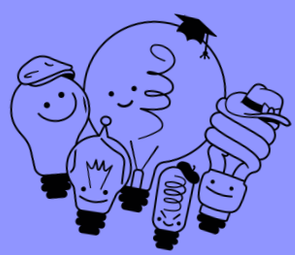
toefl®

TOEFL® Research INSIGHTS

TOEFL® Research

VOLUME 2



TOEFL® Research Insight Series, Volume 2: TOEFL Research

Preface

The TOEFL iBT® test is the world's most widely respected English language assessment, used for admissions purposes in more than 150 countries, including Australia, Canada, New Zealand, the United Kingdom, and the United States (see test review in Alderson, 2009). Since its initial launch in 1964, the TOEFL® test has undergone several major revisions motivated by advances in theories of language ability and changes in English teaching practices. The most recent revision, the TOEFL iBT test, was launched in 2005. It contains a number of innovative design features, including integrated tasks that engage multiple skills to simulate language use in academic settings and test materials that reflect the reading, listening, speaking, and writing demands of real-world academic environments.

In addition to the TOEFL iBT test, the TOEFL® Family of Assessments was expanded to provide high-quality, English proficiency assessments for a variety of academic uses and contexts. The TOEFL® Young Students Series (YSS) features the TOEFL Primary® and TOEFL Junior® tests, which are designed to help teachers and learners of English in school settings. In addition, the TOEFL ITP® program offers colleges, universities, and others affordable tests for placement and progress monitoring within English programs as a pathway to eventual degree programs. The TOEFL Essentials test evaluates the four language skills in a friendly test format, with short, engaging tasks that relate to both academic situations and everyday life.

At ETS, we understand that scores from the TOEFL Family of Assessments are used to help make important decisions about students, and we would like to keep score users and test takers up to date about the research results that help assure the quality of these scores. Through the TOEFL® Research Insight Series, we provide institutions and English teachers with information regarding the strong research and development base that underlies the TOEFL Family of Assessments, and demonstrates our continued commitment to research.

Since the 1970s, the TOEFL test has had a rigorous, productive, and far-ranging research program. But why should test score users care about the research base for a test? In short, it is only through a rigorous program of research that a testing company can substantiate claims about what test takers know or can do based on their test scores, as well as provide support for the intended uses of assessments and minimize potential negative consequences of score use. Beyond demonstrating this critical evidence of test quality, research is also important for enabling innovations in test design and addressing the needs of test takers and test score users. This is why ETS has established a strong research base as a fundamental feature underlying the evolution of the TOEFL Family of Assessments.

This TOEFL Family of Assessments is designed, produced, and supported by a world-class team of test developers, educational measurement specialists, statisticians, and researchers in applied linguistics and language testing. Our test developers have advanced degrees in fields such as English, language education, and applied linguistics. They also possess extensive international experience, having taught English on continents around the globe. Our research, measurement, and statistics teams include some of the world's most distinguished scientists and internationally recognized leaders in diverse areas such as test validity, language learning and assessment, and educational measurement.

To date, more than 300 peer-reviewed TOEFL Family of Assessments research reports, technical reports, and monographs have been published by ETS, and many more studies on the TOEFL tests have appeared in academic journals and book volumes. In addition, over 20 TOEFL test-related research projects are conducted by ETS's Research & Development staff each year and the TOEFL Committee of Examiners, comprised of language learning and testing experts from the global academic community, funds an annual program of TOEFL Family research by independent external researchers from all over the world.

The purpose of the *TOEFL Research Insight Series* is to provide a comprehensive yet user-friendly account of the essential concepts, procedures, and research results that help ensure the quality of scores for all members of the TOEFL Family of Assessments. Topics covered in these volumes feature issues of core interest to test users, including how tests were designed; evidence for the reliability, validity and fairness of test scores; and research-based recommendations for best practices.

The close collaboration with TOEFL score users, English language learning and teaching experts, and university scholars in the design of all TOEFL tests has been a cornerstone to their success and worldwide acceptance. Therefore, through this publication, we hope to foster an ever-stronger connection with our test users by sharing the rigorous measurement and research base and solid test development that continues to help ensure the quality of the TOEFL Family of Assessments.

Acknowledgements

The following ETS staff contributed to this version of Volume 2, updated in November 2024 (in alphabetical order): Larry Davis, Yoko Futagi, Lixiong Gu, Spiros Papageorgiou.

The following individuals also contributed to previous versions of this volume, by providing careful reviews and revisions, as well as editorial suggestions (in alphabetical order):
Terry Axe, Ian Blood, Jill Burstein, Ikkyu Choi, Keelan Evanini, Michelle Hampton, Marcel Ionescu, Eileen Tyson, Lin Wang, and Klaus Zechner. The primary author of the first edition was Mary K. Enright. Cristiane Breining, Brent Bridgeman, Don Powers, Rosalie Szabo, Xiaofei Tang, Eileen Tyson, Mikyung Kim Wolf, and Xiaoming Xi also contributed to the first edition.

TOEFL Research

The TOEFL program has long recognized and supported the importance of research in maintaining and improving test quality. Since the mid-1970s, a portion of the annual TOEFL budget has been committed to fund and disseminate research on issues related to language assessment. ETS supports a research program to advance knowledge in the field of language assessment and second-language acquisition.
The goals are to:

- improve language assessments and related products and services,
- help ensure that assessments, related products and services meet professional standards, and
- develop the foundation for new products and services.

The TOEFL Committee of Examiners (COE), a body of eleven individuals from around the world, each of whom has achieved professional recognition in an academic field related to English as a second or foreign language, works closely with the TOEFL program on its program of research.

The Research Process`

TOEFL research is carried out in consultation with the COE, which advises the TOEFL program about research needs and, through its research subcommittees, administers the COE research program. Through this research program, the COE solicits, reviews, and approves the funding of research proposals from experts around the globe. The TOEFL program also funds an extensive internal program of research conducted at ETS by its own staff.

To encourage external experts to conduct TOEFL research, the COE publishes an annual announcement of its research program, describing high-priority research topics. Applications are invited from research professionals who have expertise in English language learning and assessment and who are affiliated with research institutions, such as universities or not-for-profit organizations. The COE research subcommittee review the preliminary funding applications. Invitations to submit a full proposal are issued to selected applicants based on the quality of the preliminary application. Full research proposals are then evaluated in terms of their relevance to the identified research topics, the feasibility and quality of the proposed research, the qualifications of the principal investigator, organizational capacity to conduct the research, and cost effectiveness.

The quality of TOEFL research is ensured through a rigorous review process. Three to four ETS and external experts review proposals and reports. The reviewers may include applied linguists, psychologists, statisticians, psychometricians, or assessment specialists. After reports are reviewed, external researchers are encouraged to disseminate their findings in a number of ways, for example, by publishing in professional journals and the ETS Research Report series, as well as through presentations at regional and international conferences.

The TOEFL program also provides a variety of other monetary grants and awards to recognize and support significant activities or projects related to the field of English language education, and to promote high- quality language assessment research.

Grants are available to promising students working in the area of foreign- or second-language assessment, to help them finish their dissertations in a timely manner. Grants are also available to enable practitioners to become involved in ETS's efforts in promoting English learning and to encourage the broad dissemination of information on English language testing, teaching, and teacher education through presentations at conferences outside the United States.

Information about TOEFL research grants and awards is published at https://www.ets.org/toefl/grants.

Description of Selected TOEFL Research

More than 300 research reports related to the TOEFL family of assessments have been published by ETS (https://www.ets.org/toefl/research). Moreover, since the year 2000 alone, more than 100 academic journal articles and book chapters on TOEFL related research have been published, as well as six books (Barkaoui & Hadidi, 2020; Chapelle et al., 2008, Davis & Norris, 2024; Papageorgiou & Manna, 2023, Wolf & Butler, 2017; Zechner & Evanini, 2019) and more than 100 presentations at academic conferences. Certain research topics such as test validation, fairness, and reliability have been repeatedly re-examined over time as test methods and content evolved. Other topics include innovations in testing (such as advances in psychometrics, automated scoring, and computer-based testing) and projects focused on the implications of theories of language proficiency for test design.

A comprehensive summary of all the research sponsored by ETS is well beyond the scope of this document. Nevertheless, in the pages that follow, we will make a selective presentation concentrating on topics not reviewed in other publications. The extensive program of research to improve language assessment that resulted in the TOEFL iBT test is documented in a book edited by Chapelle, Enright, and Jamieson (2008). Summaries of research and procedures to ensure that the TOEFL test complies with professional standards for validity (ETS, 2020c) and reliability (ETS, 2020a) are available. In this section, we will focus on research concerning test fairness and automated analysis of writing and speaking.

Research on Test Fairness

Fairness in testing is an important measurement standard that the TOEFL program strives to meet. For the TOEFL test, test fairness means that the test scores can be interpreted as a measure of academic English language ability for various groups of test takers. Fairness requires that test scores should not be affected by factors that are not relevant to this intended interpretation. Although care is taken during test development to ensure that test content meets fairness guidelines, empirical research studies are also conducted to determine the impact of various factors on test scores. Four studies have addressed three fairness issues related to TOEFL iBT test scores: (a) the structure of the test for different groups of test takers, (b) the impact of educational and cultural background on reading performance, and (c) the performance of native English-speaking college students on the TOEFL iBT test.

One fairness issue concerns what specialists refer to as the factor structure of test scores for different groups of test takers. Factor analysis is a statistical research method that can be used to determine the underlying statistical structure of scores on a test. The factor structure of a test should be consistent with the theoretical structure implied by the test's construct—the characteristic that the test is designed to measure (e.g., English language proficiency). A test's factor structure also has implications for how scores should be reported and interpreted. Stricker and Rock (2008) analyzed the factor structure of a 2003–2004 TOEFL iBT field test form for three groups of test takers. Test takers were grouped according to (a) whether their first language was from an Indo European versus a non–Indo European language family, (b) how widely English was used in education and business contexts in their native countries, and (c) years of studying the English language in school.

The same factor structure was found for all subgroups. Analyses of operational TOEFL iBT test forms (Gu, 2014; Manna & Yoo, 2015; Sawaki & Sinharay, 2013) also showed that the test's factor structure was consistent across different groups defined by first language and test taker background characteristics. A consistent factor structure across different groups of test takers provides evidence that the test measures the same construct for the groups studied and that score aggregation and reporting procedures lead to appropriate score interpretations for these groups.

Another important question that researchers have asked about the fairness of the TOEFL iBT test is whether factors other than English language proficiency impact test performance. Liu, Schedl, Malloy, and Kong (2009) asked this question in regard to the TOEFL iBT Reading section, which has fewer but longer reading passages than previous versions of the TOEFL test. Their concern was that the decreased topic variety might increase the likelihood that test takers' familiarity with the particular topic of a given passage would influence their reading performance on the test. Accordingly, they investigated whether TOEFL iBT test reading performance was affected by test takers' outside knowledge, gained either through academic major or from immersion in a particular culture. Performance on six passages and associated questions from five TOEFL iBT test administrations were examined. Three of the passages focused on topics in physical science, and the rest emphasized European or Japanese cultures. Techniques known as differential item functioning (DIF) and differential bundle functioning (DBF) were used to investigate the impact of outside knowledge on TOEFL iBT test reading performance. DIF occurs for an item when differences in performance exist after examinees are matched on the abilities that the item is intended to measure. Liu et al. found little evidence that the sources of outside knowledge they investigated influenced overall performance on the reading passages. Further, the analysis of the items displaying DIF suggests that the differences in performance may be construct-relevant differences that TOEFL iBT test is intended to measure (e.g., vocabulary knowledge). To ensure continued fairness, the researchers recommended that passages containing technical vocabulary or culture-specific knowledge should be carefully scrutinized in the future.

Another study (Hill & Liu, 2012) explored the interaction between test takers' language proficiency and background knowledge, with the focus on their discipline-specific knowledge and cultural familiarity. The study reanalyzed the data used in Liu et al. (2009) employing DIF methods and concluded: "When examined holistically, the TOEFL iBT reading passages were neither advantageous nor disadvantageous to those who had physical science backgrounds or were familiar with a certain culture, and this holds for both the lower and higher proficiency groups" (Hill & Liu, 2012, p. 28).

A third fairness concern is that the TOEFL iBT test, with its academic content and tasks that require integrating different language skills, might be very difficult even for native English speakers. Native speakers, overall, do not represent the "ultimate criterion group for an ESL test, because they vary in formal and informal education in English and in linguistic ability" (Stricker, 2002, p. 1). Nevertheless, if educated native English speakers cannot do as well as educated non-native speakers on the TOEFL iBT test, it might be claimed that non-native speakers are being held unfairly to a higher standard in admissions decisions than native speakers. Cline and Powers (2009) compared the performance of first-year college students who were native speakers of English with that of non-native speakers. They administered one form of the 2003–2004 TOEFL iBT field test to more than 900 first-year, native English-speaking students at community colleges and nonselective 4-year colleges and compared their performance with that of the non-native speakers who had completed the field study form. Overall, the native English-speaking college students performed better than non-native speakers, although there was a reasonable amount of variation in scores within this group. The mean score differences avoring the native English speakers were moderate for listening and reading, but large for speaking and the total score. The implications are that the TOEFL iBT test is neither inappropriately difficult for non-native English speakers nor unusually easy for native English speakers. This suggests that non-native speakers are being held to a high standard, but not an unfair one.

In sum, these studies of test structure, test content, and native-speaker performance illustrate some of the fairness issues that have been addressed empirically through TOEFL research.

Automated Scoring for Writing and Speaking

Two needs arise when a test includes extended constructed-response tasks, such as the Writing and Speaking tasks on the TOEFL iBT test. One of these is the need to score the responses efficiently and reliably. The other is to provide test takers with opportunities to practice and receive feedback on their performance prior to taking the test. Through research on automated scoring of writing and speaking, ETS and the TOEFL program have been laying the foundation for new products and services that address these needs. Capabilities developed at ETS that address these needs have included the e-rater® and the SpeechRater® engines.

e-rater Engine

The e-rater engine uses natural language processing methods to automatically score written responses as well as to provide feedback on the quality of their writing. The e-rater engine identifies errors in grammar, usage, and mechanics, as well as discourse structure and undesirable stylistic features in an essay. These features, along with measures of the vocabulary and sentence variety used in an extended written response, go into the e-rater engine's statistical model to predict human holistic ratings on these responses. The engine has also been used in practice and learning products . to provide instant scoring and annotated feedback.

An extensive program of research contributed to the continuous development and refinement of these capabilities and their evaluation for use in different contexts. Although this research initially focused on analyzing and scoring essays written primarily by native English speakers (e.g., Kaplan et al., 1998), attention soon expanded to include research on essays written specifically by non-native English speakers (e.g., Chodorow & Burstein, 2004).

One area of research interest has been the validity of using the e-rater engine in conjunction with human raters to score the TOEFL iBT Writing tasks (for more information about the use of the e-rater engine in scoring TOEFL iBT Writing tasks, see Volume 3: Reliability and Comparability of TOEFL iBT® Scores). In their summary of research on the use of the e-rater engine for the independent Writing task, Enright and Quinlan (2010) reported that the e-rater engine has been found to agree with human raters as well as or better than human raters agree with each other when rating these essays. Overall, the empirical evidence summarized by Enright and Quinlan supports the use of the e-rater engine as a complement to human raters to score TOEFL test independent essays. Research has also been conducted to evaluate the use of the e-rater engine for the integrated Writing task, which requires test takers to summarize and synthesize academic reading and listening materials in writing. The areas of research included the degree of agreement of the e-rater engine with human scores, the relationships of human and e-rater engine scores to independent indicators of language ability, and the impact of the use of the e-rater engine on scores by demographic subgroup. The results yielded evidence in support of the use of the e-rater engine to complement human raters for the TOEFL test integrated Writing task as well. These studies are summarized in Ramineni, Trapani, Williamson, Davey, and Bridgeman (2012).

ETS has also conducted extensive research on the technology underlying the e-rater engine, to improve existing features as well as to expand construct coverage of the engine. Such research includes, for example, studies on preposition and comma error detection (Israel, Tetreault, & Chodorow, 2012; Tetreault, Foster, & Chodorow, 2010). Burstein, Flor, Tetreault, Madnani, and Holtzman (2012) systematically examined the paraphrase strategies used by native and non-native English speakers in a TOEFL test integrated task, as a first step toward informing the development of new e-rater engine features. Beigman Klebanov, Madnani, Burstein, and Somasundaran (2014) described a method of automatically detecting effective use of source (e.g., stimulus lecture) in a TOEFL test integrated task.

Collaboration between the TOEFL program and ETS researchers has made a unique contribution to the field of natural language processing and corpus linguistics, too, by making it possible to release the ETS Corpus of Non-Native Written English (Blanchard, Tetreault, Higgins, Cahill, & Chodorow, 2013), which is publicly available through the Linguistic Data Consortium. The corpus consists of 12,100 English essays written for the TOEFL test by speakers of eleven non-English native languages (1,100 per language) during 2006–2007. Originally developed with the specific task of native language identification in mind, the corpus can support a wide range of applications of natural language processing to the educational domain, including grammatical error detection and correction, automatic essay scoring, and studies in corpus linguistics.

SpeechRater Engine

Automated scoring of speech is a more recent development than automated scoring of writing and presents a greater challenge, in part because of the difficulty of automatically recognizing the words uttered in a response consisting of continuous speech. While speech scoring systems for simple tasks that require the production of a limited or predictable range of vocabulary have been in use for a number of years (see Zechner, Higgins, Xi, & Williamson, 2009, for a review), the tasks on the TOEFL iBT test Speaking section are more complex.
The Speaking section includes four tasks that require test takers to respond either to a relatively general question or to oral and/or written input. TOEFL iBT test spoken responses are scored holistically by human raters using a four-point scale; however, the raters are instructed to attend to three key aspects of performance: delivery, language use, and topic development (see ETS, 2024b). In addition, the SpeechRater engine also computes scores for a response to each TOEFL iBT test Speaking task, and human and automated scores are then combined using a contributory scoring approach, to produce a score for the task.

Apart from being part of a hybrid human-machine contributory scoring approach for operational TOEFL iBT test Speaking tasks, the SpeechRater engine has been used to provide sole scores for responses to TOEFL iBT test Speaking tasks in a practice environment (Zechner et al., 2009).

The engine consists of four components: a speech recognizer, a feature computation module, a filtering model, and a scoring model. The speech recognizer provides a word sequence based on the recorded response of a test taker and was trained on around 1,600 hours of responses by non-native English speakers to TOEFL iBT Speaking tasks. The feature computation module uses the output of the speech recognizer to compute a set of features related to various aspects of speaking proficiency (e.g., fluency, pronunciation, vocabulary). The filtering model flags responses that should not be scored by the SpeechRater engine (e.g., responses with no speech or with high levels of noise). The scoring model uses the features from the feature computation module to statistically predict a score for each response.

Research related to SpeechRater scoring system has addressed many aspects of system quality, including the construct coverage of the scoring features and the prediction accuracy of the scoring model (Chen et al., 2018; Loukina, Zechner, Chen, & Heilman, 2015; Zechner et al., 2009; Zechner & Evanini, 2019). The engine's speech recognizer provides information about word identity and timing. Speech scientists at ETS have developed more than 100 features that are extracted from the output of the speech recognizer and other signal processing and natural language processing software. These features are consistent with the construct of communicative language ability as embodied in the TOEFL iBT scoring guidelines. They are mainly related to the delivery and language use areas of the TOEFL iBT scoring guidelines for spoken responses, measuring aspects of fluency, pronunciation, prosody, vocabulary, and grammar. There are also some features related to the content and discourse aspects of TOEFL iBT Speaking responses. To build the SpeechRater engine's scoring model, only a subset of the available features is used. The goal here is to select features for a broad coverage of the construct, minimizing features that are highly correlated to other features in the model, and selecting features with high correlations to human rater scores (Loukina et al., 2015). For the current SpeechRater version, the correlation between the SpeechRater scores and human scores was 0.82, while the correlation between two human raters was 0.88 (for section-level scores on the Speaking section with 4 items).

Research on the SpeechRater engine is ongoing, with the goals of (a) improving the accuracy of the speech recognizer, (b) developing features to provide better coverage of the construct, and (c) improving the agreement of the SpeechRater scores with those of human raters.

Explore TOEFL Research

This brief description of a few studies does little to convey the extent of the contribution that ETS and the TOEFL program have made to advancing knowledge of language assessment. Descriptions of more than 200 research studies are available on the TOEFL website, illustrating the program's commitment to advancing the field and meeting high standards for educational measurement. To view these descriptions and download selected reports, visit the TOEFL research website (https://www.ets.org/toefl/research).

References

Alderson, J. C. Test review: Test of English as a Foreign Language™: Internet-based Test (TOEFL iBT®). *Language Testing,* 26(4), 621-631. doi:10.1177/0265532209346371

Barkaoui, K., & Hadidi, A. (2020). Assessing change in English second language writing performance. Routledge.

Beigman Klebanov, B., Madnani, N., Burstein, J., & Somasundaran, S. (2014). Content importance models for scoring writing from sources. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short papers)* (pp. 247–252). Stroudsburg, PA: Association for Computational Linguistics.

Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., & Chodorow, M. (2013). TOEFL11: *A corpus of non-native English* (Research Report No. RR-13-24). ETS. https://doi.org/10.1002/j.2333-8504.2013.tb02331.x

Burstein, J., Flor, M., Tetreault, J., Madnani, N., & Holtzman, S. (2012). *Examining linguistic characteristics of paraphrase in test-taker summaries* (Research Report No. RR-12-18). ETS. http://dx.doi.org/10.1002/j.2333-8504.2012.tb02300.x

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

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.

Cline, F., & Powers, D. E. (2009). *The new generation TOEFL: Evaluating its use with native speakers of English.* Unpublished manuscript.

Davis, L., & Norris, J. M. (Eds.). (2024). Challenges and innovations in speaking assessment (1st ed.). Routledge.

Educational Testing Service. (2024a). Reliability and comparability of *TOEFL iBT scores. TOEFL Research Insight Series Vol. 3*. https://www.ets.org/pdfs/toefl/toefl-ibt-insight-s1v3.pdf

Educational Testing Service. (2024b). TOEFL iBT test framework and test development. *TOEFL Research Insight Series Vol. 1.* https://www.ets.org/pdfs/toefl/toefl-ibt-insight-s1v1.pdf

Educational Testing Service. (2024c). Validity evidence supporting the interpretation and use of *TOEFL iBT scores. TOEFL Research Insight Series Vol. 4*. https://www.ets.org/pdfs/toefl/toefl-ibt-insight-s1v4.pdf

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

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

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

Israel, R., Tetreault, J., & Chodorow, M. (2012). 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* (NAACL-HLT) (pp. 284–294). Stroudsburg, PA: Association for Computational Linguistics.

Kaplan, R. M., Wolff, S. E., Burstein, J. C., Lu, C., Rock, D. A., & Kaplan, B. A. (1998). Scoring essays automatically using surface features (GRE Board Professional Research Report. No. 94-21P). 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.

Loukina, A., Zechner, K., Chen, L., & Heilman, M. (2015). Feature selection for automated speech scoring. In *Proceedings of the 10th Workshop on Innovative Use of Natural Language Processing for Building Educational Applications of the North American Association for Computational Linguistics and Human Language Technologies Conference* (pp. 12–19). Stroudsburg, PA: Association for Computational Linguistics.

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 No. RR-15-25). Princeton, NJ: Educational Testing Service.

Papageorgiou, S., & Manna, V. F. (Eds.). (2023). *Meaningful language test scores: Research to enhance score interpretation*. John Benjamins.

Ramineni, C., Trapani, C., Williamson, D. M., Davey, T., & Bridgeman, B. (2012). *Evaluation of the e-rater scoring engine for the TOEFL independent and integrated prompts* (Research Report No. RR-12-06). Princeton, NJ: Educational Testing Service.

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.

Stricker, L. J. (2002). *The performance of native speakers of English and ESL speakers on the computer-based TOEFL and GRE General Test* (TOEFL Research Report No. 69). 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.

Tetreault, J., Foster, J., & Chodorow, M. (2010). Using parse features for preposition selection and error detection. In *Proceedings of the 2010 Association for Computational Linguistics (ACL 2010) (pp. 353–358). Stroudsburg, PA: Association for Computational Linguistics*.

Wolf, M., & Butler, Y.G. (Eds.). (2017). *English language proficiency assessments for young learners.* Routledge

Zechner, K., & Evanini, K. (Eds.). (2019). *Automated speaking assessment: Using language technologies to score spontaneous speech.* Routledge.

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.