Princeton, NJ: Educational Testing Service.

Sawaki, Y., Stricker, L. J., & Oranje, A. (2008). *Factor structure of the TOEFL Internet-Based Test (iBT): Exploration in a field trial sample* (TOEFL iBT Research Report No. 04). Princeton, NJ: Educational Testing Service.

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

Stricker, L. J. (1997). *Using just noticeable differences to interpret Test of Spoken English scores* (TOEFL Research Report 58, RR 97–4). Princeton, NJ: Educational Testing Service.

Stricker, L. J. (2002). *The performance of native speakers of English and ESL speakers on the TOEFL CBT and GRE general test* (TOEFL Research Report No. 69). Princeton, NJ: Educational Testing Service.

Stricker, L. J. (2004). The performance of native speakers of English and ESL speakers on the computer-based TOEFL and GRE General Test. *Language Testing, 21*(2), 146-173.

Stricker, L. J., & Attali, Y. (2010). *Test takers’ attitudes about the TOEFL iBT* (TOEFL iBT Research Report No. 13). Princeton, NJ: Educational Testing Service.

Stricker, L. J., & Rock, D. A. (2008). *Factor structure of the TOEFL Internet-Based test across subgroups* (TOEFL iBT Research Report No. 07). Princeton, NJ: Educational Testing Service.

Stricker, L. J., Rock, D. A., & Lee, Y.-W. (2005). *Factor structure of the LanguEdge test across language groups* (TOEFL Monograph Series No. 32). Princeton, NJ: Educational Testing Service.

Stricker, L. J., & Wilder, G. Z. (2001). *Examinees' attitudes about the TOEFL-CBT, possible determinants, and relationships with test performance* (Research Report RR-01-01). Princeton, NJ: Educational Testing Service.

Stricker, L. J., & Wilder, G. Z. (2012). *Test takers’ interpretation and use of TOEFL iBT score reports: A focus group study* (Research Memorandum RM-12-08). Princeton, NJ: Educational Testing Service.

Stricker, L. J., Wilder, G. Z., & Rock, D. A. (2004). Attitudes about the computer-based Test of English as a Foreign Language. *Computers in Human Behavior, 20*(1), 37-54.

Swain, M., Huang, L.-S., Barkaoui, K., Brooks, L., & Lapkin, S. (2009). *The speaking section of the TOEFL iBT (SSTiBT): Test-takers’ reported strategic behaviors* (TOEFL iBT Research Report No. 10). Princeton, NJ: Educational Testing Service.

Tang, K. L., & Eignor, D. R. (2001). *A study of the use of collateral statistical information in attempting to reduce TOEFL IRT item parameter estimation sample sizes* (TOEFL Technical Report No. 17). Princeton, NJ: Educational Testing Service.

Tannenbaum, R. J., & Baron, P. A. (2011). *Mapping TOEFL ITP scores onto the Common European Framework of Reference* (Research Memorandum RM-11-33). Princeton, NJ: Educational Testing Service.

Tannenbaum, R. J., & Wylie, E. C. (2005). *Mapping English language proficiency test scores onto the common European framework* (TOEFL Research Report No. 80). Princeton, NJ: Educational Testing Service.

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 No. 06). Princeton, NJ: Educational Testing Service.

Tao, J., Ghaffarzadegan, S., Chen, L, Zechner, K. (2016). Exploring deep learning architectures for automatically grading non-native spontaneous speech. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 6140-6144.

Wagner, E. (2016). *A study of the use of TOEFL iBT test speaking and listening scores for international teaching assistant screening* (TOEFL iBT Research Report No. 27). Princeton, NJ: Educational Testing Service.

Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement, 37*(3), 203-220.

Wainer, H., & Wang, X. (2001). *Using a new statistical model for Testlets to score TOEFL* (TOEFL Technical Report No. 16). Princeton, NJ: Educational Testing Service.

Wait, I. W., & Gressel, J. W. (2009). Relationship between TOEFL score and academic success for international engineering students. *Journal of Engineering Education,98*(4), 389-398.

Wall, D., & Horák, T. (2006). *The impact of Changes in the TOEFL examination on teaching and learning in Central and Eastern Europe: Phase 1, the baseline study* (TOEFL Monograph Series No. 34). Princeton, NJ: Educational Testing Service.

Wall, D., & Horák, T. (2008). *The impact of changes in the TOEFL examination on teaching and learning in Central and Eastern Europe: Phase 2, Coping with change* (TOEFL iBT Research Report No. 05). Princeton, NJ: Educational Testing Service.

Wall, D., & Horák, T. (2011). *The impact of changes in the TOEFL exam on teaching and learning in a sample of countries in Europe: Phase 3, The role of the course book. Phase 4, Describing change* (TOEFL iBT Research Report No. 17). Princeton, NJ: Educational Testing Service.

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

Weigle, S. C. (2011). *Validation of automated scores of TOEFL iBT tasks against nontest indicators of writing ability* (TOEFL iBT Research Report No. 15). Princeton, NJ: Educational Testing Service.

Winke, P., Gass, S., & Myford, C. (2011). *The relationship between raters' prior language study and the evaluation of foreign language speech samples* (TOEFL iBT Research Report No. 16). Princeton, NJ: Educational Testing Service.

Wolfe, E. W. (2003). Examinee characteristics associated with choice of composition medium on the TOEFL writing section. *The Journal of Technology, Learning, and Assessment, 2*(4), 1-25.

Wolfe, E. W., & Manalo, J. R. (2004). Composition medium comparability in a direct writing assessment of non-native English speakers. *Language Learning & Technology, 8*(1), 53-65.

Wolfe, E. W., & Manalo, J. R. (2005). *An investigation of the impact of composition medium on the quality of scores from the TOEFL writing section: A report from the broad-based study* (TOEFL Research Report No. 72). Princeton, NJ: Educational Testing Service.

Wylie, E. C., & Tannenbaum, R. J. (2006). *TOEFL academic speaking test: Setting a cut score for international teaching assistants* (Research Memorandum RM-06-01). Princeton, NJ: Educational Testing Service.

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

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

Xi, X. (2008). *Investigating the criterion-related validity of the TOEFL speaking scores for ITA screening and setting standards for ITAs* (TOEFL iBT Research Report No. 03). Princeton, NJ: Educational Testing Service.

Xi, X. (2008). Validating TOEFL iBT Speaking and setting score requirements for ITA screening: Erratum. *Language Assessment Quarterly, 5*(1), 87.

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

Xi, X., & Mollaun, P. (2009). *How do raters from India perform in scoring the TOEFL iBT speaking section and what kind of training helps?* (TOEFL iBT Research Report No. 11). Princeton, NJ: Educational Testing Service.

Zhang, Y. (2008). *Repeater analyses for the TOEFL iBT test* (Research Memorandum RM-08-05). Princeton, NJ: Educational Testing Service.