ETS® TOEIC®

Assess to Progress.

LISTENING

SPEAKING

READING

WRITING

The Research Foundation for the *TOEIC*® Tests

A Compendium of Studies: Volume III

Donald E. Powers and Jonathan E. Schmidgall, Editors

TOEIC® Compendium of Studies: Volume III

Section I: Refinement, Revision, Renewal

Section II: Monitoring and Controlling Quality

Section III: Accumulating Evidence to Support Claims: A Validity Argument

Foreword

Organizations around the world have come to recognize that English-language proficiency is a key to global competitiveness. In response, the *TOEIC*® testing program has, since 1979, provided assessments to enable corporations, government agencies, and educational institutions throughout the world to evaluate a person's ability to communicate in English in the workplace. Today, millions of TOEIC tests are administered each year for thousands of organizations in hundreds of countries.

ETS is proud of the substantial research base that supports all of the assessments we offer. Research guides us not only as we develop new products and services but also as we continually improve existing ones, including those in the TOEIC program (e.g., the *TOEIC Bridge*™ test, the *TOEIC*® Listening and Reading test, and the *TOEIC*® Speaking and Writing tests). Offerings like these are essential to meeting our overall mission—to advance quality and equity in education for people worldwide.

This third TOEIC program compendium is a compilation of selected work conducted by ETS Research & Development staff since the second compendium was issued in 2013. The focus continues to be on making certain that TOEIC test scores remain reliable, fair, meaningful, and useful.

As we approach the TOEIC program's 40th anniversary, we are honored to be able to continue to support our clients in the global marketplace. We hope you find this compendium to be useful. As with the previous compendia, we welcome your comments and suggestions.

Ida Lawrence
Senior Vice-President
Research & Development Division
Educational Testing Service

Preface

This compendium is the third in a series that describes the research foundation for the *TOEIC*® assessments. The first volume, published in 2010, focused on three main topics: (a) a major redesign and evaluation of the existing *TOEIC*® Listening and Reading test, (b) the development and evaluation of new tests of speaking and writing, and (c) the (complementary) relationship between the existing and new measures. An overarching theme of the three major topics was the assertion that the most definitive quality of test scores is the validity of the interpretations that follow from them (i.e., the extent to which they are meaningful and useful indicators of the ability that they are designed to measure). For TOEIC, the ability is English-language proficiency in the workplace and in everyday situations. The various papers in this first compendium detail the ways in which test score validity is established and maintained throughout a test's life cycle—from the beginning of its development, to when it is actually used to facilitate decisions about test takers, to when it is revised to keep it up to date and responsive to the needs of test users.

Published in 2013, the second volume of the TOEIC program compendium continued several of the themes discussed in the first volume. Five major sections were devoted to (a) further understanding the relationships among the TOEIC tests, (b) providing information over a wide range of test-taker proficiency, (c) further establishing the meaning of test scores, (d) using test scores appropriately in decision making, and (e) maintaining and improving fairness and test quality. Concern for measurement over a wide range of proficiency levels is evident in papers describing the validation of scores for the *TOEIC Bridge*™ test, a test of the listening and reading skills of beginning and intermediate learners of English. Two papers describe efforts to further establish the meaning of TOEIC scores by mapping them to various benchmarks, performance criteria, or achievement levels, in particular to the widely used levels of the Common European Framework of Reference (CEFR) and to the levels of a lesser known framework developed for the military by the North Atlantic Treaty Organization (NATO). Besides acknowledging the need to establish test score meaning, the second compendium also recognized the need to provide practical guidance on how to use TOEIC scores appropriately. Toward this end, one section documented a set of procedures designed to facilitate the use of TOEIC scores for personnel decisions by enabling test score users to establish defensible cut scores. Finally, two of the papers in the final section focused on the (then) new writing and speaking tests. One described the extensive procedures used to ensure that raters evaluate test takers' responses consistently and accurately. The other described an evaluation of several alternative procedures for identifying tasks on the speaking and writing measures that may unfairly disadvantage some groups of test takers.

This current (third) volume of the TOEIC program compendium documents the major research that has been completed since the second volume was published. The volume begins with a brief history of the origins of the program and its evolution over the 30 some years of its existence. The remainder of the volume contains three sections, each comprising several papers that address a distinct theme. The first section (Refinement, Revision, Renewal) describes efforts concerned with keeping the TOEIC tests up to date, that is, to ensure that they remain well aligned with the most current thinking of language teaching and assessment and how English is generally used in everyday workplace situations. For example, Park and Bredlau describe in "Expanding the Question Formats of the *TOEIC*® Speaking Test" an effort to expand the variety of item formats for the TOEIC Speaking test. Their work is motivated in large measure by the notion of *washback*—that the composition of a test can affect both what is taught and what is learned. Washback can be positive or negative, and one way in which test developers can promote positive washback is by ensuring greater correspondence between test tasks and real world language tasks and situations; by preparing for the test, learners prepare for real-world communication. In their study, Park and Bredlau revisit the original test design and develop comparable additional variants for several of the fundamental TOEIC Speaking task types. Insofar as test takers would be expected to demonstrate their use of English in a wider range of situations, a greater variety of texts and topics was believed to better foster the development of communicative competence (and to discourage the memorization of task types).

In "Background and Goals of the *TOEIC*® Listening and Reading Update Project," Ashmore, Duke, and Sakano describe a study of the TOEIC Listening and Reading test that was conducted to identify any areas of linguistic competence that may have been underrepresented by the (then) current version of the test. The ultimate objective was to modify the existing listening and reading tasks in order to reflect changes in communication styles in today's workplace, such as the increasing use of electronic communication. Secondarily, the researchers explored the prospect of increasing the feedback provided to test takers and score users. As a result, pragmatic understanding has been added to the abilities measured by the TOEIC Listening test.

Both of the revision efforts described in the papers by Park and Bredlau and Ashmore et al. required empirical research to assess the effects of the proposed test modifications. These efforts are documented by Cid, Wei, Kim, and Hauck in "Statistical Analyses for the Updated *TOEIC*® Listening and Reading Test" and in "Statistical Analyses for the Expanded *TOEIC*® Speaking Test" by Qu, Cid, and Chan. Both of these evaluations were based on similar concerns—that the proposed modifications would produce (a) items with acceptable psychometric qualities and (b) test scores that could be appropriately compared with those from previous versions of the tests. Study results revealed that psychometric standards have been maintained for the revised tests. Slight differences in difficulty levels were addressed, where needed, by making appropriate adjustments to some of the new items. By monitoring operational data gathered since the launch of the updated tests, the comparability of the earlier and the updated test versions has been corroborated.

To meet the need for test security, the TOEIC program requires a substantial pool of test items from which multiple, comparable test forms can be assembled each year. This need has inspired attempts to increase the efficiency of item development while maintaining quality. The final paper in the first section ("Analyzing Item Generation With Natural Language Processing Tools for the *TOEIC*® Listening Test" by Yoon and colleagues) documents the development of automated tools to support this need for the *Listening* section of the TOEIC Listening and Reading test. These tools have been designed to help item writers by providing initial ideas, authentic language, and support for adjusting the variety and complexity of vocabulary in listening items. Item writers have found the tools to be useful, and they are now being used operationally.

The second major section (Monitoring and Controlling Quality) contains four papers dealing with several perennial quality control issues primarily related to the reliability or consistency of test scores. Three of the papers concern either the TOEIC Speaking test or the TOEIC Writing test. Unlike the TOEIC Listening and Reading test, which requires test takers to select responses that can be objectively scored by computer, the TOEIC Speaking and Writing test measures require test takers to construct responses that must be subjectively evaluated by human raters. The use of subjective scoring poses a variety of additional challenges, some of which are addressed by the efforts described in his section. A second feature of several of the papers in this section is their use of longitudinal data from test takers who take the TOEIC tests on multiple occasions over time. Several of the papers demonstrate how data from repeat test takers can be used effectively to monitor important aspects of the ongoing program.

In "The Consistency of *TOEIC*® Speaking Scores Across Ratings and Tasks," Schmidgall reports on an analysis using generalizability theory to provide information about the consistency of TOEIC Speaking scores across different aspects of the scoring procedure. Results revealed that, at the lowest level (individual tasks), most of the variation in scores can be explained by individual ability as opposed to differences between ratings. Most importantly, variation at the level of total scores is explainable largely by test takers' ability rather than by differences between ratings. In total, the results reveal the consistency of TOEIC Speaking scores, suggesting that they are determined largely by speaking proficiency rather than by any prominent features of the testing procedure that should not affect test scores.

Consistency of test scores is also a theme in "Evaluating the Stability of Test Score Means for the *TOEIC*® Speaking and Writing Tests" by Qu, Huo, and Chan. For the TOEIC assessments, it is critical to maintain consistency of various facets of the scoring procedure but also to understand the causes of any variation in test scores over time. The aim here is to ensure that interpretations about test takers' abilities are comparable from one administration (or form) to another. Using several statistical procedures, Qu and colleagues examined the stability of average TOEIC Speaking and Writing test scores for several hundred test forms administered over a 3-year period. Results indicated that fluctuations in test score averages reflect mainly real changes in test takers' speaking (or writing) ability. For both TOEIC Speaking and Writing test scores, a large proportion of the variation in score means was explained by such factors as *seasonality* (i.e., the tendency for more able test takers to take the test at particular times of the year). This finding provides evidence for the consistency of the TOEIC Speaking and Writing score scales across forms.

In "Monitoring Score Change Patterns to Support *TOEIC*® Listening and Reading Test Quality," Wei and Low examine test score consistency by analyzing the score change patterns of some 20,000 test takers (so-called test repeaters) who had taken the TOEIC Listening and Reading test at least six times over a 4-year period. The observed patterns support the assertions that TOEIC Listening and Reading scores are consistent and reliable over time and across administrations and that they are valid indicators of growth in test takers' English proficiency.

In developing multiple forms of the TOEIC Speaking Test, the current practice is to adhere to strict test specifications in order to ensure that, in terms of content and difficulty, each new form of the test is comparable to previously used forms. However, because slight differences in the difficulty of alternate forms may still occur, a statistical procedure known as test score equating is commonly used to adjust for any between-form differences in difficulty.

The focus of "Linking *TOEIC*® Speaking Test Scores Using *TOEIC*® Listening Test Scores" by Kim is maintaining the comparability of test forms across time and administrations. Kim reports an investigation that compares the current method of equating the TOEIC Speaking test with an alternative procedure that uses TOEIC Listening scores as the basis for adjusting TOEIC Speaking scores. The results suggest that the currently used procedure remains a practical choice for maintaining the comparability of TOEIC Speaking test forms over time.

The third major section (Accumulating Evidence to Support Claims: A Validity Argument) contains three papers describing efforts to generate evidence to support the various claims that are made for the TOEIC tests and to organize this information systematically in the form of a "validity argument." In "Articulating and Evaluating Validity Arguments for the *TOEIC*® Tests," Schmidgall addresses the question "How can it be determined whether a test is suitable for the purpose for which it was designed?" This fundamental question is motivated in large part by the view that test developers must convince stakeholders (i.e., anyone affected by the test) that the intended use of a test is appropriately justified. This view is formalized in the argument-based approach to justifying test use. Schmidgall provides an accessible introduction to the argument-based approach, its implementation for TOEIC tests, and its perceived benefits for stakeholders. Overall, the paper describes the approach that TOEIC research takes to support appropriate uses of the TOEIC tests.

The way in which TOEIC scores are used is also the subject of "The Case of Taiwan: Perceptions of College Students About the Use of the *TOEIC*® Tests as a Condition of Graduation" by Hsieh, who queried Taiwanese college students about their perceptions of TOEIC test scores being used to meet an English-language graduation requirement. Results indicated that, in general, students have positive views about the use of TOEIC test scores for graduation, and they believe that preparing to take the test has a positive impact on their language proficiency and future employment prospects. The study provides empirical evidence to support the use of TOEIC test scores as a college exit requirement in Taiwan and, arguably, for similar use in other countries.

Finally, in "Insights Into Using *TOEIC*® Test Scores to Inform Human Resource Management Decisions," Oliveri and Tannenbaum document their insights into TOEIC test use in another context—to inform personnel decision making. An analysis of stakeholders' use of TOEIC scores was viewed as a basis for supporting meaningful score interpretations and relevant score-based human resource decision making. Toward this end, this paper documents how managers currently tend to use TOEIC scores to inform hiring, promotion, and training decisions in the international workplace. The paper concludes by providing suggestions for future research and for possible services to test score users.

In total, the various individual papers highlight the rigorous, systematic, and evolving contribution of research to the TOEIC tests. As summarized in "Articulating and Evaluating Validity Arguments for the *TOEIC*® Tests," TOEIC research has incorporated an argument-based approach to validity that is used to monitor wide-ranging claims about the measurement quality and use of TOEIC tests. This approach begins with claims about the reliability or consistency of test scores. Test takers and score users can continue to have confidence in the consistency of TOEIC test scores across raters, tasks, test forms, and occasions of testing as demonstrated in the papers by Yoon et al.; Qu, Huo, and Chan; Wei and Lei; and Kim. A diverse group of experts in test development, psychometric analysis, and research help ensure the TOEIC tests continue to provide meaningful interpretations about English ability through the updates and enhancements described in the papers by Park and Bredlau; Ashmore et al.; Cid et al.; and Qu, Cid, and Chan. And the studies reported in the final two papers by Hsieh and by Oliveri and Tannenbaum in this compendium show how TOEIC research is investigating how TOEIC tests are used and the potential consequences of these uses. Thus, the papers in this compendium address a variety of discrete, but interrelated aspects of the TOEIC assessments. Each contributes in some way to supporting the use of TOEIC scores from each of its component tests. As we issue this third compendium, research is already well underway for a fourth volume of TOEIC research.

Donald E. Powers
Jonathan Schmidgall

Compendium Study

The *TOEIC*® Tests: A Brief History

Donald E. Powers and Jonathan Schmidgall

The *TOEIC*® program was conceived in the late 1970s when Japanese university professor Yasuo Kitaoka foresaw the need for an assessment of the ability to use English in workplace settings. In conjunction with the Japanese Ministry of International Trade and Industry, Professor Kitaoka contacted ETS to share his vision. The result was the development, and in 1979 the first administration, of the *TOEIC*® Listening and Reading test (Woodford, 1982). From its modest beginning, the TOEIC program has grown significantly and now annually serves some 7 million test takers in approximately 150 countries.

The TOEIC program's evolution has, perhaps, been quite predictable. Increasingly, today's global economy requires workers who are proficient in English. In fact, English appears to have emerged as the unofficial language of international commerce (Michaud, 2013; Nickerson, 2013). Moreover, with the vast majority of all scientific papers now published in English, English has also become the predominant language of science and technology (Montgomery, 2013; Orr, 2013; Parkinson, 2013). Thus, the need perceived by Professor Kitaoka has only increased over the years: Workers who can communicate proficiently in English are a greatly valued commodity. Consequently, to help prepare their citizens, governments around the world are promoting, and even mandating, the teaching of English in their schools ("Opportunities Abound," 2012). A resulting challenge is to help international companies (a) recruit and select individuals who have the requisite English-language skills and (b) identify those individuals whose skills are likely to improve with further instruction or training. The aim of individuals *seeking* employment is to demonstrate their English-language skills to international companies, thereby increasing their attractiveness as prospective employees.

The TOEIC tests have long played a central role in the context described above. For nearly the first thirty years of its existence, the TOEIC program offered a single test assessing only reading and listening skills. This test underwent a significant redesign in 2003 in order to ensure that it remained aligned with the most current theories of language, especially with the recognition that (a) language is used in context and (b) real-world communication typically requires the simultaneous engagement of multiple language skills. The redesign effort also afforded the opportunity to enhance test score interpretation. This enhancement was accomplished by analyses that identified the specific skills and deficiencies of test takers at various TOEIC score levels. The results of both the redesign effort and the concomitant score interpretation analyses have been documented in the first TOEIC program compendium (Educational Testing Service, 2010).

Despite high demand for the TOEIC Listening and Reading test, the selective coverage of only reading and listening occasionally gave rise to complaints from TOEIC users, who observed that test takers sometimes achieved high test scores despite being seriously deficient in their overall ability to communicate in English. In response to such concerns, in 2004 the TOEIC program undertook the development of standardized measures of productive language skills, that is, speaking and writing. The effort took account of various factors such as market needs, business requirements, and design issues. For instance, concern for test security dictated the need for a significant number of parallel test forms, which in turn required detailed test specifications that could be clearly communicated to, and efficiently implemented by, test developers. An additional requirement (which is not a concern for the objectively scored multiple choice listening and reading items) was that the responses elicited from test takers should be amenable to consistent and accurate subjective scoring by trained raters. Research addressing these issues is documented in the first TOEIC program compendium, as is research to elucidate the meaning of scores from these new measures once they became fully operational (Powers, Kim, Yu, Weng, & van Winkle, 2009). Thus, after first offering the new tests in 2006, the TOEIC program now provides measures of English proficiency in all four language domains: speaking, listening, writing, and reading.

Because the construct assessed by TOEIC tests (the ability to use English in a workplace setting) is complex and multifaceted, it's impossible to measure every important aspect in the relatively limited time that is available for testing. Instead, the aim of TOEIC program developers is to ensure that those facets thought to be most critical are covered in proportion to their importance in the workplace. The detailed test specifications, or blueprints, that guide TOEIC assessment developers are intended to meet this goal. As documented in the paper by Park and Bredlau and in the paper by Ashmore, Duke, and Sakano in the current compendium, these blueprints are revisited periodically in order to ensure that they meet the most current state-of-the-art expectations in language assessment.

The current TOEIC tests assess a reasonably diverse sample of the everyday English skills used in the international workplace. Focusing on skills thought to be essential in this environment, the tests provide opportunities for test takers to demonstrate their proficiency in a variety of ways using each of the four traditional language skills. Although testing is set in the context of everyday and workplace communication, the focus is on general purposes rather than workplace-specific knowledge and skills.

Except in some regions where it is computer-administered, the TOEIC Listening and Reading test is a paper-and-pencil multiple-choice test based on both aural and written stimuli. Listening questions are based on a variety of statements, questions, conversations, and talks recorded in English. Reading questions require understanding of a variety of written material. The TOEIC Listening and Reading test questions are also designed to measure skills like listening for a purpose and reading for required details, which are needed in order to communicate in the real world.

The speaking and writing tests are computer-based tests that require test takers to construct responses. The speaking test contains 11 spoken or written prompts that require oral responses such as (a) reading a text aloud, (b) describing pictures, (c) responding to questions after reading a short text and listening to a related spoken text, (d) listening to a short spoken text and then proposing a solution, and (e) expressing an opinion on a specific topic. The writing test contains eight written prompts that require test takers to (a) write a sentence based on a picture, (b) respond to written requests in an e-mail format, and (c) write an argument that states, explains, and supports an opinion. Speaking and writing tasks are expressly designed to elicit the most important aspects of communicative ability. In addition, when test-taker responses are scored, raters are trained explicitly to evaluate these aspects—for example, task completion, organization, vocabulary use, and correct grammar. In short, the TOEIC tests provide a sampling of the most important aspects of speaking, writing, reading, and listening skills. (For more details, visit https://www.ets.org/toeic/test-takers/listening-reading/about/content-format and https://www.ets.org/toeic/test-takers/speaking-writing/about/content-format.)

The availability of the new speaking and writing tests prompted additional researchable questions. For example, there was interest in establishing the unique contribution of each of the new measures beyond the existing listening and reading test and in learning how each of the four measures complement one another. These questions have been addressed both logically and empirically in the first compendium (Liao, Qu, & Morgan, 2010).

The validity of test scores is a concern not only when tests are developed, but also when they are scored. For the multiple-choice listening and reading questions, computers enable straightforward, completely objective scoring. For the speaking and writing tests, however, scoring is necessarily more subjective. As mentioned earlier, in order to accommodate the inherent subjectivity involved in scoring test-taker-produced responses, trained raters use detailed scoring guidelines to evaluate test takers' responses. Only raters who meet certain qualifications and pass a rater certification test are hired, and they are trained to apply rigorous scoring guidelines. Moreover, each day before they begin scoring, raters must pass a calibration test to demonstrate that they have maintained their scoring accuracy. In addition, raters' performance is monitored continuously in real time, to ensure that accuracy is maintained and that scoring guidelines are applied consistently. Further details about these procedures have been provided by Everson and Hines (2010) in the first TOEIC program compendium. Raters' scores are also subjected to a variety of other statistical analyses and quality control procedures, as described in several of the papers in the current compendium ("The Consistency of *TOEIC*® Speaking Scores Across Ratings and Tasks" by Schmidgall, "Evaluating the Stability of Test Score Means for the *TOEIC*® Speaking and Writing Tests" by Qu, Huo, and Chan, and "Monitoring Score Change Patterns to Support *TOEIC*® Listening and Reading Test Quality" by Wei and Low).

One critical kind of evidence of test score validity is the extent to which test scores relate to various criteria of success. For the TOEIC tests, one such criterion is on-the-job performance with respect to the ability to perform tasks that require English-language skills. Another criterion that has been employed for the TOEIC tests is self-assessment by the test takers themselves, who have been asked how well they can perform a variety of different language tasks in English. Research has shown that test takers' responses to these self-assessments are reasonably trustworthy, and so the extent to which these reports agree with TOEIC scores is evidence that TOEIC scores are meaningful indicators of English proficiency. The results of this research have been documented in the first TOEIC program compendium (Powers et al., 2008; Powers et al., 2009). An alternative approach to anchoring the meaning of TOEIC scores has been to map them to levels of English-language proficiency as specified in widely accepted frameworks such as the Common European Framework of Reference (CEFR). This approach is exemplified in studies such as one conducted by Tannenbaum and Wylie (2013).

Besides the need to establish firmly what test scores mean, professional standards are also concerned with the actual consequences of testing, including, for example, so-called *washback*. Though somewhat difficult to research rigorously, washback is an established phenomenon: Tests can and do influence teachers and learners to engage in activities that either facilitate or inhibit language learning (e.g., Choi, 2008; Messick, 1996). For instance, a focus mainly on test-taking skills and test-question formats may improve test scores without improving language proficiency. In addition, focusing disproportionately on some language skills at the expense of others may result in uneven profiles of proficiency. There is good reason to believe that, by tapping skills in all four domains, the TOEIC tests can contribute to improving overall English-language proficiency: Positive washback is more likely for test takers who prepare for all four TOEIC measures than for those who prepare more selectively (Stoynoff, 2009). How test takers prepare for and approach testing is determined to some degree by their perceptions of the tests, which may also be an important factor in determining other consequences of testing. Such perceptions are the focus of a study in the paper by Hsieh in the current compendium.

Arguably, the most consequential effects of the TOEIC tests, however, concern their use by employers to make high-stakes decisions about prospective employees. The effect of test-based decisions is, for a number of reasons, generally far less researched than are a variety of other important topics in standardized testing. A study by Oliveri and Tannenbaum (a paper in the current compendium) makes modest inroads into the manner in which TOEIC scores are used in practice to make personnel decisions.

Each kind of evidence mentioned above provides partial justification for the use of TOEIC scores. However, as important as these distinct aspects of support may be, they are not entirely sufficient. Still needed is a cohesive argument that incorporates the various kinds of support. Providing such an argument is an ongoing process that, as documented in "Articulating and Evaluating Validity Arguments for the *TOEIC*® Tests," a paper by Schmidgall in this compendium, is now well underway.

References

Choi, I.-C. (2008). The impact of EFL testing on EFL education in Korea. *Language Testing*, *25*, 39–62. https://doi.org/10.1177/0265532207083744

Educational Testing Service. (2010). *The research foundation for the TOEIC® tests: A compendium of studies.* Princeton, NJ: Author.

Everson, P., & Hines, S. (2010). How ETS scores the *TOEIC*® Speaking and Writing tests' responses. In *The research foundation for the TOEIC® tests: A compendium of studies* (pp. 8.1–8.9). Princeton, NJ: Educational Testing Service.

Liao, C.-W., Qu, Y., & Morgan, R. (2010). Relationships of test scores measured by the *TOEIC*® Listening and Reading Test and the *TOEIC*® Speaking and Writing Tests. In *The research foundation for the TOEIC tests: A compendium of studies* (pp. 13.1–13.15). Princeton, NJ: Educational Testing Service.

Michaud, C. (2013, May 16). *English the preferred language for world business: Poll.* Retrieved from http://www.reuters.com/article/2012/05/16/us-language-idUSBRE84F0OK20120516

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*, 241–256. https://doi.org/10.1177/026553229601300302

Montgomery, S. L. (2013). *Does science need a global language? English and the future of research.* Chicago, IL: University of Chicago Press. https://doi.org/10.7208/chicago/9780226010045.001.0001

Nickerson, C. (2013). English for business. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics.* New York, NY: Blackwell Publishing.

Opportunities abound in digital English language learning market. (2012, October). *ICEF Monitor.* Retrieved from http://monitor.icef.com/2012/10/opportunities-abound-in-digital-english-language-learning-market

Orr, T. (2013). English for science and technology. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics.* New York, NY: Blackwell Publishing.

Parkinson, J. (2013). English for science and technology. In B. Paltridge & S. Starfield (Eds.), *The handbook of English for specific purposes* (pp. 155–173). New York, NY: John Wiley.

Powers, D. E., Kim, H.-J., & Weng, V. Z. (2008). *The redesigned TOEIC® Listening and Reading test: Relations to test-taker perceptions of proficiency in English* (Research Report No. RR-08-56). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2008.tb02142.x

Powers, D. E., Kim, H.-J., Yu, F., Weng, V. Z., & vanWinkle, W. (2009). *The TOEIC® Speaking and Writing tests: Relations to test-taker perceptions of proficiency in English* (Research Report No. RR-09-18). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2009.tb02175.x

Stoynoff, S. (2009). Recent developments in language assessment and the case of four large-scale tests of ESOL ability. *Language Teaching*, *42*, 1–40. https://doi.org/10.1017/S0261444808005399

Tannenbaum, R. J., & Wylie, C. E. (2013). Mapping TOEIC and *TOEIC Bridge*™ test scores to the Common European Framework of Reference. In D. E. Powers (Ed.), *The research foundation for the TOEIC® tests: A compendium of studies* (Vol. 2, pp. 6.1–6.10). Princeton, NJ: Educational Testing Service.

Woodford, P. E. (1982). *The Test of English for International Communication*™. In C. Brumfit (Ed.), *English for international communication* (pp. 61-72). Oxford, England: Pergamon Press.

Compendium Study

Expanding the Question Formats of the *TOEIC®* Speaking Test

Elizabeth Park and Elizabeth Bredlau

The *TOEIC*® program has assessed the English-language proficiency of nonnative speakers of English since 1979. Used in 150 countries to inform hiring, employee placement and promotion, training, and learning progress, the TOEIC tests measure the everyday English-language skills of people currently working in international settings or preparing to enter the global workforce. Originally testing only the receptive skills of listening and reading, the TOEIC program introduced the *TOEIC*® Speaking and Writing tests in 2006, responding to the market's need for fair, valid, and reliable assessments of productive English-language skills.

The composition of a test can affect both what is taught and what is learned. If a test is too narrow in its scope, teaching and learning may become correspondingly narrow. Hence, an obstacle of the effort described here was to expand the scope of the *TOEIC*® Speaking test, at least modestly, in order to decrease the likelihood of instruction that is constricted or preparation that is geared solely toward a limited number of task types. Companies and other institutions that use the TOEIC tests are interested in whether or not the test taker has demonstrated necessary English-language communicative skills.

However, test-preparation strategies that favor teaching mastery of a short list of communication tasks and scenarios may end up ignoring the wide range of skills necessary for English-language proficiency, and test takers who rely on rote memorization and test-taking strategies may fall short of the communicative competence desired by employers and educational institutions.

For tests to effect positive change in learning, they must encourage learning. One way to foster communicative language learning rather than memorization is for tests to present a variety of topics and texts. However, the need for variation must be balanced with the need for validity and reliability.

Detailed specifications for developing the tasks, or types of questions on the test, can assist in the development of valid and reliable tests. The templates used to generate parallel test questions, also known as task blueprints, describe the elements, rubric, and range of acceptable variants for a given task type.

Although the original pilot in 2006 confirmed the viability of the TOEIC Speaking test design, a group of content experts reviewed the TOEIC Speaking task blueprints in 2013 to evaluate how well the current specifications had succeeded in balancing the need for specificity with the desire for variety. Of particular note during the review were the lists of variants generated during the original design phase to illustrate a range of topic areas and types of texts that could be used when developing test questions for a task. While the fixed elements (e.g., nature of the task) of the blueprint focus on the fundamental structures (e.g., propose a solution based on a problematic situation) that are shared by all questions of the same task type, the lists reviewed by the expansion team present variants, or options for features that may vary from question to question. These lists of acceptable variations in question formats include a diverse range of possible topics areas and text types, and the lists were intended to be helpful, but not exhaustive. As such, would it be possible to expand the list of variants with alternate but comparable options? A team of content experts, statisticians, and product managers was formed to consider this possibility.

The purpose of this paper is to document the process of developing an expanded list of test question variants for the TOEIC Speaking test—specifically, a reexamination of the original test design and task blueprints as well as a summary of the prototyping and piloting phases.

Revisiting the Original Test Design Analysis

Content experts began by revisiting the original test design analysis (TDA) conducted in 2005. Any new variant, no matter how similar to existing variants, would need to flow naturally from the test design. Derived from the principles of evidence-centered design, the six steps of the TDA process serve as the foundation for the templates used to generate parallel tasks, also known as task blueprints (Hines, 2010). The original analysis steps were as follows in Table 1.

Table 1

Example Task Design Analysis for Speaking

Step in task design analysis	Outcome for the speaking test
Reviewing previous research and other relevant assessments	Ideas about language proficiency and potential test tasks
Articulating claims and subclaims:	Claim: The test taker is able to communicate in spoken English, which is needed to function effectively in the context of a global workplace. Subclaims 1. The test taker can generate language intelligible to native and proficient nonnative English speakers. 2. The test taker can select appropriate language to carry out routine social and occupational interactions (such as giving and receiving directions; asking for and giving information; asking for and providing clarification; making purchases; greeting and introductions; etc.). 3. The test taker can create connected, sustained discourse appropriate to the typical workplace.
Listing sources of evidence	Task appropriateness, delivery, relevant vocabulary and use of structures
Listing real world tasks in which test takers can provide relevant evidence	Asking and responding to questions based on written information in a workplace setting, participating in a discussion that requires problem solving, and exchanging information one-on-one with colleagues, customers, or acquaintances
Identifying aspects of situations that would affect their difficulty	Characteristics of reading and listening material; the nature of their connections to each other (referring to Subclaim 2)
Identifying criteria for evaluating performance on the tasks	Range and complexity of vocabulary and structures; clarity and pace of speech; coherence and cohesion; progression of ideas in response; relevance and thoroughness of the content of the response

Note. From "Evidence-Centered Design: The *TOEIC*® Speaking and Writing Tests" (p. 7.8), by S. Hines, 2010, in *The research foundation for the TOEIC® tests: A compendium of studies*, Princeton, NJ: Educational Testing Service. Used with permission.

A desired outcome of the proposed expansion was to retain comparability with the original formats; thus the statements, or claims, the test makes about test taker's performance should remain the same. As there were no changes to those statements in step two of TDA, the evidence needed to support the test claim and subclaims would remain unchanged from step three of the original design. The proposed alternate question formats should require test takers to demonstrate the same proficiencies in the same language skills as the original formats.

Upon reexamining step four from the original analysis, the exploratory team decided not to revise the list of real-world tasks. Effective communication in a global workplace continues to require speakers to participate in discussions, to solve problems, and to exchange information one-on-one. Similarly, the design team decided not to change the factors contributing to the difficulty of communication tasks listed in step five of the original analysis. Finally, as the claims, subclaims, evidence, tasks, and aspects of situations affecting difficulty would remain the same, so, too, should the scoring criteria expressed in step six of the original TDA.

After reviewing the analysis that informed the creation of the TOEIC Speaking test, it became clear that for the current and proposed question formats to be comparable, the claims, evidence, task, and scoring criteria should not vary from the original test design. Any proposed variation on question formats should be firmly rooted in the original test design analysis.

Reviewing Task Specifications and Expanding the List of Variants

Having confirmed that the general design of the tasks and test should not be altered, the expansion team focused its attention on the detailed task specifications (i.e., task blueprints). These useful tools help to ensure that necessary elements, as determined by the TDA process, are included in each test question. To achieve this goal, task blueprints articulate four components: (a) the fixed elements (e.g., nature of the task, order of question elements) common to all questions of the same task type; (b) the elements, such as topic and type of the stimulus text, that can acceptably vary from question to question, also known as variable elements; (c) the rubric; and (d) the list of variants that provide possible options for the variable elements. For example, a task blueprint that lists the category of *topic* as a variable element could list advertisements, entertainment, health, shopping, and travel as possible variants.

In the interest of comparability, the expansion team agreed that the fixed elements, or the aspects that are always associated with this type of task, should remain the same as the original. Thus, the nature of the task and sequence of its elements are unchanged. Similarly, to maintain comparability, the expansion team determined that no modifications to the rubric should occur. For example, in the current Propose a Solution (Test Question 10) task type, the following elements are fixed and should appear in the same order in all test questions within this class of task:

1. Test takers listen to an extended audio stimulus, approximately 120 to 135 words in length, which presents a problem or issue.

2. Test takers have 30 seconds to prepare a response.
3. Test takers have 60 seconds to:
    - use connected, sustained discourse appropriate to the typical workplace,
    - summarize the aforementioned problem or issue, and
    - propose a solution to the aforementioned problem or issue.
4. Test taker's responses are scored by qualified, trained, certified, and calibrated raters, using a rubric with a range of points between 0 and 5.

Reviewing the final two components of the task blueprint (variable elements and variants), content experts hypothesized that it was feasible to expand the current list of possible varieties with comparable topics and types of listening and reading stimuli. As the variants contained within the blueprint were intended to be illustrative and not restrictive, the expansion team proposed adding different but parallel options to the list. To do so, content experts considered typical real-world communication tasks that occur in daily life and the international workplace, this time to explore diversity in settings, situations, and topic areas.

Prototyping and Piloting

Content experts began the prototyping phase by identifying the different ways someone in the context of a global workplace could participate in the real-world tasks from step four of the test design analysis. In what scenarios might a person need to participate in discussions in order to solve problems? What are the different forms in which two people might exchange information?

Using the evidence paradigm as a starting point, test developers listed different types of problem-solving discussions and one-on-one informational exchanges that routinely occur in global workplaces and daily life. Table 2 contains some of the communication formats considered.

Table 2

Sample Communication Formats of Real-World Tasks

Real-world tasks (Step 4 of TDA)	Communication formats
Participating in a discussion that requires problem solving	• conversations with one or more speakers • meetings with one or more speakers • teleconferences with one or more speakers • voicemail messages from one or more speakers • telephone conversations with one or more speakers • radio talk shows with one or more speakers • etc.
Exchanging information one-on-one with colleagues, customers, or acquaintances	Face-to-face conversations • friend talking to a friend • employee talking to a boss • company or organization talking to a client • etc. Telephone conversations • customer talking to a company • market researcher talking to a person • friend talking to a friend • company or organization talking to a client • family member talking to a family member • employee talking to an employee • etc.

Questions exploring the alternate communication formats in Table 2 were developed, and the prototypes were tried out among a group of content experts specializing in English-language education and assessment. Based on the feedback from the content experts, two key workplace communication tasks—telephone calls and meetings—were identified as being:

- most meaningful for the test-taking population and clients of the TOEIC Speaking test, and
- most comparable to the current formats for the *Respond to Questions* (test questions 4–6) and *Propose a Solution* task types.

Per step three of TDA, the successful completion of authentic workplace tasks, such as participating in a problem-solving discussion or exchanging information one-on-one, provides evidence in support of the test claims. As these types of discussions are likely to occur via voicemail messages as well as meetings with one or multiple speakers, expanding the *Propose a Solution* task type to include meetings seemed to align well with the original design as well as allow the score users to gather meaningful evidence from test takers regarding their ability to discuss and communicate solutions to problems in the global workplace. Figure 1 presents a sample of a *Propose a Solution* test question generated from the expanded list of variants.

Respond as if you work with Melanie in the event-planning department at the hotel.

In your response, be sure to:

- show that you recognize the problem, and
- propose a way of dealing with the problem.

Script for Audio Stimulus:

(Woman): Before we end our event-planning meeting, let's discuss a problem with an upcoming reservation at our hotel. I just talked to Ms. Ortega to confirm her reservation for the Lake Room for a family reunion on June first. It seems like we have double booked the Lake Room for that day.

(Man): That's right, Melanie. The Stevens Company reserved that room for an awards ceremony on the same night. And even though we have other rooms available, none of them are big enough for either group.

(Woman): We need to fix this problem of the Lake Room being double booked, but our meeting time is over. I'd like everyone to call me later with a detailed plan for how we should solve this problem.

***Figure 1*. Example of *Propose a Solution* alternate question format.**

Similarly, for another task type, as information is likely to be exchanged among colleagues, customers, and acquaintances as well as market researchers, adding these different parties to the list of *Respond to Questions* seemed a suitable route. Figure 2 presents a sample of a *Respond to Questions* task generated from the expanded list of variants.

Imagine that a friend will be moving to your neighborhood. You are having a telephone conversation about where you live.

Question 4: How many grocery stores are in your neighborhood, and can you walk to them?

Question 5: What's the best time of day to go to the grocery store, and why?

Question 6: Do you usually buy all your groceries from the same store? Why or why not?

***Figure 2*. Sample of *Respond to Questions* alternate question format.**

Following prototyping, the expansion team entered the pilot phase. Using information from the prototyping stage and input from content experts, the lists of variants for the *Propose a Solution* and *Respond to Question* tasks were expanded to include the alternates. Using the longer lists of variants, alternate question formats were developed and included in two pilot test forms (Forms B and C). In order to confirm the comparability of the alternate formats, the two pilot forms and one form (Form A) using only the current question formats were administered to 992 candidates from Korea and Taiwan between October and November of 2013. Responses from the pilot study were scored by qualified, certified, trained, and calibrated TOEIC Speaking test raters. Although modifications to existing task types appeared to affect the difficulty of the alternate question formats (some new variants proved somewhat more difficult and others somewhat less difficult), the effects tended to cancel out when aggregated at higher levels of performance (i.e., total score and scores for claims based on multiple items; see "Statistical Analyses for the Expanded *TOEIC*® Speaking Test" by Qu, Cid, and Chan in the present compendium). Furthermore, the effects observed in the rigorously designed study were within the range of variation typically seen across multiple, parallel forms of the TOEIC Speaking test.

Nonetheless, in order to ensure that test form difficulty is controlled as tightly as possible, performance on the modified question types should be subjected to ongoing monitoring.

Based on the results of the pilot study, test developers have added examples of alternate variants listed in task blueprint used by item writers. The rubrics and the fixed elements of the task remain the same.

Next Steps

The primary objective throughout the expansion project was to ensure the positive effect of the TOEIC Speaking test on learners and score users. An expanded list of variants with new and comparable formats representative of routine workplace communication tasks allows test takers the opportunity to demonstrate proficiency in a range of situations. Score users similarly benefit from knowing that the test scores are based on evidence that the test taker can effectively use spoken English in a greater variety of authentic communication activities. By periodically evaluating real-world communication tasks in everyday life and in the international workplace, the TOEIC program can continue to meet the needs of score users and English-language learners by making informed adjustments to question formats that move language learning forward. With the pilot study confirming the expanded question formats as comparable, next steps include informing test takers, score users, and the public about the expansion.

References

Bailey, K. M. (1999). *Washback in language testing* (TOEFL Monograph No. 15). Princeton, NJ: Educational Testing Service.

Hines, S. (2010). Evidence-centered design: The *TOEIC*® Speaking and Writing tests. In *The research foundation for the TOEIC® tests: A compendium of studies* (pp. 7.1–7.31). Princeton, NJ: Educational Testing Service.

Powers, D. E. (2010). Validity: What does it mean for the *TOEIC*® tests? In *The research foundation for the TOEIC® tests: A compendium of studies.* Princeton, NJ: Educational Testing Service.

Appendix. Summary of Specifications for Speaking Measure

Speaking claim	Test taker can communicate in spoken English to function effectively in the context of a global workplace.					
Subclaims	Test taker can generate language intelligible to native and proficient nonnative English speakers.		Test taker can select appropriate language to carry out routine social and occupational interactions (such as giving and receiving directions; asking for and giving information; asking for and providing clarification; making purchases; greetings and introductions, etc.)		Test taker can create connected, sustained discourse appropriate to the typical workplace.	
Nature of speaking task	Read a text aloud	Describe a picture	Respond to a short question based on a personal experience in the context of a telephone market survey [or telephone call]	Respond to short questions based on information from a written schedule/ agenda	Propose a solution based on a problematic situation stated in the context of a voice mail message [or meeting]	Describe and support opinion with respect to a given pair of behaviors or course of action
Scoring rubric	Analytic 0–3	Independent 0–3	Integrated 0–3	Integrated 0–3	Integrated 0–5	Integrated 0–5
Number of questions	2	1	3	3	1	1
Nature of stimulus material	Reading text that contains: • complex sentence • list of three items • transition • 40–60 words Text must be accessible to low-level speakers	Photograph that represents high-frequency vocabulary or activities	Listening stimuli made up of three short, related questions that are both seen and heard by the candidate; lead-in sets context for the topic of the questions; voices represent English speaking voices from the United States, Australia, Britain and Canada	Reading passage: Telegraphic text in the form of an agenda or schedule (65–75 words; 12 line max). Listening stimulus: Three short questions based on a written schedule. Q1 asks about basic information. Q2 is based on an incorrect assumption or requires the test taker to make an inference. Q3 is a summary of multiple pieces of information.	Listening stimulus: Voice mail message [or meeting] that represents a problem or issue that requires the test taker to summarize and propose a solution (120–135 words).	Listening stimulus: Prompt that is both seen and heard and requires test taker to take stance on an issue or topic.
Prep time	45 seconds	30 seconds	0 second	0 second	30 seconds	15 seconds
Response time	45 seconds	45 seconds	15, 15, and 30 seconds	15, 15, and 30 seconds	60 seconds	60 seconds
Total time	Approximately 30 minutes for 11 questions					

Note. Alternate question format descriptions in brackets. From "Evidence-Centered Design: The *TOEIC®* Speaking and Writing Tests" (p. 7.15), by S. Hines, 2010, in *The research foundation for the TOEIC® tests: A compendium of studies*, Princeton, NJ: Educational Testing Service. Adapted with permission.

Compendium Study

Background and Goals of the *TOEIC®* Listening and Reading Update Project

Elizabeth Ashmore, Trina Duke, and Jennifer Sakano

The suite of the *TOEIC*® tests assesses the English-language proficiency of nonnative speakers in the global workplace and everyday life. Test takers are (a) working in international companies where English is necessary to communicate with native and other nonnative speakers, (b) preparing to enter the global workforce, or (c) wanting to improve their general English-language proficiency for daily life. The *TOEIC*® Listening and Reading test has been in use since 1979; as the need for direct measures of speaking and writing skills emerged, the *TOEIC*® Writing and Speaking tests were added in 2006. TOEIC test scores are used to inform decisions affecting recruitment, job placement, promotion, training, and evaluation. Used in 150 countries around the world, the TOEIC tests are widely recognized as a global standard for general English-language proficiency.

The TOEIC program recognizes that the use of English communication continually evolves, particularly in international contexts. Thus, in the spring of 2013, it revisited the current TOEIC Listening and Reading test to examine what, if any, updates could and should be made. The project was a collaboration among the Assessment Development division, the Statistical Analysis, Data Analysis and Psychometric Research division, the Research division, and the Global TOEIC business management unit at Educational Testing Service.

Goals of *TOEIC*® Listening and Reading Test Update

The goals of the project can be summarized as follows:

- To ensure that test tasks are aligned with current theories of language learning and language testing
- To create positive washback for test takers and teachers—to encourage preparation for the TOEIC Listening and Reading test that aligns more consistently with sound pedagogical practice to increase English-language skills
- To modify tasks in order to reflect changes in communication styles and methods in the modern workplace and daily life
- To increase the amount of individualized feedback and the utility of the information provided to test takers and score users

Stages of Development for the Updated Test

The stages of development for the updated test were:

1. Proposing changes to the test content to achieve the goals stated above
2. Gathering feedback from clients and score users
3. Gathering data on the performance of possible new item types via a pilot study
4. Gathering data on the revised test as a whole via field studies
5. Finalizing the operational test design for the updated test

Revisiting the Test Design

Test development experts at ETS examined key areas such as (a) language tasks currently carried out in the global workplace and everyday life, (b) the needs of score users, and (c) any areas of linguistic competence that may have been underrepresented by the current testing of the construct. They also considered the possibility of making changes to the look and feel of the test.

An example of a language task that was found to have gained greater prominence in the workplace since the 2006 revision of the test is the understanding of forms of electronic communication such as text messages, online chat discussions, and electronic bulletin boards. Areas that were considered to expand the linguistic competencies tested in the construct included using slightly more informal language and testing connections between information that is heard and information that is read.

Client and Score User Feedback

After identifying aspects of the construct that could be represented in the new test, the team wrote examples of question types that would provide evidence to support claims about test-taker abilities. Sample questions were shared with clients and feedback was collected in writing and in face-to-face discussion. While some practical considerations were raised, feedback from clients was largely positive.

Pilot Study

A pilot study was conducted to gather empirical information on the performance of the possible new question types. The test forms administered in the study were not necessarily intended to reflect the eventual composition of an updated test.

Two parallel forms were constructed using a mixture of previously administered questions, new question types, and new questions of existing types. The forms were administered in Japan and Korea in late 2013 and early 2014 to 2486 test takers. The population was split almost equally between the two test forms and between the two countries. Data collected were sufficient to evaluate the performance of the questions.

The new questions tested in the pilot study represented the tasks proposed in the preceding activities.

Pilot Study Forms

Listening—Candidates for new tasks in the listening sections included:

- Conversations and talks that refer to written information (typically in the form of a short schedule or graph) that require the test taker to understand what the speaker(s) is (are) saying about the written information

- Conversations among three speakers
- Sets composed of a conversation and a related talk, or two related conversations
- Questions testing pragmatic understanding that included the replay of a word or phrase from the stimulus and a question about the speaker's intended meaning.

Reading—Candidates for new tasks in the reading sections included:

- Text completion sets including a question testing discourse organization. These questions required the selection of a complete sentence that best fit in a provided blank.
- Reading comprehension sets based on text messages or online chat chains. The passages included entries by more than one writer and informal language. The sets included a question testing pragmatic understanding by asking the intended meaning of a word or phrase in context.
- Reading comprehension sets based on *three* passages on a related topic. These sets included two questions that required a connection between two of the passages.
- Reading comprehension sets testing the understanding of numerical data (such as survey results) presented in prose. The questions required drawing conclusions based on the material and interpreting the relationship between the prose and a table used to present the data.
- Reading comprehension sets with a question testing discourse organization by having the test taker choose the best position in the passage in which to place a provided sentence.

Pilot Study Results

All of the new question types tried out in the pilot were successful from a content and test design point of view. They performed within the normal range for TOEIC Listening and Reading questions with respect to both difficulty and discrimination.

In the *Listening* section, questions connecting an oral text with a graphic, questions testing pragmatic understanding, and questions about conversations with three speakers all performed as expected, and within the expected range of difficulty. When the TOEIC test update project began, the business partners wanted to leave open the possibility of increasing the overall difficulty of the test. Test development staff created tasks for the pilot which were likely to be more difficult. The listening sets that used a talk and related conversation as stimuli were created for this purpose and, in fact, were consistently difficult. When the decision was made that the updates should *not* increase overall test difficulty, this question type was excluded from subsequent development efforts.

In the *Reading* section, the two varieties of discourse analysis questions (those in the text completion sets and those in the reading comprehension sets) were slightly easier than predicted. The questions about the text message and online chat chains showed statistically similar performance to that of questions about existing passage types. The questions about triple passages were more difficult than the double passages; in subsequent content development, the reading load across three related passages was reduced to be comparable to the reading load in two related passages. The questions about numerical information (the survey sets) also performed well, but the question type was not used in further testing because of concerns that there was a limited range of language to test and thus the question type might become quite coachable.

Questionnaire

A questionnaire was administered to collect test taker feedback about the pilot forms and test questions. Test takers were asked about their background and their overall impression of test difficulty, as well as some specific questions about new question types. Test takers in both Japan and Korea indicated that they found the *Listening* section more difficult than the *Reading* section, but to different degrees. The survey responses indicated that the directions to the test were clear and that test takers understood what they were required to do for the new question types.

Field Study

Decisions made subsequent to the pilot affected the design of the field study forms. Feedback from clients indicated that the updated TOEIC Listening and Reading test would be most useful to score users if the following criteria were met:

- The test continued to be administered in a paper and pencil format and consisted entirely of selected response questions.
- The number of questions in each section (*Listening* and *Reading*) remained the same as in the existing test.
- Test administration time remained the same as in the existing test.
- The reported section and total score scales remained the same and overall test difficulty did not change.

Changes to test content affected no more than 20–30% of the questions in each section. All existing question types are represented in the updated test, as the information generated from them is still relevant and useful to test takers and score users.

The field study was the final stage of the project. The study was conducted in two rounds in late 2014 and early 2015. In each round, parallel test forms were administered in Japan and Korea. Each field study form combined previously administered questions, new questions of existing types, and examples of new question types created as part of the update project.

This study was intended to assess the adjustments to the test as a whole. The focus was not only on the performance of individual questions and the new question types but on the specifications and performance of complete test forms. In response to the study, small refinements were made to the distribution of the question types and to question difficulty.

New question types again performed within the existing TOEIC score ranges for difficulty and discrimination. New and newly coded questions supported additional information that could be included on the score report.

Post Field Study

Following the field study, some further updates were made to the specifications for the test. In the Conversation and Talk parts of the *Listening* section, the text of the spoken phrase as well as the audio playback of the word or phrase tested was included in questions that assess understanding of intended meaning in context (pragmatic understanding). This is to help the test taker focus on the correct portion of the conversation or talk, while still allowing the question type to provide information about test taker ability.

In each section of the test, measures were instituted to control the amount of listening or reading required. Guidelines were put in place to limit the length of conversations and talks, and to limit the reading load in all parts of the *Reading* section. The aim of these guidelines is to ensure that new forms are comparable with existing forms, and to aid in the ongoing assembly of parallel forms.

Updates to Score Report

Working with the Research division, assessment specialists and psychometricians investigated possibilities for adding to feedback provided to test takers and score users on the score report. A category of pragmatic understanding was added to the "Abilities Measured" for the *Listening* section. Some existing Question-Response questions already test pragmatic understanding; those questions, with the new pragmatic understanding questions in the Conversations and Talk test parts, provide information that supports separately reporting the ability to understand a speaker's purpose or intended meaning. For a discussion of the existing "Abilities Measured" (also referred to as *claims*), please see Schedl (2010).

Conclusion

The TOEIC Listening and Reading Test update has resulted in test forms that resemble the existing test but that more closely reflect current real-world language and tasks. The expansion of the kinds of real-world tasks represented on the test increases the likelihood of positive test washback; that is, test takers preparing to take the test are likely, in the process, to develop skills that are useful in the real world. A single *Ability Measured* has been added to the score report. Careful psychometric analysis supports all changes to the test, including multi-coding of some questions to allow the addition to the score report. Test takers will find a test that has not changed in length, number of questions, or the scale on which scores are reported.

Table 1

Comparison of the Composition of the Existing TOEIC Listening and Reading Test With Final TOEIC Listening and Reading Test Design: Listening Section

		Existing version	Updated version
Part 1:	Stand-alone questions	Photographs: 10 questions	Photographs: 6 questions
Part 2:	Stand-alone questions	Question-response: 30 questions	Question-response: 25 questions
Part 3:	Set-based questions	Conversations: 30 questions • 10 conversations • 3 questions per conversation	Conversations: 39 questions • 13 conversations • 3 questions per conversation
Part 4:	Set-based questions	Talks: 30 questions • 10 talks • 3 questions per talk	Talks: 30 questions • 10 talks • 3 questions per talk
		Listening Comprehension: 100 questions	Listening Comprehension: 100 questions

Table 2

Comparison of the Composition of the Existing TOEIC Listening and Reading Test With Final TOEIC Listening and Reading Test Design: Reading Section

		Existing version	Updated version
Part 5:	Stand-alone questions	Incomplete sentences: 40 questions	Incomplete sentences: 30 questions
Part 6:	Set-based questions	Text completion: 12 questions	Text completion: 16 questions
Part 7:	Set-based questions	Single passage: 28 questions • 9 single passages • 2–5 questions per passage	Single passage: 29 questions • 10 single passages • 2–4 questions per passage
	Set-based questions	Double passages: 20 questions • 4 double passages • 5 questions per set	Multiple passages: 25 questions • 2 set-based double passages • 3 set-based triple passages • 5 questions per set
		Reading Comprehension: 100 questions	Reading Comprehension: 100 questions

References

Schedl, M. (2010). Background and goals of the *TOEIC®* Listening and Reading Test Redesign Project. In *The research foundation for the TOEIC® tests: A compendium of studies* (pp. 7.1–7.31). Princeton, NJ: Educational Testing Service.

Wei, Y., & Cid, J. (2014). *TOEIC® Listening and Reading pilot test analysis report.* Unpublished manuscript. Educational Testing Service.

Appendix. Summary of *TOEIC*® Listening and Reading Update Refinements

Overall Comments

The range of proficiency tested has not changed, and the scale for each test section remains the same as for the existing test. There are still 100 questions in each test section; time for delivery of the test remains unchanged. Some questions are multi-coded and contribute to more than one ability measured (claim).

Listening Section

The Listening test continues to be paced, with a time of 45 minutes +/- 59 seconds. There are still four accents in the Listening section. The language in the Listening section is more reflective of speech used in the real world—the kind of language that test takers will hear in their everyday and working lives. Therefore, there may be fewer complete sentences, more fragments, more reduced speech, some interruptions or hesitations, meaning communicated through tone and/or stressed language, and some variation in pace.

1. Photographs

   Photograph questions offer visual variation and are a pure test of listening, since they require no reading. Six photographs remain in the test. Photographs support the claims that the test taker can understand (a) gist and (b) detail in short spoken texts.

2. Question Response

   Question Response (QR) questions are also a pure test of listening and require no reading. Twenty-five QR questions remain in the test to allow support of existing claims—understanding (a) gist and (b) detail in short spoken texts—as well as the new pragmatic understanding claim about test-taker performance. Some QR questions are multicoded to support more than one claim.

3. Conversations

   Conversation sets allow the testing of a variety of subskills. The conversations sound more natural with the inclusion of elision, sentence fragments, interruptions, and meaning indicated through intonation and stress. Some sets allow the testing of pragmatic understanding; some, the connection with a visual image.

   (Some conversations refer to a visual image such as a sign or map, so that test takers are carrying out real-world tasks that require them to make a connection between an oral text and a visual image. The reading load is not different from what is currently required of the test taker when reading question stems and options in the test book.) The photographs and QR questions that have been removed from earlier in the test are replaced with more conversation questions that allow the inclusion of these tasks; there is now a total of 39 conversations.

4. Talks

   As in the present TOEIC test, there are 30 talk questions. Some sets allow the testing of pragmatic understanding; some, the connection with a visual image.

Reading Section

The Reading test will continue to take 75 minutes of testing time. Question types added to the Reading section allow for the testing of pragmatic understanding (understanding intended meaning), discourse organization, and connections across multiple texts.

5. Incomplete Sentences

   The number of these questions has been reduced to accommodate new question types that reflect more authentic tasks. Enough Incomplete Sentence questions remain to support current claims.

6. Text Completion

   There are still 4 text completion sets, with an additional question added to each set. The added question tests the ability to recognize which of four complete sentences belongs in a particular part of a text. There are 16 text completion questions in all. In addition, in a response to comments from some test takers, the question options now appear below the text, without interrupting the flow of the reading.

7. Reading Comprehension

   There is a one-question increase in the number of single-passage questions. Three triple-passage texts replace two of the double-passage sets and one single passage text. In other words, there are single-text and multiple-text (2- or 3- related-text) reading comprehension sets. Five questions are associated with each of the 5 multiple-passage texts. The overall number of questions in the reading comprehension section has increased slightly to allow the testing of pragmatic understanding and discourse organization. Two reading sets will reflect more recently adopted real-world communication styles, such as instant messaging, where there may be more than one writer and sequencing is not always linear.

Compendium Study

Statistical Analyses for the Updated *TOEIC*® Listening and Reading Test

Jaime Cid, Youhua Wei, Sooyeon Kim, and Claudia Hauck

The *TOEIC*® program is designed to measure the English-language proficiency of nonnative speakers of English engaged in the global workplace, where English is the language of communication. The *TOEIC*® Listening and Reading test consists of two separately timed sections, *Listening Comprehension* and *Reading Comprehension* with 100 items in each section. The *Listening* section is paced by audiotape recording.

In May 2016, the TOEIC program announced updates to the TOEIC Listening and Reading test to keep up with the changing use of English and the ways in which individuals commonly communicate in the global workplace and everyday life. New item types were included, but there was no change in total testing time, number of items, test difficulty, or score scale. The updated test included communication formats, such as text messaging and instant messaging, which are in current use. It also placed greater emphasis on connecting information across multiple sources, such as what is seen in a visual image and what is heard in a related conversation (pragmatics). A pilot study conducted in May 2015 evaluated the statistical properties of the updated TOEIC Listening and Reading test. The purpose of this report is to document the results of such statistical analyses.

Background

Table 1 presents the composition of the *Listening* and *Reading* sections, in the preupdated and updated (new) specifications. The changes in the *Listening* section require a greater emphasis on *Short Conversations* but less emphasis on *Photographs and Question–Response*. The changes on the *Reading* section require a greater emphasis on *Reading Comprehension* and *Text Completion* but less emphasis on *Incomplete Sentences*. Approximately one-quarter of items in each of the *Listening* and *Reading* sections are new-type items. These item types include to some extent the new features aforementioned (e.g., new communication formats in the *Reading* section, such as text messages, instant messages, and online chat conversations with multiple writers). The preupdated score reports included scale scores for both the *Listening* and *Reading* sections and the percentage of questions answered correctly for each of four ability claims in the *Listening* section and each of five ability claims in the *Reading* section. The updated *Listening* section includes an additional ability claim (Ability 5, pragmatic understanding). The reporting scale for each section of the updated test remains the same as for the preupdated test, with a score scale ranging from 5 to 495 in increments of 5.

Table 1

Composition of Each Section Under the Updated Specification and the Preupdated Specification of the TOEIC Test

Section: Part	Updated test	Preupdated test
Listening: Part 1. Photographs	6	10
Listening: Part 2. Question–Response	25	30
Listening: Part 3. Short Conversations	39	30
Listening: Part 4. Short Talks	30	30
Reading: Part 5. Incomplete Sentences	30	40
Reading: Part 6. Text Completion	16	12
Reading: Part 7. Reading Comprehension	54	48
Reading: Part 7A. Single Passages	29	28
Reading: Part 7B. Multiple Passages	25	20

Note. Total number of items in each section was 100.

Pilot Form Design

Two parallel TOEIC Listening and Reading test pilot forms (Forms E and F) were assembled based on the updated specifications (see Table 1). Forms E and F were designed to be parallel from statistical and content perspectives. The pilot forms were randomly distributed to the test takers to make the two pilot form groups comparable in ability. To establish a strong score connection between the reference and the pilot forms, 50 *Listening* items and 45 *Reading* items were used as common items in both Form E and Form F. The common item sets were designed to be miniature versions of the reference form in terms of the content and statistical specifications.

As mentioned earlier, for both sections of the updated test, five ability claims are reported in the score report using the percentage correct score. Tables 2 and 3 present the number of items associated with each of the five abilities measured in each section. Although some of the abilities had fewer than 15 items in the pilot forms, the minimum number of items included currently in operational forms for each ability claim is 15.

Table 2

Number of Items for Each Ability Claim in the Listening Section

Ability	Form E	Form F	Reference
1. Can infer gist, purpose, and basic context based on information that is explicitly stated in short spoken texts	16	15	19
2. Can infer gist, purpose, and basic context based on information that is explicitly stated in extended spoken texts	19	16	17
3. Can understand details in short spoken texts	15	16	21
4. Can understand details in extended spoken texts	50	53	43
5. Can understand a speaker's purpose or implied meaning in a phrase or sentence (pragmatic understanding)	11	13	-

Note. Because some items measure more than one ability in the *Listening* section, the total number of items in each form will not be equal to 100.

Table 3

Number of Items for Each Ability Claim in the Reading Section

Ability	Form E	Form F	Reference
1. Can locate and understand specific information in tables and passages	18	20	16
2. Can connect information across multiple sentences in a single text and across texts	13	11	16
3. Can make inferences based on information in written texts	35	35	25
4. Can understand vocabulary in workplace texts	28	26	29
5. Can understand grammar in workplace texts	20	20	27

Note. Because some items measure more than one ability in the *Reading* section, the total number of items in each form will not be equal to 100.

Data Collection

A total of 3,673 test takers from Japan (*n* = 2,045) and Korea (*n* = 1,628) participated in the pilot study in May 2015. To evaluate the representativeness of the pilot samples, standardized mean differences[1] (SMD) were calculated based on the total score of each group. As shown in Table 4, in the reference form group, who were administered the May 2014 operational form, Korean test takers were much more able than Japanese test takers in both sections, and their ability difference was much larger in *Listening* (SMD = .53) than in *Reading* (SMD = .22). In the pilot study, a different trend emerged. The Japanese pilot form groups were more able than the Korean groups, and their ability differences were larger on the *Reading* section than on the *Listening* section. Therefore the pilot samples were not completely representative of the TOEIC population. A possible reason is that in the pilot samples, the percentage of repeaters was larger than the percentage observed in operational practice in Japan than in Korea. However, the operational trend of the Korean group performing comparatively better on the *Listening* section than on the *Reading* section was present in the pilot study (.35 and .21 better for Form E and Form F, respectively). The descriptive statistics of raw scores for *Listening* and *Reading* sections by country and form are presented in Table 5.

Table 4

Standardized Mean Differences of Groups and of Forms Based on the Total Test Score of Each Group

Difference	Listening	Reading
Form E (Korea—Japan)	–.05	–.40
Form F (Korea—Japan)	–.12	–.33
Reference (Korea—Japan)	.53	.22

Table 5

Descriptive Statistics of Raw Scores by Country and Form

Statistic	Form E: Japan	Form E: Korea	Form E: Combined	Form F: Japan	Form F: Korea	Form F: Combined	Reference: Japan	Reference: Korea	Reference: Combined
Sample size	1,019	824	1,843	1,026	804	1,830	48,745	38,500	87,245
Listening mean	65.11	63.63	64.45	66.34	63.90	65.27	66.72	73.81	69.85
Listening *SD*	15.35	18.14	16.67	15.74	18.51	17.05	15.95	16.31	16.49
Reading mean	55.96	50.99	53.74	60.31	54.97	57.96	57.05	62.42	59.42
Reading *SD*	15.1	16.98	16.16	15.87	19.02	17.52	16.95	18.14	17.69

Note. SD = standard deviation.

Statistical Analyses and Results

Equating

The comparability of the pilot and operational testing samples was further evaluated by examining the performance of the pilot and reference groups on the common items. Then, for each pilot form, equating—under the nonequivalent groups with anchor test design—was conducted through the common items shared in both the pilot (updated) forms and operational reference (preupdated) form. Equating is used to adjust the difficulty level of a form and derive the scaled scores from test takers' raw scores in order that the reported scaled scores obtained from different test forms are comparable, regardless of any potential differences in form difficulty. The number of common items was 50 in *Listening* and 45 in *Reading*. The equating relationship between the new forms and the operational reference form was based on the combined group of Japanese and Korean test takers. Table 6 presents the descriptive statistics of the anchor scores in Forms E and F (combined group) and the operational reference form. As indicated by the negative SMD between the new and operational reference groups in Table 6, the operational reference group was somewhat more able than the combined pilot groups in both sections. Likewise, the Form F group was somewhat more able than the Form E group.

Table 6

Summary of Anchor Statistics and Group Differences

Statistic	Form E	Form F	Reference form
Sample size	1,843	1,830	87,245
Listening number of anchor items	50	50	50
Listening mean	34.76	35.42	35.54
Listening *SD*	8.60	8.59	8.25
Listening standardized difference [a]	−0.09	−0.02	
Reading number of anchor items	45	45	45
Reading mean	25.71	26.64	26.97
Reading *SD*	8.01	8.19	7.82
Reading standardized difference [a]	−0.16	−0.04	

Note. SD = standard deviation.

[a] Denotes standardized mean difference between the pilot form (E or F) and reference form.

Table 7 provides the summary statistics (mean and standard deviation) of the scaled scores for each group taking each form. Recall that the Japanese pilot groups were more able than the Korean pilot groups. As expected, after adjusting the test form difficulty, the scaled score means of the Japanese pilot form groups were higher than the scaled score means of the Korean pilot form groups. Likewise, the scaled score mean of the combined pilot group was somewhat lower (Japan and Korea) than for the reference group. Therefore the group differences based on reported scores were consistent with the group differences based on anchor raw scores.

Table 7

Summary Statistics of Test Takers' Scale Scores

Statistic	Form E: Japan	Form E: Korea	Form E: Combined	Form F: Japan	Form F: Korea	Form F: Combined	Reference: Japan	Reference: Korea	Reference: Combined
Sample size	1,019	824	1,843	1,026	804	1,830	48,745	38,500	87,245
Listening mean	329.28	320.23	325.23	338.96	324.27	332.50	316.1	354.40	333.00
Listening *SD*	84.13	100.58	91.86	83.98	100.98	92.01	85.21	88.73	88.84
Reading mean	277.09	246.64	263.50	288.19	257.73	274.87	264.86	294.83	278.09
Reading *SD*	93.38	103.70	99.24	93.39	109.01	101.6	94.52	100.48	98.33

Note. SD = standard deviation.

Item Difficulty

The difficulty of the items was evaluated by examining two types of statistical indices: *p*-value (defined as the proportion of test takers who answer an item correctly in a given population) and delta (defined as 13 + 4*z*, where *z* is the normal deviate corresponding to proportion correct). *P*-values range from 0 to 1, with a higher value indicating that a greater proportion of test takers responded to the item correctly, and it was thus an easier item. Delta values typically range from 6 for a very easy item to 20 for a very difficult item, with a mean of 13 (50% correct).

Table 8 presents the *p*-values and equated deltas[2] in each section of the pilot forms and operational reference form. In *Listening*, the mean *p*-value for the operational reference form was .70, and the mean *p*-values for Forms E and F were .64 and .65, respectively. In *Reading*, the mean *p*-value for the operational reference form was .60, and the mean *p*-values for Forms E and F were .55 and .59, respectively.

The equated deltas provide us with a difficulty metric that accounts for the different ability levels among the two pilot test groups and the operational test group. The *Listening* sections for the pilot forms were slightly more difficult than the operational reference form. In *Reading*, the overall difficulty of the pilot forms was more comparable to the overall difficulty of the operational reference form. This finding is not unexpected given that test takers were not as familiar with the new item types in the pilot forms as they were with the items of the operational reference form.

Table 8

Summary of Item Statistics for Each Section Based on Combined Group

Statistic	*p*-value: Form E	*p*-value: Form F	*p*-value: Reference	ED: Form E	ED: Form F	ED: Reference	*R*-biserial: Form E	*R*-biserial: Form F	*R*-biserial: Reference
Listening mean	0.64	0.65	0.70	13.1	13.2	12.7	0.48	0.50	0.47
Listening *SD*	0.16	0.15	0.13	1.6	1.5	1.3	0.11	0.10	0.11
Listening min	0.26	0.24	0.40	9.8	9.5	9.3	0.20	0.27	0.19
Listening max	0.92	0.94	0.95	16.7	17.0	15.2	0.70	0.74	0.67
Reading mean	0.55	0.59	0.60	12.5	12.3	12.3	0.45	0.49	0.47
Reading *SD*	0.18	0.18	0.16	1.9	1.8	1.7	0.14	0.13	0.11
Reading min	0.19	0.20	0.22	8.3	8.7	8.0	0.09	0.10	0.15
Reading max	0.89	0.89	0.92	16.4	16.4	16.1	0.73	0.72	0.70

Note. ED = equated delta; SD = standard deviation.

Table 9 shows *p*-values and equated delta values for the different parts of the test on the pilot forms and the operational reference form. Overall, in comparison to the operational reference form, in *Listening*, *Short Conversations* (Part 3) and *Short Talks* (Part 4) were more difficult on the pilot forms than on the operational reference form. The same was observed in *Reading for Multiple Passages* (Part 7B). However, in general, all forms produced similar difficulty patterns. That is, in *Listening*, *Photographs* (Part 1) and *Short Talks* (Part 4) were, on average, the easiest and hardest parts, respectively. In *Reading*, as observed on the operational reference form, *Incomplete Sentences* (Part 5) and *Multiple Passages* (Part 7B) were, on average, the easiest and most difficult parts, respectively.

Table 9

Means of Item Statistics for Each Part Based on Combined Group

Section: Part	*p*-value: Form E	*p*-value: Form F	*p*-value: Reference	ED: Form E	ED: Form F	ED: Reference	*R*-biserial: Form E	*R*-biserial: Form F	*R*-biserial: Reference
Listening: Part 1	0.80	0.82	0.74	11.4	11.4	11.9	0.39	0.39	0.40
Listening: Part 2	0.67	0.70	0.68	12.9	12.7	12.8	0.45	0.47	0.44
Listening: Part 3	0.66	0.62	0.73	13.0	13.5	12.6	0.52	0.50	0.50
Listening: Part 4	0.57	0.62	0.66	13.8	13.6	13.1	0.47	0.53	0.50
Reading Part 5	0.67	0.68	0.65	11.2	11.4	11.7	0.52	0.51	0.50
Reading Part 6	0.55	0.57	0.51	12.5	12.5	13.0	0.42	0.46	0.40
Reading Part 7	0.48	0.55	0.57	13.2	12.8	12.7	0.42	0.48	0.47
Reading Part 7A	0.53	0.62	0.61	12.7	11.9	12.2	0.45	0.54	0.48
Reading Part 7B	0.42	0.45	0.51	13.8	13.7	13.4	0.39	0.42	0.45

Note. ED = equated delta; Part 1 = Photographs; Part 2 = Question–Response; Part 3 = Short Conversations; Part 4 = Short Talks; Part 5 = Incomplete Sentences; Part 6 = Text Completion; Part 7 = Reading Comprehension; Part 7A = Single Passages; Part 7B = Multiple Passages.

Item Discrimination

Item discrimination is evaluated by the *R*-biserial correlation coefficient. The *R*-biserial correlation is the relationship between test takers' scores on a particular item (e.g., 0 for an incorrect response or 1 for a correct response) with the corresponding total score (e.g., total score for a section). The *R*-biserial correlation indicates how well an item serves to discriminate between low- and high-ability test takers. Table 8 presents the summary statistics for the *R*-biserial correlations for the pilot and operational reference forms. In general, for both *Listening* and *Reading*, the means of *R*-biserial values were comparable between the pilot forms and the operational reference form. Overall, these results indicate that the three forms were, on average, equally discriminating.

Table 9 provides *R*-biserial values for the different parts of the test in Forms E and F and the reference form. Overall, the values suggest that for both *Listening* and *Reading*, on average, the items of the different parts of Forms E and F were very close in discrimination to the items of the operational reference form.

Differential Item Functioning

Differential item functioning (DIF) analyses were performed to ensure that all new item types were fair to both men and women. DIF analyses involve the statistical analysis of test items for evidence of differential item difficulty related to subgroup membership. The two groups of interest (e.g., male/ female) are matched with respect to ability on a criterion (e.g., total test score) and then compared to see if the item is performing similarly in both groups. The probability that a test taker answers an item correctly should be independent of his or her group membership. The DIF analysis methodology employed (Dorans & Kulick, 1986; Holland & Thayer, 1988) uses statistics that describe the amount of DIF for each item as well as the statistical significance of the DIF effect. The DIF classification followed the ETS system as described by Zwick (2012), in which items are classified into three levels: A (least), B, and C (most). Items identified as C-level DIF should be referred to fairness committees for further evaluation. No item showed C-level DIF. Therefore no item was differentially more difficult for one gender than the other.

Test Parts and Abilities

As mentioned earlier, the *Listening* section of the updated test includes four parts and provides five ability scores, whereas the *Reading* section includes three parts and provides five ability scores. The fifth ability of the *Listening* section of the updated test is a new ability claim. The correlation between each item score and its ability score measures how well each item is related to its corresponding ability claim. As shown in Table 10, the average item–ability correlations were generally moderate in the *Listening* and *Reading* sections. Forms E and F and operational reference form showed similar patterns. The newly added *Listening* ability claim (Ability 5, pragmatic understanding) yielded item correlations comparable to the ones observed in the other Listening claims.

Table 10

Summary of Item–Ability Correlations Based on Combined Group

Form: Ability	Listening mean	Listening *SD*	Listening min	Listening max	Reading mean	Reading *SD*	Reading min	Reading max
Form E: Ability 1	0.53	0.08	0.35	0.65	0.49	0.12	0.28	0.68
Form E: Ability 2	0.55	0.07	0.31	0.67	0.58	0.08	0.45	0.68
Form E: Ability 3	0.54	0.06	0.42	0.62	0.46	0.14	0.17	0.71
Form E: Ability 4	0.52	0.12	0.23	0.71	0.51	0.13	0.27	0.72
Form E: Ability 5	0.56	0.08	0.44	0.71	0.54	0.09	0.28	0.71
Form F: Ability 1	0.55	0.09	0.35	0.66	0.55	0.13	0.32	0.75
Form F: Ability 2	0.58	0.07	0.50	0.74	0.64	0.07	0.49	0.72
Form F: Ability 3	0.52	0.07	0.40	0.64	0.50	0.14	0.18	0.74
Form F: Ability 4	0.53	0.10	0.34	0.77	0.52	0.11	0.28	0.69
Form F: Ability 5	0.56	0.07	0.42	0.64	0.56	0.10	0.29	0.73
Reference: Ability 1	0.54	0.10	0.25	0.70	0.58	0.10	0.36	0.68
Reference: Ability 2	0.54	0.06	0.40	0.64	0.58	0.09	0.40	0.72
Reference: Ability 3	0.50	0.07	0.38	0.61	0.53	0.12	0.28	0.69
Reference: Ability 4	0.54	0.09	0.25	0.69	0.50	0.13	0.23	0.69
Reference: Ability 5	–	–	–	–	0.53	0.10	0.26	0.73

Note. Ability 5 is the new Listening ability added to the updated TOEIC test. For Listening: Ability 1, can infer gist, purpose, and basic context based on information that is explicitly stated in short spoken texts; Ability 2, can infer gist, purpose, and basic context based on information that is explicitly stated in extended spoken texts; Ability 3, can understand details in short spoken texts; Ability 4, can understand details in extended spoken texts; Ability 5, can understand a speaker's purpose or implied meaning in a phrase or sentence (pragmatic understanding). For Reading: Ability 1, can locate and understand specific information in tables and passages; Ability 2, can connect information across multiple sentences in a single text and across texts; Ability 3, can make inferences based on information in written texts; Ability 4, can understand vocabulary in workplace texts; Ability 5, can understand grammar in workplace texts.

Tables 11–14 present the intercorrelations of the different parts of the test and the abilities in Forms E and F. The lower part below the diagonal presents the correlations from the Form E group, and the upper part above the diagonal presents the correlations of the Form F group. As expected, in *Listening* (Tables 11–12), Photographs (Part 1), with only six items, yielded the lowest correlations. The correlations among parts (Parts 1–4 for *Listening* and Parts 5–7B for *Reading*) and abilities (Abilities 1–5 for each section) were moderate to high. The newly added *Listening* ability (Ability 5, pragmatic understanding) yielded correlations comparable to those of the other abilities. Although not reported in the tables owing to space constraints, the intercorrelations of parts and abilities of the operational reference form in the *Listening* and *Reading* sections are consistent with the trends observed for the pilot forms.

Table 11

Intercorrelations of Parts Based on Combined Group for Listening

Part	Total	Part 1	Part 2	Part 3	Part 4
Total	–	.53	.87	.94	.93
Part 1	.54	–	.47	.44	.44
Part 2	.86	.47	–	.73	.72
Part 3	.95	.45	.73	–	.82
Part 4	.91	.42	.69	.80	–

Note. Part 1 = Photographs; Part 2 = Question–Response; Part 3 = Short Conversations; Part 4 = Short Talks.

Table 12

Intercorrelations of Abilities Based on Combined Group for Listening

Ability	Total	Ability 1	Ability 2	Ability 3	Ability 4	Ability 5
Total	–	.80	.88	.80	.97	.83
Ability 1	.80	–	.65	.67	.69	.82
Ability 2	.88	.60	–	.64	.83	.70
Ability 3	.79	.63	.64	–	.69	.69
Ability 4	.97	.69	.82	.67	–	.77
Ability 5	.79	.73	.63	.63	.74	–

Note. Ability 1, can infer gist, purpose, and basic context based on information that is explicitly stated in short spoken texts; Ability 2, can infer gist, purpose, and basic context based on information that is explicitly stated in extended spoken texts; Ability 3, can understand details in short spoken texts; Ability 4, can understand details in extended spoken texts; Ability 5, can understand a speaker's purpose or implied meaning in a phrase or sentence (pragmatic understanding).

Table 13

Intercorrelations of Parts Based on Combined Group for Reading

Part	Total	Part 5	Part 6	Part 7	Part 7A	Part 7B
Total	–	.88	.81	.95	.91	.79
Part 5	.87	–	.68	.71	.71	.56
Part 6	.79	.69	–	.68	.70	.50
Part 7	.93	.66	.61	–	.92	.88
Part 7A	.88	.67	.64	.90	–	.63
Part 7B	.76	.49	.43	.87	.58	–

Note. Part 5 = Incomplete Sentences; Part 6 = Text Completion; Part 7 = Reading Comprehension; Part 7A = Single Passages; Part 7B = Multiple Passages.

Table 14

Intercorrelations of Abilities Based on Combined Group for Reading

Ability	Total	Ability 1	Ability 2	Ability 3w	Ability 4	Ability 5
Total	–	.89	.85	.93	.91	.86
Ability 1	.83	–	.76	.86	.73	.67
Ability 2	.79	.64	–	.78	.68	.64
Ability 3	.91	.81	.72	–	.79	.69
Ability 4	.90	.66	.60	.77	–	.79
Ability 5	.85	.60	.53	.66	.78	–

Note. Ability 1, can locate and understand specific information in tables and passages; Ability 2, can connect information across multiple sentences in a single text and across texts; Ability 3, can make inferences based on information in written texts; Ability 4, can understand vocabulary in workplace texts; Ability 5, can understand grammar in workplace texts.

Reliability

Reliability provides an indication of the extent to which test scores are consistent across different conditions of administration of the same form or alternate forms. In general, when all else is equal, more items tend to lead to higher reliability. The reliability of the TOEIC Listening and Reading test is estimated using an internal consistency method (reliability coefficient called alpha) based on the correlations between different items on the same test. The reliability estimate ranges from 0 to 1. The higher the reliability coefficient for a section, part, or test, the higher the consistency of test takers' responses to the items of that section, part, or test.

Tables 15–16 display the reliability estimates for the total test and for different parts of the test and abilities for the pilot forms and the reference form in *Listening* and *Reading*. Overall, the reliabilities of the total test were nearly the same for the pilot forms and the operational reference form (.94 on average for *Listening* and *Reading*). Photographs (Part 1) produced the lowest reliability in the pilot forms. The reliability coefficients of the other parts of the test in both *Listening* and *Reading* were aligned with the reliabilities observed in the reference form and in typical operational forms.

The reliabilities of the ability scores were moderate to high and also comparable between the pilot forms and reference form. The newly added *Listening* ability (Ability 5, pragmatic understanding) yielded the lowest reliabilities because the number of items included in this ability was lower than the minimum number used in operational practice (i.e., 15) in Forms E and F (see Table 2).

Table 15

Reliability Estimates for Listening

Part or ability	Form E	Form F	Reference form
Total test	.94 (100)	.94 (100)	.94 (100)
Part 1. Photographs	.43 (6)	.37 (6)	.50 (10)
Part 2. Question–Response	.78 (25)	.79 (25)	.82 (30)
Part 3. Short Conversations	.88 (39)	.87 (39)	.84 (30)
Part 4. Short Talks	.82 (30)	.86 (30)	.85 (30)
Ability 1	.68 (16)	.68 (15)	.74 (19)
Ability 2	.76 (19)	.73 (16)	.70 (17)
Ability 3	.69 (15)	.68 (16)	.72 (21)
Ability 4	.89 (50)	.90 (53)	.89 (43)
Ability 5	.59 (11)	.67 (13)	–

Note. Numbers in parentheses indicate number of items. Ability 1, can infer gist, purpose, and basic context based on information that is explicitly stated in short spoken texts; Ability 2, can infer gist, purpose, and basic context based on information that is explicitly stated in extended spoken texts; Ability 3, can understand details in short spoken texts; Ability 4, can understand details in extended spoken texts; Ability 5, can understand a speaker's purpose or implied meaning in a phrase or sentence (pragmatic understanding).

Table 16

Reliability Estimates for Reading

Part or ability	Form E	Form F	Reference form
Total test	.93 (100)	.94 (100)	.94 (100)
Part 5. Incomplete Sentences	.86 (30)	.85 (30)	.88 (40)
Part 6. Text Completion	.68 (16)	.73 (16)	.60 (12)
Part 7. Reading Comprehension	.88 (54)	.91 (54)	.90 (48)
Part 7A. Single Passages	.81 (29)	.87 (29)	.83 (28)
Part 7B. Multiple Passages	.81 (25)	.82 (25)	.83 (20)
Ability 1	.64 (18)	.76 (20)	.74 (16)
Ability 2	.69 (13)	.73 (11)	.75 (16)
Ability 3	.80 (35)	.83 (35)	.80 (25)
Ability 4	.80 (28)	.80 (26)	.81 (29)
Ability 5	.77 (20)	.78 (20)	.81 (27)

Note. Numbers in parentheses indicate number of items. Ability 1, can locate and understand specific information in tables and passages; Ability 2, can connect information across multiple sentences in a single text and across texts; Ability 3, can make inferences based on information in written texts; Ability 4, can understand vocabulary in workplace texts; Ability 5, can understand grammar in workplace texts.

Speededness

The TOEIC *Listening* section is paced by a tape recording, and thus speededness is not a concern. Four types of statistics frequently used to evaluate the speededness of the *Reading* section are presented in Table 17: (a) percentage of test takers reaching all items, (b) percentage of test takers completing 75% of the items, (c) number of items reached by 80% of the test takers, and (d) ratio of not reached variance (NRV) to total variance (TV). Typically, a test is regarded as unspeeded for a group if (a) nearly all test takers complete 75% of the items, (b) at least 80% of the test takers reach all items, and (c) the ratio of NRV to TV is less than 0.15. As shown in Table 17, *Reading* was speeded for Forms E and F for Japan. The percentage reaching all items was 79% in both forms. Typically, this percentage in operational settings is about 95% for Japan. In this pilot study, the last five items had nonreached rates of about 20%. The values of the speededness index (i.e., ratio of NRV to TV) for Japan were much higher than a conventional criterion of .15. For the combined group, the *Reading* section was slightly speeded.

Table 17

Statistics of Speededness for Reading Sections

Statistic	Form E: Japan	Form E: Korea	Form E: Combined	Form F: Japan	Form F: Korea	Form F: Combined	Reference form: Japan	Reference form: Korea	Reference form: Combined
Number of test takers	1,019	824	1,843	1,026	804	1,830	48,745	38,500	87,245
% reaching all items	79.1	96.5	86.9	79.0	96.8	86.8	96.4	99.4	97.9
% reaching 75% of items	97.2	98.9	97.9	96.7	99.5	97.9	99.6	99.8	99.7
Number of items reached by 80%	97	100	100	97	100	100	100	100	100
Ratio of NRV to TV	0.27	0.10	0.18	0.25	0.06	0.15	0.04	0.02	0.02

Note. NRV = not reached variance; TV = total variance.

One of the chief purposes of the updated TOEIC Listening and Reading test was to ensure that the psychometric properties of the updated test were comparable to those of the preupdated test. The results presented in this report suggest that the updated pilot forms were equally discriminating on average on the total test and on different parts of the test as the operational reference form. The correlations among parts and ability scores were similar to the correlations observed in the operational reference form. Likewise, the newly added Listening ability (Ability 5, pragmatic understanding) produced correlations comparable to those of the other abilities. Overall, the reliabilities of *Listening* and *Reading*, parts, and ability scores in the updated pilot forms were similar to the reliabilities of the operational reference form. However, the updated Reading pilot forms appeared to be speeded for

Japan. Additionally, the results of the pilot study indicate that for both *Listening* and *Reading*, the items on the updated pilot forms were, on average, slightly more difficult than the items on the operational reference form. These findings were shared with test developers in order that they could make the appropriate adjustments to the difficulty of some items.

Operational Results

Since the launch of the updated TOEIC Listening and Reading test, the difficulty of the updated test and reliability of its scores have been closely monitored. To illustrate how the TOEIC test has continued to maintain the psychometric properties of the preupdated test, Table 18 provides a test performance comparison between preupdated and updated operational TOEIC Listening and Reading test forms based on Japan. The difference in average equated delta between preupdated and updated forms is .23 for both *Listening* and *Reading*. This difficulty difference is considered small.[3] In this regard, it is important to note that operational data have shown that, unlike for the Reading pilot forms, the percentage of reaching all items for Japan has been the same as the percentage observed for the preupdated forms (about 95%). Test discrimination and reliability have also not changed since the updates to the TOEIC test. The test continues to be equally discriminating (*R*-biserial ranges from .45 to .47 in *Listening* and *Reading*) and equally reliable (average reliability of .93). The average scaled scores are also quite stable. After forms are equated and the test scores are adjusted based on the difficulty levels of the forms, the average scaled scores for each section are relatively close.

Table 18

Summary Statistics of Preupdated and Updated TOEIC Listening and Reading Forms

Statistic	Preupdated form: Equated delta mean	Preupdated form: *R*-biserial mean	Preupdated form: Reliability	Preupdated form: Scale score mean	Updated form: Equated delta mean	Updated form: *R*-biserial mean	Updated form: Reliability	Updated form: Scale score mean
Listening mean	12.66	0.47	0.93	320.46	12.89	0.47	0.93	317.35
Listening *SD*	0.15	0.01	0.01	4.75	0.12	0.01	0.00	5.14
Listening min	12.30	0.44	0.92	312.89	12.60	0.45	0.92	306.50
Listening max	13.00	0.49	0.94	331.93	13.30	0.49	0.94	325.89
Reading mean	12.23	0.47	0.93	263.37	12.46	0.45	0.93	261.98
Reading *SD*	0.21	0.02	0.01	4.32	0.17	0.01	0.01	4.75
Reading min	11.90	0.44	0.92	253.35	12.00	0.42	0.91	252.37
Reading max	12.80	0.51	0.94	271.38	12.80	0.49	0.94	273.87

Note. N = 49. Preupdated forms are forms administered between November 2013 and April 2016. Updated forms are forms administered between May 2016 and May 2017. SD = standard deviation.

Similar trends were observed in a test performance comparison of 23 preupdated and 23 updated operational forms based in Korea. Average difficulty, discrimination, and reliability of the forms were also consistent between the preupdated and updated forms.

In summary, given the difficulty, discrimination, reliability, and scaled score values observed in operational practice, one can say that the updated TOEIC test continues to have the same psychometric quality as the preupdated TOEIC test.

Conclusion

Beginning with the public test in May 2016, the TOEIC Listening and Reading test included some updates to the question formats to reflect the changing use of English and the ways in which individuals commonly communicate in everyday social and workplace situations around the world. A pilot study conducted in May 2015 to evaluate the statistical properties of the updated TOEIC Listening and Reading test demonstrated that the psychometric properties of the updated pilot forms were comparable to those of the preupdated reference form. Overall, discrimination of items and sections; correlations among parts and ability scores; and reliabilities of sections, parts, and ability scores were similar to the ones observed in the operational reference form. The slight differences in difficulty levels observed in the pilot study were addressed by test developers, who made appropriate adjustments to the difficulty levels of some items. Operational data gathered after the launch of the updated test suggest that the TOEIC Listening and Reading test continues to have the same appropriate psychometric properties (e.g., difficulty, discrimination, reliability) as the preupdated test.

References

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing differential item functioning on the Scholastic Aptitude Test. *Journal of Educational Measurement*, *23*, 355–368. https://doi.org/10.1111/j.1745-3984.1986.tb00255.x

Holland, P. W., & Thayer, D. T. (1988). *An alternative definition of the ETS delta scale of item difficulty* (Research Report No. RR-85-43). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1985.tb00128.x

Zwick, R. (2012). *A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement* (Research Report No. RR-12-08). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2012.tb02290.x

Notes

[1] A summary statistic that expresses the mean difference between two groups in standard deviation units.

[2] Type of delta that indicates how difficult an item would be after placed on the same scale for all forms.

[3] A difference of .23 in equated delta converts approximately to a difference of .02 in *p*-value or proportion correct or to a difference of 2% in percentage correct.

Compendium Study

Statistical Analyses for the Expanded *TOEIC*® Speaking Test

Yanxuan Qu, Jaime Cid, and Eric Chan

Since the introduction in 2006 of the *TOEIC*® Speaking and Writing tests, the *TOEIC*® program has periodically evaluated the test content specifications to ensure that they continue to meet the needs of test takers and test users. To better foster communicative language learning and to discourage the use of memorization and test-taking strategies, Educational Testing Service (ETS) expanded the existing format of some items of the TOEIC Speaking test in May 2015. Specifically, additional formats were added to four of the existing speaking items (Items 4, 5, 6, and 10). Items 4, 5, and 6 are often called an item set because they share the same item stem. The appendix describes the existing formats and the new formats. The purpose of the expansion was not to replace existing formats but rather to supplement them with new alternative formats. More details about the process that ETS followed to develop the expanded item formats were provided in the paper "Expanding the Question Formats of the *TOEIC*® Speaking Test" by Park and Bredlau in the present compendium.

To ensure that these modifications would not significantly alter the difficulty of the items, a pilot study was conducted in November 2013. The purpose of the pilot study was to evaluate the comparability of existing formats with new formats in terms of difficulty and to determine if forms with the new formats had adequate reliability. In this paper, we summarize the analyses and results of the pilot study and the monitoring of the performance of the new formats in operational administrations.

The *TOEIC*® Speaking Test

The first TOEIC Speaking test was launched in December 2006. It was designed to measure test takers' ability to communicate in spoken English in the context of daily life and the global workplace. The test has 11 items. Items 1 and 2 are each scored on two dimensions: pronunciation and intonation. Each dimension has a score scale from 0 to 3. Items 3 to 9 are rated on a scale of 0 to 3. Items 10 and 11 are rated on a scale of 0 to 5. Raw scores on each item are weighted when calculating the total test score (Qu, Liu, & Chan, 2013). The reported scaled scores range from 0 to 200 in increments of 10.

Pilot Forms

Three test forms (A, B, and C) were used in the pilot study (Table 1). Form A was selected as the base form. This existing TOEIC Speaking-only form received relatively low exposure (i.e., only a small number of test takers have taken this form). Items 4, 5, 6, and 10 in Forms B and C used the new formats but differed in terms of content. The other seven items were common across the three forms.

Table 1

Outline of the Three Forms for the Pilot Study

Item type	Form A	Form B	Form C
Read a text aloud	Item 1 and Item 2	Same items as in Form A	Same items as in Form A
Describe a picture	Item 3	Same item as in Form A	Same item as in Form A
Answer 3 questions using information provided	Item 4, Item 5, Item 6 with existing format	New item formats	Same item format as in Form B, but different content
Answer 3 questions using information provided	Item 7, Item 8, Item 9	Same items as in Form A	Same items as in Form A
Propose a solution	Item 10 with existing format	New item formats	Same item format as in Form B, but different content
State an opinion	Item 11	Same item as in Form A	Same item as in Form A

Note. Items in Form A were all in existing format.

Data Collection

Data for the pilot study were collected from multiple test administrations that took place in October and November 2013 in Korea and Taiwan. Each test taker was asked to answer a background questionnaire before taking the TOEIC Speaking test. All forms were administered according to the same test administration procedures in place for operational administrations of the TOEIC Speaking test. Test takers who had previously taken Form A were not part of the study, and no test takers took more than one form. When recruiting the pilot samples, efforts were made to represent the current test-taking population in terms of demographic characteristics. As a result, given its representation in operational samples, more than 80% of test takers came from Korea. All of the test takers' responses were scored by two certified, trained, and calibrated TOEIC Speaking raters (Everson & Hines, 2010).

Tables 2 and 3 present summaries of the total number of test takers by gender, form, and country. Tables 4 and 5 display the sample sizes of our analysis sample by gender, form, and country. In the analysis sample, test takers with a score of zero on any of the speaking items were screened out from the total sample (except when calculating reliability indices).

Table 2

Sample Size of the Full Data Set by Gender and Form, Korea

Gender	Form A (%) *n* = 319	Form B (%) *n* = 377	Form C (%) *n* = 296	Total
Female	172 (57.9)	263 (73.5)	194 (69.8)	629 (67.4)
Male	125 (42.1)	95 (26.5)	84 (30.2)	304 (32.6)
Total	297	358	278	933

Note. Percentages of male and female test takers within each country are provided in parentheses.

Table 3

Sample Size of the Full Data Set by Gender and Form, Taiwan

Gender	Form A (%) *n* = 319	Form B (%) *n* = 377	Form C (%) *n* = 296	Total
Female	16 (72.7)	17 (89.5)	11 (61.1)	44 (74.6)
Male	6 (27.3)	2 (10.5)	7 (38.9)	15 (25.4)
Total	22	19	18	59

Note. Percentages of male and female test takers within each country are provided in parentheses.

Table 4

Sample Size of the Analysis Sample by Gender and Form, Korea

Gender	Form A (%) *n* = 278 (87)	Form B (%) *n* = 322 (85)	Form C (%) *n* = 274 (93)	Total
Female	150 (58.4)	226 (74.6)	177 (69.1)	553 (67.8)
Male	107 (41.6)	77 (25.4)	79 (30.9)	263 (32.2)
Total	257	303	256	816

Note. Column headings show percentages of test takers remaining in the analysis sample after data screening in parentheses. Percentages of male and female test takers within each country are provided in parentheses in the data cells.

Table 5

Sample Size of the Analysis Sample by Gender and Form, Taiwan

Gender	Form A (%) *n* = 278 (87)	Form B (%) *n* = 322 (85)	Form C (%) *n* = 274 (93)	Total
Female	15 (71.4)	17 (89.5)	11 (61.1)	43 (74.1)
Male	6 (28.6)	2 (10.5)	7 (38.9)	15 (25.9)
Total	21	19	18	58

Note. Column headings show percentages of test takers remaining in the analysis sample after data screening in parentheses. Percentages of male and female test takers within each country are provided in parentheses in the data cells.

Statistical Analyses

Difficulty

To compare the performance of the new and existing formats in terms of difficulty, the following statistics were calculated for each form and compared across forms administered during the pilot study.

1. Standardized score mean difference across test forms. Standardized mean differences among the pilot groups for Forms A, B, and C were calculated based on weighted raw scores on all common items. These score mean differences on common items reflect differences in group ability.
2. Average item scores. Items 4, 5, and 6 are based on the same item stimuli and are usually considered an item set. Because the difficulty level of these three items is controlled at set level instead of item level when assembling forms, average item score is provided only at the set level (denoted as "Avg_456" in Table 6). Similarly, average item score is only provided for Items 7, 8, and 9 as a set (denoted as "Avg_789" in Table 6). In general, items with higher average scores are easier than items with lower average scores.
3. Mean and standard deviation of scaled scores.
4. Adjusted item score means by ANCOVA after controlling group differences on common items and gender. In the ANCOVA model, test form was the treatment factor, weighted common test score was the covariate, and gender was a controlling factor.

Reliability

The reliability of the items and forms with the new formats was evaluated by examining the following statistics.

1. Pearson correlations between item raw scores and weighted total raw scores. An item with a high correlation with the total test score is a good item that can discriminate high-ability test takers from low-ability test takers and can contribute more to the overall test reliability.
2. Interrater agreement measures for evaluating scoring reliability. Because the TOEIC Speaking test is evaluated by raters, it is important to evaluate the consistency of the ratings given by the two raters. These measures included percentage of exact agreement between two raters' ratings for each item: weighted kappa (Haberman, 2012) and correlations between two raters' scores. The two scores (pronunciation and intonation) of Item 1 and Item 2 were considered independently when calculating these item level statistics.

3. Total test reliability coefficient. Reliability refers to the extent to which the assessment scores are consistent over repeated administrations of the same test or alternate forms. Stratified coefficient alpha (Rajaratnam, Cronbach, & Gleser, 1965) was used as a reliability estimate in this study. A high coefficient alpha reliability is desired because it indicates that scores obtained remain consistent over repeated administrations of the same or alternate forms of the test.

Results of the Pilot Study

Evaluation of Difficulty Level

Table 6 presents average scores for each item, item set, common items, and total test. The standardized mean difference (SMD) in weighted raw scores on all common items was 0.17 for Forms B and A and 0.30 for Forms C and A, which indicates that the three pilot groups were not equivalent in terms of ability. Therefore, it was necessary to control group ability differences before making comparisons on the difficulty levels between the new and existing formats. For this reason, an ANCOVA (Howell, 2002) was conducted to take into account group ability differences when comparing item difficulty across forms. The following section introduces the ANCOVA analyses and results.

Table 6

Comparison of Scores Across Three Forms

Scores	Form A Mean (*SD*)	Form B Mean (*SD*)	Form C Mean (*SD*)	*SMD* for Forms B and A	*SMD* for Forms C and D
Weighted score on common items	142.55 (26.22)	147.24 (27.55)	150.65 (27.67)	0.17	0.30
Weighted score on Items 4,5,6, and 10	76.91 (17.87)	77.03 (18.02)	82.01 (16.17)	0.01	0.30
Scaled score	125.5 (29.51)	128.79 (30.75)	134.93 (30.11)		
P1	2.28 (0.55)	2.39 (0.55)	2.39 (0.55)		
P2	2.36 (0.55)	2.44 (0.52)	2.50 (0.55)		
I1	2.23 (0.55)	2.28 (0.57)	2.33 (0.58)		
I2	2.17 (0.51)	2.28 (0.51)	2.30 (0.51)		
3	2.37 (0.68)	2.32 (0.63)	2.46 (0.62)		
Avg_456	2.05 (.60)	2.33 (0.47)	2.58 (0.42)		
Avg_789	1.96 (0.43)	1.97 (0.44)	2.03 (0.41)		
10	3.08 (0.76)	2.80 (0.84)	2.89 (0.78)		
11	2.98 (0.83)	3.16 (0.87)	3.22 (0.92)		

Note. P1 = Item 1 Pronunciation; P2 = Item 2 Pronunciation; I1 = Item 1 Intonation; I2 = Item 2 Intonation; SMD = standardized mean difference.

Controlling Group Ability Differences by ANCOVA Analysis

A further examination of the background data revealed that the three forms had similar background distributions except on gender. Table 7 indicates that Forms B and C had higher percentages of female test takers than Form A. In addition, female test takers performed better than male test takers (see Table 8) on all items. Therefore, gender was selected as a controlling factor, and the weighted raw scores on common items was treated as the covariate in the ANCOVA model.

Two ANCOVA models were run. Both models had form as the treatment variable, gender as a controlling factor, and weighted score on common items as the covariate. The first model used the average score of Item Set 456 as the dependent variable, and the second model used the raw score of Item 10 as the dependent variable. Adjusted group means on the average score of Item Set 456 and the raw score of Item 10 were obtained in each ANCOVA analysis. Table 9 shows the results of the two ANCOVA models.

Table 7

Average Item Score by Gender, Female

Item	Form A (*N* = 165)	Form B (*N* = 243)	Form C (*N* = 188)
Pronunciation	2.42	2.48	2.53
Intonation	2.30	2.35	2.40
3	2.50	2.40	2.55
Avg_456	2.16	2.37	2.65
Avg_789	2.01	2.00	2.08
10	3.19	2.87	2.99
11	3.16	3.26	3.38

Note. Pronunciation is the average of Item 1 and Item 2 pronunciations; intonation is the average of Item 1 and Item 2 intonations.

Table 8

Average Item Score by Gender, Male

Item	Form A (*N* = 113)	Form B (*N* = 79)	Form C (*N* = 86)
Pronunciation	2.17	2.23	2.25
Intonation	2.04	2.08	2.14
3	2.17	2.08	2.24
Avg_456	1.89	2.22	2.44
Avg_789	1.89	1.86	1.91
10	2.92	2.59	2.65
11	2.73	2.86	2.88

Note. Pronunciation is the average of Item 1 and Item 2 pronunciations; intonation is the average of Item 1 and Item 2 intonations.

Table 9

Summary Results for ANCOVA (N = 874)

Model	R^2	Form A[a]	Form B[a]	Form C[a]	Significance test for mean difference
Avg_456 = Form + Weighted Common Test Scores + Gender	.52	2.10 (2.05)	2.32 (2.33)	2.53 (2.58)	$p < .001$ for all pairs
Item 10 = Form + Weighted Common Test Scores + Gender	.47	3.17 (3.08)	2.81 (2.80)	2.82 (2.89)	$p < .0001$ for A vs. B and A vs. C

[a] Adjusted means with unadjusted means in parentheses.

To decide how meaningful these differences in the mean scores were for Item Set 456 and Item 10, we compared the score variations for Item Set 456 and Item 10 in the pilot forms against the score variations of Item Set 456 and Item 10 across all forms administered from January 2012 through November 2013 (see Table 10). As Table 10 shows, the covariate adjusted average scores for Item Set 456 on the pilot forms varied from 2.10 to 2.53, which is within three standard deviations of the mean of Item Set 456 in operational administrations. The difficulty difference between existing and new formats can be considered reasonable on Item Set 456. For Item 10, the average score on the pilot forms varied from 2.81 to 3.17. The average score of Item 10 on the existing Form A (3.17) was more than three standard deviations above the operational mean. The average scores for Forms B and C were well within historical averages.

Table 10

Adjusted Item Scores Compared to Operational Scores

Item/Item set	Form A	Form B	Form C	Mean[a] (*SD*)
Avg 456	2.10	2.32	2.53	2.30 (0.16)
Item 10	3.17	2.81	2.82	2.70 (0.15)

[a] Operational data based on 46 forms administered from January 2012 to November 2013.

Evaluation of Reliability

Table 11 presents correlations of Items 4, 5, 6, and 10 with the weighted total score of the seven common items. The correlations in Forms B and C were similar to those in Form A, and all of the correlation coefficients were larger than 0.30. Items with the new formats performed as well as items with the existing formats in discriminating high- and low-ability test takers.

The total test coefficient alpha reliability information in Table 12 shows that all forms had adequately high reliability. The reliability of the forms with the new item formats (Forms B and C) were higher than the reliability of the form with the existing item formats (Form A).

Tables 13, 14, and 15 present the interrater agreement measures, including percentages of exact agreement, weighted kappas, and correlations between the two ratings. All three pilot forms had adequate to high rater agreement coefficients, indicating that the overall scoring reliability was adequately high for forms with both new and existing formats.

Table 11

Correlations Between New Format Item and Weighted Common Test Scores

Form	N	Item 4	Item 5	Item 6	Item 10
A	278	0.47	0.48	0.64	0.65
B	322	0.46	0.50	0.60	0.67
C	274	0.45	0.49	0.60	0.71

Table 12

Total Test Coefficient Alpha Reliability

Form	Reliability
A	0.87
B	0.91
C	0.91

Table 13

Interrater Reliability: Exact Agreement

Item	Form A	Form B	Form C
I1: Item 1 Intonation	62	66	64
P1: Item 1 Pronunciation	67	72	71
I2: Item 2 Intonation	71	76	72
P2: Item 2 Pronunciation	66	70	71
3	66	64	67
4	76	75	72
5	70	67	80
6	67	63	72
7	88	86	90
8	73	69	72
9	89	85	86
10	66	74	69
11	70	68	70

Sample sizes for Form A ranged from 317 to 319 across different items, from 375 to 377 for Form B, and from 294 to 296 for Form C.

Table 14

Interrater Reliability: Weighted Kappa

Item	Form A	Form B	Form C
I1: Item 1 Intonation	0.40	0.50	0.47
P1: Item 1 Pronunciation	0.49	0.57	0.57
I2: Item 2 Intonation	0.49	0.59	0.53
P2: Item 2 Pronunciation	0.41	0.52	0.56
3	0.64	0.64	0.67
4	0.82	0.76	0.60
5	0.73	0.64	0.65
6	0.67	0.63	0.70
7	0.86	0.86	0.84
8	0.84	0.82	0.79
9	0.83	0.75	0.78
10	0.71	0.87	0.80
11	0.77	0.84	0.86

Note. Sample sizes for Form A ranged from 317 to 319 across different items, from 375 to 377 for Form B, and from 294 to 296 for Form C.

Table 15

Interrater Reliability: Correlation

Item	Form A	Form B	Form C
I1: Item 1 Intonation	0.40	0.50	0.47
P1: Item 1 Pronunciation	0.50	0.57	0.57
I2: Item 2 Intonation	0.49	0.60	0.53
P2: Item 2 Pronunciation	0.42	0.52	0.56
3	0.64	0.64	0.67
4	0.82	0.76	0.60
5	0.73	0.65	0.66
6	0.67	0.64	0.70
7	0.86	0.86	0.85
8	0.84	0.82	0.79
9	0.83	0.75	0.78
10	0.71	0.87	0.80
11	0.77	0.84	0.86

Note. Sample sizes for Form A ranged from 317 to 319 across different items, from 375 to 377 for Form B, and from 294 to 296 for Form C.

Difficulties of the Expanded Item Formats in Operational Administrations

The new formats of Item Set 456 and Item 10 have been used in operational practice along with the existing formats since May 2015. To monitor the difficulties of the new formats, the scores of Item Set 456 and Item 10 with the new formats were compared to the scores of Item Set 456 and Item 10 with the existing formats. In this paper, item scores were compared separately for two types of operational forms: SP (secured program) and SSP (special secured program) forms. Although SP and SSP forms have the same test specifications, SP forms are administered once a month in Korea and other Asian countries, whereas SSP forms are administered only in Korea.

Figure 1 shows plots of the scores of Item Set 456 in all SP forms administered from February 2014 through June 2016 in Asian countries. The red diamonds note the scores for Item Set 456 with the new formats. Forms administered before May 2015 were included to provide a reference for the comparison between new and existing item formats. In total, 58 SP forms were administered from February 2014 through June 2016, including the 11 forms with new formats for Item Set 456 administered after May 2015. The average score of Item Set 456 with the new formats ranged from 2.06 to 2.51, with a mean of 2.33. Similarly, the average score of Item Set 456 with the existing formats ranged from 2.02 to 2.60, with a mean of 2.31. Figure 1 shows that Item Set 456 with new formats was similar in difficulty level to those with existing formats.

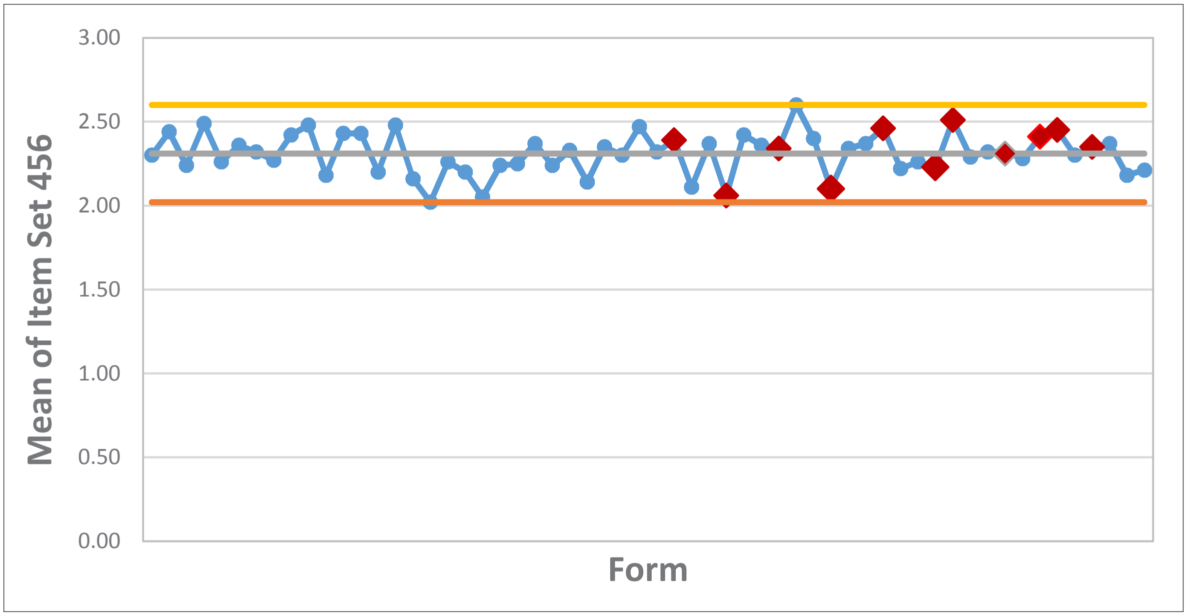


***Figure 1*. Comparison of the means of Item Set 456 from February 2014–June 2016 in operational administrations (SP forms). *Note*. Red diamonds denote new formats.**

Figure 2 shows plots of the scores of Item 10 in all SP forms administered from February 2014 through June 2016 in Asian countries. After May 2015, nine forms used the new formats for Item 10. The scores of Item 10 with the new formats ranged from 2.40 to 2.84 with a mean of 2.62. The scores of Item 10 with the existing formats ranged from 2.32 to 2.95 with a mean of 2.62. Figure 2 shows that Item 10 with the new formats also had a difficulty level similar to those with the existing formats.

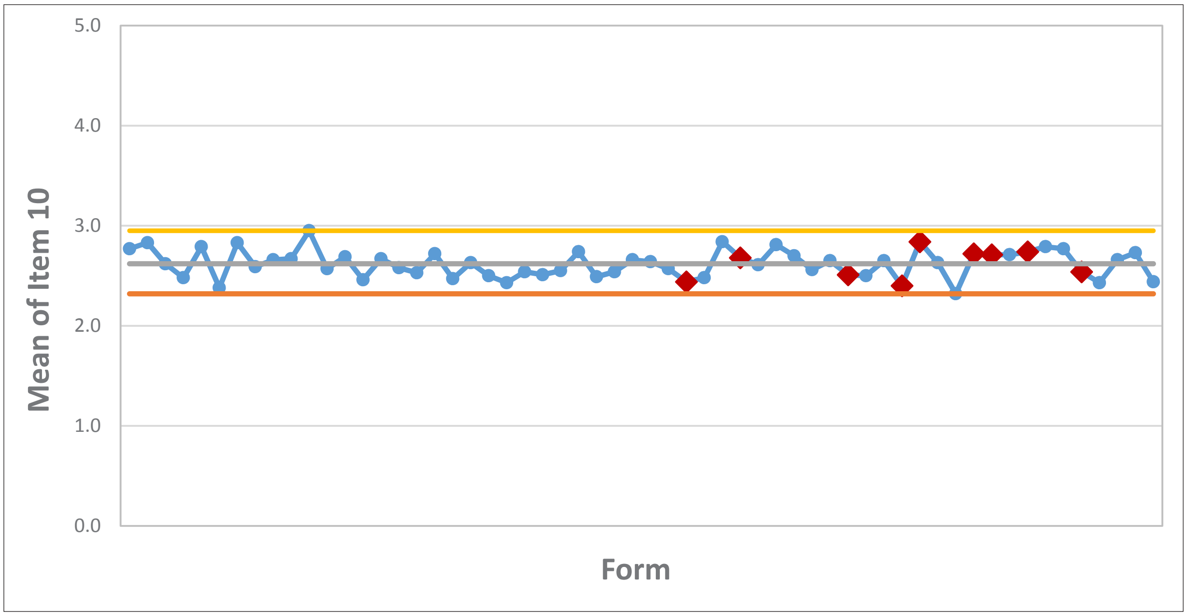


***Figure 2*. Comparison of the means of Item 10 from February 2014–June 2016 in operational administrations (SP forms). *Note*. Red diamonds denote new formats.**

Figure 3 shows plots of the scores of Item Set 456 for all SSP forms administered in Korea from February 2014 through June 2016. Nine forms (out of 358) had Item Set 456 in the new formats. The scores of Item Set 456 with the new formats ranged from 2.31 to 2.52 with a mean of 2.41. This is within the range of the scores of Item Set 456 with the existing formats (1.81 to 2.61, with a mean of 2.32).

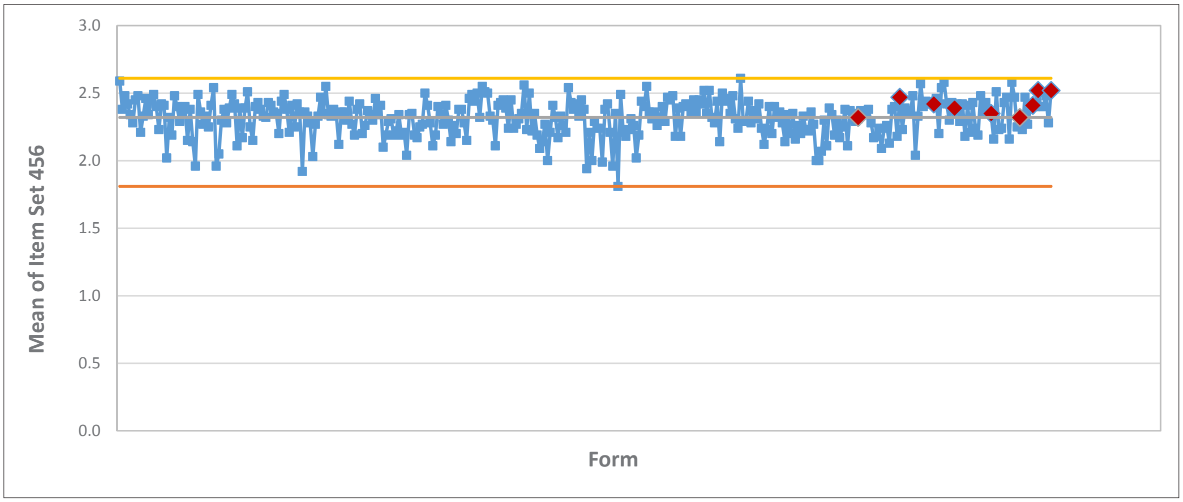


***Figure 3*. Comparison of the means of Item Set 456 from February 2014–June 2016 in operational administrations (SSP forms). *Note*. Red diamonds denote new formats.**

Figure 4 shows the scores of Item 10 for all SSP forms administered from February 2014 through June 2016. Four forms (out of 358) had Item 10 in the new formats. The scores of Item 10 with the new formats ranged from 2.57 to 2.79 with a mean of 2.72. The scores of Item 10 with the existing formats ranged from 2.10 to 2.98 with a mean of 2.60. Therefore, in both SP and SSP administrations, the score means were similar between new and existing formats for both Item Set 456 and Item 10.

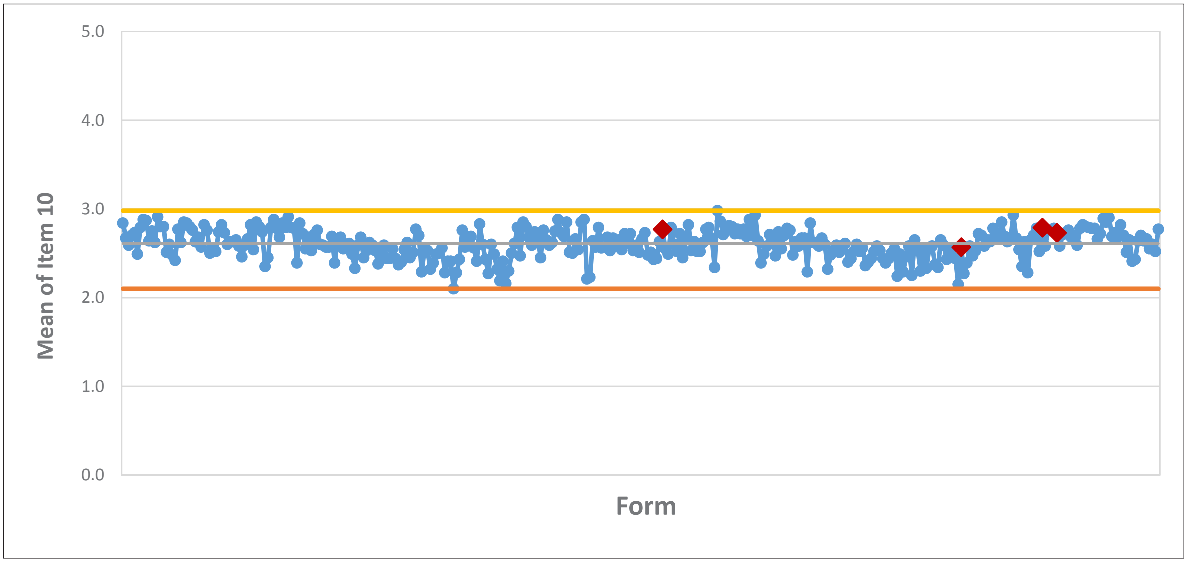


***Figure 4*. Comparison of the means of Item 10 from February 2014–June 2016 in operational administrations (SSP forms). *Note*. Red diamonds denote new formats.**

Reliabilities for Tests With Existing and New Formats in Operational Administrations

Reliability estimates averaged across operational forms administered from February 2014 through June 2016 are provided in Table 16 for the new and existing formats separately. The SP and SSP forms are included in the comparison. Both interrater reliability estimates (interrater correlation) at the item level and the internal consistency (coefficient alpha) reliability estimate at the test level are similar between forms with the new formats and forms with the existing formats.

Table 16

Average Reliability Estimates for Operational Forms With New and Existing Formats From February 2014 Through June 2016

Format	Interrater reliability estimate				Internal consistency reliability estimate for total test
	Item 4	Item 5	Item 6	Item 10	
New formats	0.67 (*n* = 20)	0.61 (*n* = 20)	0.48 (*n* = 20)	0.71 (*n* = 13)	0.81
Existing formats	0.62 (*n* = 383)	0.62 (*n* = 383)	0.52 (*n* = 383)	0.67 (*n* = 383)	0.80

Concluding Remarks

In this report, we describe an evaluation of whether expanded item formats of the TOEIC Speaking test impacted item difficulty and test reliability. As noted at the outset, these expanded formats were intended to expand coverage in a way that was thought to foster language learning and to discourage the use of undesirable test-taking strategies. The results of this study suggest that modifications to existing item formats had a slight effect on the difficulty of items, as some items were more difficult and others were less difficult. However, the effects observed were basically within the range of variation typically observed across alternate forms of the test. Further monitoring of the difficulties of the new item formats in operational practice also indicates that items with the new formats have performed similarly to items with existing formats in operational practice. In operational administrations, forms with the new formats have also had reliability estimates similar to those with the existing formats. In conclusion, efforts to improve selected TOEIC Speaking items so as to better foster communicative language learning appears not to have had any significant undesirable effects on item difficulty or test score reliability.

References

Everson, P., & Hines, S. (2010). How ETS scores the *TOEIC®* Speaking and Writing tests' responses. In *The research foundation for the TOEIC® tests: A compendium of studies* (pp. 8.1–8.9). Princeton, NJ: Educational Testing Service.

Haberman, S. (2012). *Measures of agreement* [PowerPoint presentation]. Princeton, NJ: Educational Testing Service.

Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: Duxbury/ Thomson Learning.

Qu, Y., Liu, J. & Chan, E. (2013). *Changing the current weights of the TOEIC Speaking test* [Internal memorandum]. Princeton, NJ: Educational Testing Service.

Rajaratnam, N., Cronbach, L. J., & Gleser G. C. (1965). Generalizability of stratified-parallel tests. *Psychometrika*, *30*, 39–56. https://doi.org/10.1007/BF02289746

Appendix. Expanded Question Formats

Table A1

Expanded Question Formats for Items 4–6

Task: Respond to questions	Existing formats	New formats
Familiar topics and personal experiences	x	
Begin with "Imagine that . . ."	x	
Talk on the telephone with a marketing firm	x	
Hear and read the questions	x	
Two 15-second and one 30-second response	x	
No preparation time	x	
Current rubric and scoring rules	x	
Familiar topics and personal experiences		x
Begin with "Imagine that . . ."		x
Talk on the telephone with an employee, colleague, friend, etc.		x
Hear and read the questions		x
Two 15-second and one 30-second response		x
No preparation time		x
Current rubric and scoring rules		x

Note. Bolded parts note the difference between the two formats.

Table A2

Expanded Question Formats for Item 10

Task: Propose a solution	Existing formats	New formats
Single speaker	x	
Recognize the problem	x	
Propose a way of dealing with the problem	x	
Listen to the question, no reading	x	
60-second response	x	
30 seconds of preparation time	x	
Current rubric and scoring rules	x	
Two people speaking at a meeting		x
Recognize the problem		x
Propose a way of dealing with the problem		x
Listen to the question, no reading		x
60-second response		x
30 seconds of preparation time		x
Current rubric and scoring rules		x

Note. Bolded parts are the difference between the two formats.

Compendium Study

Analyzing Item Generation With Natural Language Processing Tools for the *TOEIC*® Listening Test

Su-Youn Yoon, Chong Min Lee, Patrick Houghton, Melissa Lopez, Jennifer Sakano, Anastassia Loukina, Bob Krovetz, Chi Lu, and Nitin Madnani

ETS® TOEIC®

Assess to Progress.

Recent developments in natural language processing (NLP) technology and massive online resources have substantially changed the environment of language learning. Online materials are useful sources of authentic situations and language use, and they have been frequently used in generating vocabulary lists (Capel, 2010; Coxhead, 2000; Fuentes, 2002) and in examples of collocation expressions (Chen, Huang, Chang, & Liou, 2015; Liou et al., 2013).

Another frequent use of online resources and computerized corpora is the development of listening and reading materials. Several studies have explored the application of NLP technologies to the selection of appropriate reading or listening materials for students who have English as a second language (ESL), with most studies focused on evaluating the difficulty of materials. Graesser, McNamara, Louwerse, and Cai (2004) and Sheehan, Kostin, Napolitano, and Flor (2014) developed automated systems to provide the overall difficulty score of written text in English. One of the primary goals of these systems is to provide native and ESL students with reading or listening materials according to their grade or language proficiency level. These studies focused on estimating difficulty for the given texts.

In order to create high-quality reading or listening materials from online resources, aspects other than difficulty also need to be considered. Heitler (2005) developed a manual on how to prepare classroom materials from online resources and provided useful strategies such as adjusting text length, replacing vocabulary, simplifying syntactic structure, and resolving proper names and abbreviations. To reduce this manual effort, an initial selection of appropriate materials that can be quickly and easily adapted into learning materials is necessary. However, few studies have discussed the characteristics of such appropriate materials. Furthermore, previous studies have mostly focused on generating learning materials (e.g., classroom materials) and have not discussed characteristics of appropriate materials for language assessments. As mentioned in Hoshino and Nakagawa (2007), automated material selection is more difficult to apply to language assessments than learning materials because the former is subject to greater strictures on language variety, type, and difficulty. There may be additional requirements that the online resources must meet in order to be considered appropriate assessment materials, and these requirements increase the difficulty of fully automated material selection.

In this study, we developed assistive tools based on NLP technology and online resources to support listening item generation for *TOEIC®* Listening, a large-scale international English proficiency test. In contrast to previous studies, which focused on the automated generation of limited item types such as a cloze test for vocabulary and prepositions (e.g., Heilman & Smith, 2010; Huang, Chen, & Sun, 2012; Huang, Tseng, Sun, & Chen, 2014), our tools support diverse tasks for a multitude of different item types.

We developed three tools: an automated system that retrieves appropriate real-world videos, a list of vocabulary associated with difficulty levels, and a tool that suggests words that occur frequently in similar contexts. These tools were expected to improve the quality of items by increasing the diversity and authenticity of contexts and vocabulary, which would also increase the efficiency of item generation because diversity and authenticity prevent overlap among items and reduce the amount of revision as a result.

This study addresses the following points:

- We provide a discussion about the characteristics of appropriate materials for listening items based on an annotation study with two expert language test developers.
- We provide three tools that support the main tasks related to listening item generation: passage generation, adjustment of the word difficulty used in items, and distractor generation.
- We examine the usefulness of these tools through a small-scale item generation study.

Assistive Tools Used in This Study

In this study, we classified listening item generation process into three stages and developed a tool to support each stage as follows:

- Brainstorming and idea generation: seed video retrieval system
- Distractor generation: word similarity tool
- Revision and adjustment of the created item: vocabulary list

Word Similarity Tool

We used a tool to identify words that convey similar meanings (e.g., student and learner) or related meanings (e.g., student and school) developed by Heilman and Madnani (2012). Using NLP techniques, researchers employed empirical approaches to assess lexical associations. Based on the intuition that words with similar contextual distribution (i.e., the linguistic contexts that they appear in) will have similar or related meanings, they calculated distributional similarities among words from large text corpora. Following this line of research, we first estimated distributional similarities among words based on Dekang Lin's Distributional Thesaurus and stored them in a large database. We provided a web-based user interface, and it returned the 10 most similar words, based on similarity score, given the query word provided by the item writers.

Vocabulary List

We created a vocabulary list by combining the following three vocabulary lists:

- New General Service List (NGSL): A word list designed for general service purposes. The list is composed of the 2,800 most frequently occurring words extracted from a subset of the Cambridge English corpus, which includes approximately 270 million words.
- Lemmatized British National Corpus (BNC) frequency list: A word list including the top 6,318 most frequently occurring words from the BNC
- Corpus-based list: A word list including the top 7,699 most frequently occurring words from an English Gigaword corpus

The words from these three lists were classified into four groups. First, we made a separate category for 368 function words such as articles, prepositions, and pronouns. Next, we classified the remaining vocabulary into three tiers: Tier 1 for basic vocabulary, Tier 2 for intermediate level vocabulary, and Tier 3 for topic-specific vocabulary. As the tier increases, the difficulty of the vocabulary also increases. The difficulty level (the tier the word belongs to) was determined based on the source, rather than its frequency in the specific corpus. 2,551 words in the NGSL list excluding function words were assigned to Tier 1; 1,712 words in the BNC but not in the NGSL and function words were assigned to Tier 2; and 3,478 words in the mixed corpora-based list but not in NGSL, BNC, or function words were assigned to Tier 3. The final list was composed of 8,109 unique words.

We also provided a separate vocabulary list created using a large pool of listening items. The item corpus was composed of 19,460 listening items extracted the TOEIC Listening test. The list included a total of 3,503 unique words, their tier information when available, and the number of items that include this word.

Automated Seed Video Retrieval System

Good items make use of authentic language used in varied situations. Writing a listening item that provides an appropriate level of difficulty, reflecting authentic language use while avoiding duplicates, is a difficult task. Writing such an item about unfamiliar topics is an especially challenging task for item writers and it requires a substantial amount of research to find initial ideas.

The most frequently used approach to develop items in unfamiliar contexts is searching web resources. However, finding the appropriate materials with unguided searching is not an easy task, and item writers tend to spend time reviewing useless web resources retrieved from search engines. Results from web searches usually contain a large amount of irrelevant material to sift through, including redundant or unrelated content as well as content inappropriate for language assessments.

In order to address this issue, we designed an assistive tool, called the Seed Video Retrieval System, for test item writers. This tool provides web resources that have a greater likelihood of being useful for writing test items. When more helpful web resources are available, item writers can reduce time wasted finding resources from retrieved search results. This system is designed to retrieve only YouTube® videos that meet certain constraints, when users enter search keywords.

Although text resources are also available as web resources, we decided to focus on YouTube videos due to a few advantages of videos over text resources. After an initial attempt to use web pages as resources for item writers, we observed challenges and drawbacks in extracted text resources. Item writers wanted a small number of concise data sources relevant to their keywords. They also wanted data containing contexts with enough development to allow them to understand the content. The power of videos in conveying content is well expressed in an idiom: a picture is worth a thousand words. Visual images in videos can provide more information to the viewer than words in texts. So, videos can be more concise while providing as much or more information. For example, it can be easier to figure out which vocabulary words need to be used in which situation when an item writer watches a video. The images in a video inherently contain lots of contextual information on places, tools, roles, and so forth. Furthermore, extracted text data usually contained too much text to read and texts on topics outside item writers' fields of expertise, which required further research to understand. Moreover, web pages usually contained redundant data such as HTML tags and content irrelevant to search keywords. It was a technical challenge to automatically remove the redundant or irrelevant data from the set of retrieved pages, in order to provide a useful tool.

During our initial exploration of YouTube videos as a resource, we discovered some challenges for item writers who might seek to use them. Some videos were too long, too difficult to understand, or too incoherent to make items. We will further discuss the characteristics of appropriate videos in the Participants section. Based on a qualitative analysis using a subset of data, we found that a higher percentage of videos with manual transcriptions contained coherent content and better audio quality. Manual transcription means that the video's owners provided transcripts of speech in the videos when they were uploaded. The existence of manual transcriptions could be indirect evidence that uploaders paid more attention to the quality of the videos and that they also considered their audiences. As a result, a higher percentage of such videos were appropriate for item generation than was the case for videos lacking these manual transcriptions. Based on these findings, we developed an automated system using the YouTube API with refined search conditions. The system retrieved videos with a manual transcription shorter than 4 minutes in length.

When we tested our video retrieval system, we found that the search skills differed greatly across the individual item writers, and the usefulness of the tool also substantially varied depending on their skills. Therefore, instead of providing the tool itself, we created a set of key words by concatenating topic and genre words provided by expert item writers. We selected four topics and collected a total of 664 videos. For each video, the title, the key words used in the video search, and the link to the video were provided in an Excel spreadsheet. The quality of this video data collection is analyzed in the Participants section.

User Study

In order to investigate the usefulness of the assistive tools, we conducted a small-scale pilot study. The participants took part in an 8-week item-writing program, and during the program they were asked to use the tools described above to assist them in creating items. At the end of the program, the participants completed a survey and answered questions during a follow-up interview about the usefulness of these tools.

Participants

Applicants filled out a form where they created several types of common listening items. These were scored blindly by experienced item writers, organizers of the 8-week item-writing program, without any personal information about the applicant. The selected item writers consisted of six women and one man. Their educational backgrounds included undergraduate students with different majors (e.g., French, history, education, and journalism), a university professor teaching English to speakers of other languages, and a public school teacher of bilingual education. Two of the participants had substantial experience in item generation and participated in the same item-writing program for 3 years. The other five item writers were first-time participants.

Tasks

The participants were asked to create the following three types of TOEIC Listening items:

- Type 1: The test taker is presented with a picture and four recorded statements and asked to select the statement that best describes the picture.
- Type 2: The test taker listens to a conversation between two speakers and answers a series of written questions about the content of the conversation.
- Type 3: The test taker listens to a recording of a single speaker (e.g., announcement or advertisement) and answers a series of written multiple-choice questions about the content of the recording.

The tools were introduced to participants in the second week of the program. We provided a 30-minute presentation and question-and-answer session, as well as written manuals. Both the seed videos and vocabulary list were presented as spreadsheets, and the word similarity tool was presented as a website. All participants were requested to use the tools during first 2 weeks of the test period. After this initial test period, the use of tools was optional, but all participants used at least one tool throughout the entire program. During the 8-week program, each participant created 18 Type 1, 40 Type 2, and 40 Type 3 items, on average.

We asked participants about their usage of the tools using a survey and interview on the last day of the program. The survey was composed of two questions about the participants' background (experience in item generation) and 21 questions about their experience with the tools, divided over four sections in the survey: frequency of use, perceived usefulness, method of use, and future improvements. Multiple-choice questions were used for the frequency of use section. For the perceived usefulness section, we used 12 Likert-type questions (four questions for each tool). Higher point responses indicated a higher degree of usefulness in item generation. Finally, open-ended questions were used for both the method of use and future improvement sections in the survey.

There were follow-up interviews after the survey responses were collected. Participants' survey responses were reviewed before the interviews, and two researchers in this study asked questions to understand survey responses further. Participants were asked to clarify why they did or did not use particular tools and how the tools were used in the item writing process and to expand on some of the shorter responses. In this way, we were able to pinpoint the ways in which the tools were successful and the aspects we could focus on improving. The interviews also allowed some context in which to evaluate the multiple-choice and Likert responses qualitatively.

In the next section in this report, we provide some insight into the participants' evaluation of the usefulness of the tools. Therefore, we focus on frequency of use, perceived usefulness, and method of use.

Results

To What Extent Are Automatically Retrieved Resources Appropriate for Item Generation?

In order to evaluate the quality of videos retrieved by the automated seed video retrieval system, two experienced item writers were recruited to rate 664 videos. First, they were asked to rate the holistic quality of each video with regard to its appropriateness as a seed video (Is the video helpful in item writing?). In addition, they answered the following five subquestions:

- Does the video contain a new context? (new context)
- Does the video contain sufficient information to understand it? (sufficient info)
- Is the content of the video appropriate for the test? (appropriate content)
- Does the video provide good examples of formal language? (formal language)
- Is the video generally appropriate in terms of vocabulary difficulty? (vocabulary difficulty)

The first question in the list above is about whether a retrieved video contains a new context that reflects contemporary language expressions and situations that have not been frequently used in existing items. The second question is about whether an annotator understands a given video without referring to other resources. The third question is about whether the content of a video could be used in test items. The fourth question serves to help figure out if the video contains words of a level of formality that is useful for test item writing. The fifth question is designed to explore the influence that the vocabulary difficulty of a video has on the usefulness of that video.

For each question, annotators were asked to choose one answer: yes, maybe, or no. *Yes* means that a video is highly likely to be qualified for the stated characteristic, *no* means that a video is highly unlikely to be qualified for the stated characteristic, and *maybe* means that a video is likely to be somewhat qualified for the stated characteristic.

Table 1 shows the distributions of annotators' answers on the questions. In addition, each cell contains a count and its ratio (a count of yes, maybe, or no divided by the count of all videos).

Table 1

Distribution of Annotations

Question	Annotator 1			Annotator 2		
	Yes	Maybe	No	Yes	Maybe	No
Seed video	286 (43%)	210 (32%)	168 (25%)	397 (60%)	209 (31%)	58 (9%)
New context	428 (64%)	164 (25%)	72 (11%)	509 (77%)	147 (22%)	8 (1%)
Sufficient info	283 (43%)	199 (30%)	182 (27%)	263 (40%)	258 (39%)	143 (22%)
Appropriate content	389 (59%)	161 (24%)	114 (17%)	488 (73%)	152 (23%)	23 (4%)
Formal language	597 (90%)	44 (7%)	23 (3%)	546 (82%)	68 (10%)	50 (8%)
Vocabulary difficulty	531 (80%)	73 (11%)	60 (9%)	511 (77%)	77 (12%)	76 (11%)

Annotator 1 and 2 considered 43% and 60%, respectively, of 664 videos to be appropriate seed videos that could be helpful in writing test items. The number of videos for which both annotators answered yes on the main question about appropriateness as a seed video was 243 (36.6%). The number of videos for which at least one annotator marked yes was 440 (66.3%). So, depending on item writers' needs, over half of the retrieved videos could be helpful in writing test items. In order to calculate the interannotator agreement, we converted ratings into a numeric scale: 1 for yes, 2 for maybe, and 3 for no. The quadratic weighted kappa on the main question was 0.51.

Both annotators thought that most videos (from 60% to 90%) met criteria for new context, appropriate content, formal language, and vocabulary difficulty; however the proportion of videos that contained sufficient information was substantially lower (ranging from 40% to 43%). A possible reason that the majority of retrieved videos could meet the prescribed criteria was the search conditions we adopted. We only selected videos with manual captions and these results were in line with our expectations.

As an initial effort to develop an automated classifier that predicts the holistic quality of seed videos, we investigated to what extent the manual annotations of the five subquestions could accurately categorize the retrieved videos into *appropriate*, *maybe*, or *inappropriate* seed videos. We converted yes, maybe, and no answers into 1, 2, and 3, respectively, and then trained multiple linear regression models with the seed video question as a dependent variable and the five subquestions as independent variables. We used all 664 videos for the model building and reported model fits in the training data. We tried all combinations of the five subquestions and reported the best performers for each size of combination in terms of the coefficient of determination (R-squared, R^2). We excluded vocabulary difficulty because including it in the model did not result in significant improvement in R^2 score. All of the regressions in Table 2 were significant at p-value < 0.001.

Table 2

Regression Analysis Using Annotated Data

Size of combination	Annotator 1			Annotator 2		
	Features	R^2	Adjusted R^2	Features	R^2	Adjusted R^2
1	New context	0.753	0.752	Appropriate content	0.562	0.561
2	New context + sufficient info	0.866	0.865	Appropriate content + new context	0.647	0.647
3	New context + sufficient info + appropriate content	0.880	0.880	Appropriate content + new context + sufficient info +	0.700	0.700
4	New context + sufficient info +appropriate content + formal language	0.884	0.883	Appropriate content + new context + sufficient info + formal language	0.717	0.717

Table 2 shows which combinations of subquestions lead to improvements of both R^2 and adjusted R^2 values. For example, when only one factor is considered, new context and appropriate content were the best factors for Annotators 1 and 2, respectively. The best adjusted R^2 scores for Annotators 1 and 2 (0.883 and 0.717, respectively) were achieved using the combination of new context, sufficient info, appropriate content, and formal language. Previous studies about automated listening and reading material selection mostly focused on difficulty, and other dimensions such as topics and content have been neglected. This analysis shows the importance of these dimensions. In particular, for assistive item generation, they are the most important factors in determining the appropriateness of the materials.

Can Assistive Tools Improve the Item Generation Process?

Survey Questions 1, 2, and 3 from the pilot test of the tools solicited information about the frequency of use for each tool. All participants used at least one tool, once to a few times per week. Frequency of use for each tool is presented in Table 3. In general, the word similarity tool was the most frequently used among the three tools, and four participants (57%) used it more than once a day. It was followed in popularity by the seed video list and the vocabulary list.

Table 3

Frequency of Use for Each Tool

Tool	Never	Once	Once to a few times per week	Daily	Multiple times a day	Total
Word similarity tool	1	1	1	1	3	7
Vocabulary list	1	2	4	0	0	7
Seed videos	0	1	6	0	0	7

Next, we asked the participants about how they used each tool during item generation. The participants provided a short description, and we got further detailed explanation during the follow-up interviews. The word similarity tool was used to find similar words (much like a thesaurus) and avoid repetition of words both in stimuli and items. One participant specifically mentioned it was used for distractor generation. The vocabulary list was used to adjust the difficulty of vocabulary in both stimuli and items. The tool was used both in addition of words (low or medium frequency words) and removal of words (high frequency words, removed to avoid repetition). One participant used it to create a list of context ideas by using a random word and phrase generator that was provided with the tool. The seed video tool was used primarily for idea generation of Item Types 2 and 3. In addition, three participants used videos to extract authentic language expressions and terms for specific fields.

Twelve survey questions solicited the perceived usefulness for each tool. We asked questions about the following topics:

- Speed of item generation: Using the tool enabled me to write items more quickly.
- Quality of created item: Using the tool improved the feedback I received for quality of my items.
- Diversity in context and vocabulary (subquestion for quality): Using the tool made it easier to create a larger variety of items (with regard to contexts, difficulty, etc.).
- Overall usefulness: I found the tool useful in my job.

Each question had a 4-point Likert scale, where 1 indicated strong disagreement and 4 indicated strong agreement. In addition, the participants could select not applicable if, for instance, they had not used a particular tool beyond the initial requested period or did not effectively use the tools in generating any items. Indeed, some participants found some of the tools to be more time-consuming than useful.

We investigated the usefulness of each tool, and the four questions for each tool were combined into a single composite score during analysis. A total of 28 responses (7 participants multiplied by 4 questions) were available for each tool. Table 4 summarizes the results.

Table 4

Frequency of Use for Each Tool

Tool		Strongly disagree	Somewhat disagree	Somewhat agree	Strongly agree	Not applicable
Word similarity tool	No. of responses	1	0	9	12	6
	%	4%	0%	32%	43%	21%
Vocabulary list	No. of responses	0	3	9	2	14
	%	0%	11%	32%	7%	50%
Seed videos	No. of responses	1	13	8	5	1
	%	4%	46%	29%	18%	4%

We found the most positive responses for the word similarity tool. For the general usefulness question (Q4), six participants agreed or strongly agreed that the tool was useful in item generation, with four participants indicating strong agreement.

The vocabulary list was less frequently used than the other tools, and the participants chose not applicable for 50% of responses across four questions. When we excluded not applicable responses, participants provided substantially more positive reactions than negative reactions. For the general usefulness question (Q4), four participants (75% after excluding participants who chose not applicable) agreed or strongly agreed that the tool was useful in item generation. Thus, we can see that the tool was useful for the smaller group of participants who actually used it.

The seed video list was the most widely used of the tools, and the ratings varied across different question types. For the general usefulness question (Q4), four participants agreed or strongly agreed that the tool was useful in item generation, and the positive response was slightly more frequent than negative responses (three participants disagreed). The tool received the most positive evaluation for diversity of context and vocabulary (Q3), and five participants agreed or strongly agreed that the tool increased the variety in created items. The tool received the least positive evaluation for speed of item generation (Q1), with five participants disagreeing that the tool increased the speed of item generation. The tool was favored by one of the experienced item writers, who strongly agreed that the tool was useful in item generation. She pointed out that the seed video list may be more useful for experienced item writers who may have exhausted ideas for new items. The vocabulary tools, she felt, may be useful for novice item writers who are not yet familiar with the tasks and the kinds of vocabulary that are appropriate for potential test takers. This suggests potential differences in the usefulness of tools between experienced item writers and new item writers.

In an additional analysis, we converted each option to a numeric value and calculated the mean of Likert-scale scores for each tool. Strongly disagree, disagree, somewhat agree, and strongly agree were mapped into 1, 2, 3, and 4, respectively, and not applicable was excluded from analysis. The mean scores for the word similarity tool, vocabulary list, and seed video list were 3.45, 2.93, and 2.63, respectively, on a 4-point scale.

Finally, participants provided comments about how to improve tools. Many comments were related to the organization and presentation of the seed video collection. Because the participants were not assigned to create items on a specific topic, we initially hypothesised that participants may use any video if it included appropriate materials for the target language proficiency test. However, in reality, participants first made a decision about a narrow topic of the item and started searching videos relevant to the specific topic. As a result, an efficient interface to help search within the video data collection was required. Here are some detailed comments:

- Descriptions: In addition to the YouTube video title, the participants requested a short summary for each video.

- Content overlap: The video collection included multiple videos that were not identical but similar in content. The participants suggested removing videos with similar content to reduce the overlap.

Based on these comments, we are currently improving the automated seed video retrieval system. First, we will provide the category and video uploader information for each video, in order to improve the descriptions of videos. To reduce the overlapping content, we set a limit on the number of videos from any particular video uploader. Additionally, we calculated the similarity of different videos by applying a vector space model and selected only one video from sets of overly similar videos.

Conclusions

In this study, we explored the use of existing resources and NLP technology to support listening item generation for the TOEIC Listening test. Good items need to use authentic situations and language in a wide variety of contexts. However, creating items for less familiar topics is a challenging task for item writers. As a result, the item writers tend to create items for familiar topics, and this can result in an imbalance in contexts and vocabulary. Most item writers tend to have expertise in the education and English-language fields, which leads to overlap in experience from which to draw item ideas. To address this issue, we developed an automated seed video retrieval system, a list of vocabulary, and a word similarity tool. To examine the usefulness of these tools, we conducted a small-scale pilot study. Seven item writers created TOEIC Listening items using our tools and responded to a survey and interview on the last day of the pilot study. We evaluated the usefulness and impact of these tools on item generation based on the survey responses. In general, all tools were considered useful, and the word similarity tool in particular was rated the most useful. The preference of particular resources may vary across different item writers. The word similarity tool was most favored overall (four novice item writers), and the seed video collection was most useful to one of the experienced item writers. This finding suggests potential differences in the usefulness of tools between experienced item writers and new item writers. In our future exploration of this topic, we will extend our study and further investigate the impact of our resources with experienced item writers.

References

Capel, A. (2010). A1–B2 vocabulary: Insights and issues arising from the English Profile Word-lists project. *English Profile Journal, 1.* https://doi.org/10.1017/S2041536210000048

Chen, M.-H., Huang, S.-T., Chang, J. S., & Liou, H.-C. (2015). Developing a corpus-based paraphrase tool to improve EFL learners' writing skills. *Computer Assisted Language Learning*, *28*(1), 22–40. https://doi.org/10.1080/09588221.2013.783873

Coxhead, A. (2000). A new academic word list. TESOL *Quarterly*, *34*, 213–238. https://doi.org/10.2307/3587951

Fuentes, A. C. (2002, April). Exploitation and assessment of a business English corpus through language learning tasks. *ICAME Journal*, *26*, 5–32.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Cohmetrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*, 193–202. https://doi.org/10.3758/BF03195564

Heilman, M., & Madnani, N. (2012). *A unified resource for distributional lexical similarity*. Unpublished manuscript.

Heilman, M., & Smith, N. A. (2010). Good question! Statistical ranking for question generation. In *Human language technologies: The 2010 annual conference of the North American Chapter of the Association for Computational Linguistics* (pp. 609–617). Stroudsburg, PA: Association for Computational Linguistics.

Heitler, D. (2005). *Teaching with authentic materials*. Retrieved from the Pearson Intelligent Business website: http://www.pearsonlongman.com/intelligent_business/images/teachers_resourse/pdf4.pdf

Hoshino, A., & Nakagawa, H. (2007). Sakumon: An assistance system for English cloze test. In R. Carlsen, K. McFerrin, J. Price, R. Weber, & D. A. Willis (Eds.), *Proceedings of Society for Information Technology & Teacher Education International Conference 2007*. San Antonio, TX: Association for the Advancement of Computing in Education.

Huang, Y.-T., Chen, M. C., & Sun, Y. S. (2012). Personalized automatic quiz generation based on proficiency level estimation. In *Proceedings of the 20th International Conference on Computers in Education* (pp. 553–560). Retrieved from http://autoquizhttp.iis.sinica.edu.tw/docs/personalized.pdf

Huang, Y.-T., Tseng, Y.-M., Sun, Y. S., & Chen, M. C. (2014). TED quiz: Automatic quiz generation for TED Talks video clips to assess listening comprehension. In *Proceedings of the 14th IEEE International Conference on Advanced Learning Technologies* (pp. 350–354). Piscataway, NJ: IEEE. https://doi.org/10.1109/ICALT.2014.105

Liou, H.-C., Chang, J. S., Chen, H.-J., Lin, C. C., Liaw, M.-L., Gao, Z.-M., & You, G.-N. (2013). Corpora processing and computational scaffolding for a web-based English learning environment: The CANDLE project. *CALICO Journal*, *24*(1), 77–95.

Sheehan, K. M., Kostin, I., Napolitano, D., & Flor, M. (2014). The *TextEvaluator*® tool. *The Elementary School Journal*, *115*, 184–209. https://doi.org/10.1086/678294

Compendium Study

The Consistency of *TOEIC*® Speaking Scores Across Ratings and Tasks

Jonathan Schmidgall

An important quality of test scores is their reliability or consistency across different aspects of the measurement procedure (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014). Researchers have observed that reliability is a prerequisite to validity (Haertel, 2006), and this "conventional wisdom" is made explicit in argument-based approaches to validity in which claims about score consistency underlie subsequent inferences about the interpretation of scores (e.g., Bachman & Palmer, 2010). In a validity argument, an overall claim that scores are consistent is dependent on a series of more detailed statements about specific aspects of consistency (e.g., agreement across raters). These statements (or assertions) are backed by evidence, often in the form of reliability coefficients. A test administrator's claim that test scores are consistent is supported or weakened by the extent to which the evidence supports the various assertions regarding score consistency.

An assessment use argument (Bachman & Palmer, 2010) is an argument-based approach to validity in which claims about the meaning and use of scores rest on the foundational claim that scores should be consistent. We have utilized this approach to specify claims about the measurement quality and use of *TOEIC*® test scores. In this report, I focus on the overall claim that TOEIC Speaking test scores are consistent. Table 1 summarizes the various assertions used to advance the claim that *TOEIC*® Speaking test scores are consistent as well as the evidence that supports each assertion.

Table 1

Underlying Assertions and Evidence to Support the Overall Claim That TOEIC Speaking Test Scores Are Consistent

Underlying claim or assertion	Published source of evidence
Administration procedures are followed consistently.	Hines (2010)
Scoring procedures are followed consistently.	Everson and Hines (2010); Hines (2010); Qu and Ricker-Pedley (2013)
Raters are trained, certified, calibrated, and monitored.	Everson and Hines (2010)
Scores are internally consistent.	Reasonably high internal consistency (Liao & Wei, 2010)
Scores from different raters (ratings) are consistent.	Reasonably high rater agreement rates (Liao & Wei, 2010; Qu & Ricker-Pedley, 2013); reasonably high generalizability of task scores (Liao & Wei, 2010)
Scores from different test forms and occasions of testing are consistent.	Liao and Qu (2010)

The TOEIC Speaking test requires test takers to demonstrate their English speaking ability across 11 speaking tasks that are scored by trained raters. This scoring process transforms a test taker's speaking performance into a scale score that is an indicator of his or her English speaking ability in the context of the workplace and everyday life. As stated in Table 1, the essential claim made about this scale score is that it is consistent (or reliable) across different aspects of the measurement procedure. This claim is supported by a series of assertions that are backed by evidence from the test design process and research. For example, one way to help ensure the consistency of scores is to follow administration and scoring procedures consistently. All TOEIC Speaking tests are administered on a computer that requires the use of headphones with a microphone, and test tasks are administered using a standardized format across occasions for all groups of test takers (Hines, 2010). A standardized procedure clearly specifies the steps involved in obtaining a scale score for each test taker and is carefully implemented and monitored to ensure compliance (Everson & Hines, 2010; Hines, 2010; Qu & Ricker-Pedley, 2013).

Raters themselves can be a source of either systematic bias or random error, and careful selection and training of qualified raters are critical (Brown, 2012; Engelhard, 2002). Everson and Hines (2010) described the path to becoming a TOEIC Speaking rater, which includes a number of steps designed to ensure consistent and high-quality ratings. Potential raters must (a) be qualified professionals (college graduates with experience teaching English as a second language/English as a foreign language); (b) complete a training course (which includes reviewing the purpose of each task type, sample and benchmark responses, and written explanations of scores for responses); and then (c) pass a certification test to demonstrate their rating proficiency. Applicants who pass the certification test qualify to work as raters but must subsequently pass a calibration test prior to every scoring session. The function of the calibration test is to ensure that raters maintain consistent standards for each new scoring session. Finally, each rater's performance is monitored by a scoring leader during the scoring session. All these policies and procedures are designed to promote the consistency and accuracy of rater scoring.

The use of highly trained raters and monitoring procedures helps to reduce the random error and bias introduced by human raters, but it is still essential to empirically quantify various aspects of reliability or score consistency (AERA, APA, & NCME, 2014). Score consistency can be quantified in a variety of ways. Prior research has found that TOEIC Speaking test scores are internally consistent (Liao & Wei, 2010), that scores from different raters are consistent (Liao & Wei, 2010; Qu & Ricker-Pedley, 2013), and that scores from different test forms and occasions of testing are consistent (Liao & Qu, 2010).

In an analysis of TOEIC Speaking pilot test data, Liao and Wei (2010) examined the interrater reliability and internal consistency of two test forms. Interrater reliability was evaluated by looking at rater agreement for each task and by using generalizability theory (G-theory) to estimate a generalizability coefficient for each task. Internal consistency for claim scores and weighted total scores was estimated using Cronbach's alpha (Cronbach, 1951) and stratified alpha (Rajaratnam, Cronbach, & Gleser, 1965), respectively. The analysis of rater agreement found acceptably high levels of rater agreement across most tasks, with exact agreement ranging from 50% to 81% and agreement within one score point ranging from 98% to 100%. In other words, very few test takers were given scores that were more than one score point apart. Generalizability coefficients for individual tasks ranged from .58 to .91 and were reasonably high for most. Estimates of internal consistency for claim scores were slightly lower for Claim 1 (.66–.68) and slightly higher for Claim 2 (.66–.80) and Claim 3 (.71–.74). The internal consistency of total scores ranged from .82 to .86, acceptable estimates according to traditional rules of thumb (Knapp & Mueller, 2010). Ultimately, because total scores are used to make interpretations about speaking ability, these estimates are the most critical.

Test takers complete a particular form of the TOEIC Speaking test on a particular occasion, but their scores should not be unduly influenced by the particular test form or occasion of testing. Liao and Qu (2010) examined the so-called alternate form test–retest reliability of TOEIC Speaking raw and scale scores across different occasions (e.g., 1–30 days, 31–60 days) and test forms. The test–retest reliability coefficients estimated across occasions of five different lengths ranged from .75 to .83, which supports the claim that scale scores are consistent across test forms and occasions.

To help stakeholders better understand the measurement facets of a TOEIC Speaking scale score, Figure 1 illustrates the design of the test.

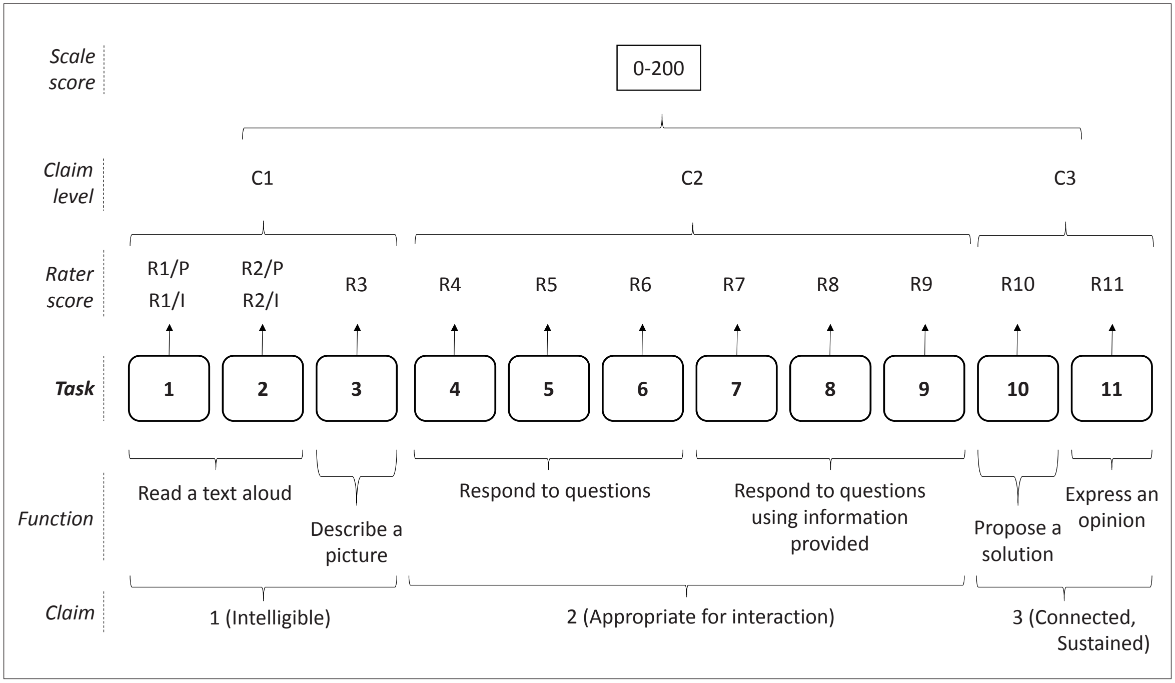


***Figure 1*. Design of the TOEIC Speaking test. R = rating; C = claim; P = pronunciation subscale; I = intonation subscale.**

Figure 1 should be viewed from the bottom to the top to understand how intended claims about a test taker's speaking ability informed the design of the test and the scale score that reflects these claims. The TOEIC Speaking test is designed to provide an interpretation about English speaking ability with respect to three claims: generating speech that is intelligible (Claim 1), appropriate for routine social and occupational interactions (Claim 2), and connected and sustained for typical workplace tasks (Claim 3; see Hines, 2010). As shown in Figure 1, 11 tasks were designed that targeted communicative functions that were representative of these claims. Each task is scored by a rater and assigned a single score, except for Tasks 1 and 2, which are given two scores: one for pronunciation and one for intonation. Different raters score each of the tasks, and a minimum of three different raters contribute to the final score of an individual test taker. As the figure indicates, variation in scores at the claim level reflects a test taker's performance on tasks that correspond to that claim as evaluated by different raters, for example, ratings from Tasks 1 to 3 reflect performance with respect to Claim 1. Finally, the scale score reflects a test taker's performance with respect to all three claims about speaking ability, which includes ratings of individual tasks that correspond to the claims. Ultimately, it is the scale score that is the basis for making an interpretation about someone's speaking proficiency, and so evidence of reliability or consistency is most critical for scale scores.

Research Questions

Prior research has produced evidence for the consistency of TOEIC speaking rater scores, claim-level information, and total scores (raw or scale). However, some of that evidence is based on an analysis of pilot study data (i.e., Liao & Wei, 2010), not on operational test scores that "count." To provide updated estimates of the consistency of TOEIC speaking scores across different phases of the scoring procedure, this study addresses the following research questions:

1. How consistent are ratings on individual tasks, as measured by generalizability and dependability coefficients?
2. How consistent is performance at the claim level across ratings, as measured by generalizability and dependability coefficients?
3. How consistent are scale scores across ratings, as measured by generalizability and dependability coefficients?

Methodology

Participants, Instrument, and Procedure

A previously administered and scored TOEIC Speaking test form was rescored in its entirety. The form and set of responses that were selected to be rescored were representative of TOEIC Speaking test form administrations in terms of sample size (N = 1,390), internal consistency reliability (α = .85), and scale score distribution (M = 15.71, SD = 3.58). Operational scoring conditions were maintained for the rescoring study (see Everson & Hines, 2010, for a description of the scoring procedure), and raters were not aware that scoring was being performed for a research study. The number of raters scoring each set of test-taker responses varied as per operational practice but was roughly comparable across the original and rescored samples.

Analysis

The framework of G-theory (Brennan, 2001) was used to identify sources of variances associated with test-taker ability (p) and facets of the measurement procedure, which may include ratings (r') and tasks (t). The ratings and tasks facets are considered random, as they are conceptualized as representative of the population from which they are drawn without exhaustively defining it.

Although most facets of measurement are self-explanatory, a brief overview of the ratings (*r′*) facet is needed. Each task for each person is scored by two raters, but the combination of raters differs across tasks. This approach of assigning multiple raters to each person is by design to minimize systematic bias that may arise from having the same rater or pair of raters score all of a person's responses. Thus this is a partially nested rating design in which each person is scored by multiple raters. Implementing this design using G-theory requires very large sample sizes depending on the number of rater combinations. This approach was impractical for this data set, where a large number of rater combinations was possible.

To provide a simplified approach to partially nested designs involving raters, researchers have proposed using a fully crossed design, $p \times t \times r'$, where r′ represents *ratings*, not *raters* (Lee, 2006; Lee & Kantor, 2005). With ratings as a facet, some researchers have argued that a main effect cannot be interpreted as differences between people who score (raters) but simply as differences between a first and second rating (Lee, 2006). However, researchers have shown that under certain conditions, this conceptual distinction may be negligible (Lin, 2013; Sawaki, 2017; Schmidgall, 2013). For example, Lin (2013) conducted a series of simulation studies under conditions that varied sample size, number of raters, and rating conditions; he concluded that when raters are relatively homogenous (i.e., when they have similar levels of experience), the rating method is sufficient for operational use, as it sacrifices little precision. In a related effort, Schmidgall (2013) examined a number of fully crossed pairs of raters within a larger data set and found negligible rater effects, which he used to partially justify a rating method. Sawaki (2017) performed multiple analyses in which a fully crossed rating method ($p \times t \times r'$) was used for an entire data set and separate analyses were conducted for each rater pair in the data set using the rater method ($p \times t \times r$); results were largely consistent across the analyses. The purpose of the present analysis was to estimate the amount of variation across ratings irrespective of which specific raters made these ratings, so the use of this fully crossed design ($p \times t \times r'$) is appropriate.

G-theory requires the researcher to specify the relationship between the object of measurement and facets. The following sections specify the G-study designs used to estimate generalizability coefficients and variance components associated with facets of measurement for scale scores, raw scores, claim scores, individual task scores and the five different rubrics used for the TOEIC Speaking test. G-studies were then performed using Edu-G software (Cardinet, Johnson, & Pini, 2010). Decision studies (D-studies; Brennan, 2001) were also performed to provide an estimate of reliability based on the operational scoring design of the TOEIC Speaking test, which typically uses a single rater to score each task. G-studies provide estimates that reflect the actual measurement design of a data set (e.g., 11 tasks and 2 raters), whereas D-studies provide estimates for different variations of the original measurement design (e.g., 11 tasks and 1 rater).

Individual Task Scores

Tasks 3 through 11 were assigned two ratings ($r' = 2$) using a holistic (i.e., one score) rubric, which is characterized by the design $p \times r'$. Tasks 1 and 2 use an analytic rubric in which test takers were assigned two ratings ($r' = 2$). In the scoring procedure, this results in four scores that equally contribute to the Claim 1 score. The measurement design of each of these four ratings is also $p \times r'$.

Claim-Level Performance

Performance at the claim level can be characterized by the G-study $p \times t \times r'$, in which a set of tasks are assigned two ratings ($r' = 2$). There are six tasks with Claim 2 ($t = 6$, Tasks 4–9) and two tasks with Claim 3 ($t = 2$, Tasks 10 and 11). There are three tasks associated with Claim 1 (Tasks 1–3), but four scores are produced for Tasks 1 and 2, because these tasks are scored separately for pronunciation and intonation. Thus, for the purpose of the analyses, there are five rated tasks associated with Claim 1 ($t = 5$). This approach may introduce the halo effect and underestimate variance components associated with tasks for Claim 1, a potential limitation of this analysis.

Scale Scores

Scale scores are based on linear combinations of raw scores and can be characterized using the fully crossed G-study design $p \times r'$. Because the research question examining scale scores is concerned with consistency across occasions of ratings, task was not specified as a facet of the G-study design.

Results

Consistency of Individual Task Scores

The percentage of total variance accounted for by each facet of measurement for each task or score is summarized in Table 2. The generalizability coefficient ($\hat{\rho}^2$) based on the G-study indicates the reliability of each individual task or score assigned two ratings (i.e., for this study), whereas the coefficient based on the D-study extrapolates the reliability estimate to operational scoring conditions in which one rater is typically used to score each task.

Table 2

Individual Task G-Study Percentage of Total Variance for Each Facet of Measurement, G-Study Generalizability Coefficient, and D-Study Generalizability Coefficient for Design With r′ = 1

Source	Task/score												
	1/P	1/I	2/P	2/I	3	4	5	6	7	8	9	10	11
Person (*p*)	54.2	42.1	40.0	39.3	52.2	74.5	56.7	59.8	84.9	76.0	64.1	71.9	72.4
Rating (*r*′)	0.1	0.6	0.1	0.5	0.7	1.2	0.2	1.2	0.1	0.3	0.0	0.0	0.0
pr′, *e*	45.7	57.2	59.9	60.2	47.1	24.3	43.1	39.1	15.0	23.6	35.8	28.1	27.5
$\hat{\rho}^2$ (G-study)	0.70	0.60	0.57	0.57	0.69	0.86	0.72	0.75	0.92	0.87	0.78	0.84	0.84
$\hat{\rho}^2$ (D-study)	0.54	0.42	0.40	0.40	0.53	0.75	0.57	0.61	0.85	0.76	0.64	0.72	0.72

Note. P = pronunciation subscale; I = intonation subscale; G-study = generalizability study; D-study = decision study.

As shown in Table 2, a person's ability (*p*) explains more of the variance in scores than the rating he or she received (*r*′) or the combination of unexplained error (*e*) and the particular rating for a particular person (*pr*′). For 3 of 11 scores (1/I, 2/P, 2/I), a greater percentage of variance in scores was accounted for by the combination of unexplained error and the particular rating provided for a particular person.

Generalizability coefficients ($\hat{\rho}^2$) based on G-studies were adequate (median = .75, range, .57–.92), particularly for Tasks 4 through 11. Generalizability coefficients based on D-studies that reflect the operational rating design (r′ = 1) were slightly lower (median = .61, range, .40–.85). One possible explanation for the higher proportion of variance explained by the combination of error and the particular rating provided for a particular person for Tasks 1 through 3—and, thus, lower generalizability coefficients—is restriction of range. The variance of ratings for Tasks 1 through 3 was comparatively lower than for other tasks, which may help explain the comparatively lower generalizability coefficients.

Consistency of Claim-Level Performance

The percentage of total variance accounted for by each facet of measurement for each claim is summarized in Table 3, along with generalizability coefficients based on G- and D-studies.

Table 3

Claim-Level G-Study Percentage of Total Variance for Each Facet of Measurement, G-Study Generalizability Coefficient, and D-Study Generalizability Coefficient for Design With r′ = 1

Source	Claim		
	1	2	3
Person (*p*)	31.8	32.3	54.5
Rating (*r′*)	0.0	0.3	0.0
Task (*t*)	0.3	6.1	1.5
pr′	2.0	0.0	1.0
pt	14.1	34.4	16.6
tr′	0.4	0.2	0.0
ptr′, e	51.5	26.7	26.4
$\hat{\rho}^2$ (G-study)	0.78	0.80	0.78
$\hat{\rho}^2$ (D-study)	0.68	0.76	0.71

Note. G-study = generalizability study; D-study = decision study.

As seen in Table 3, most of the variance in claim-level performance was explained by ability (*p*); the interaction between ability and task (*pt*); and the combination of unexplained error (*e*) and the three-way interaction between person, task, and rating (*ptr′*). The interaction between task and ability (*pt*) should be interpreted as the extent to which different test takers (*p*) performed differently on tasks associated with that claim (e.g., Tasks 4–9 for Claim 2); in other words, the rank ordering of persons varied across tasks within a claim. For Claim 2, a relatively large percentage of total variance (34.4%) was explained by person–task interaction or by differences in the rank ordering of performances across different tasks. Overall, though, differences in task difficulty did not account for a high percentage of total variance (6.1%).

Unexplained error and the three-way interaction between person, task, and rating (*ptr′, e*) accounted for a relatively large percentage of total variance (51.5%) in Claim 1 performance, although ability (*p*) still explained a sizable percentage of the total variance (31.8%). The opposite pattern was observed in Claim 3 performance, with ability accounting for the largest percentage of total variance in performance (54.5%). Generalizability coefficients based on G-studies for each of the claim scores were reasonably high (.78–.80), and those based on D-studies were lower (.68–.76).

Consistency of Scale Scores

The full results of the G-study for scale scores using the $p \times r$ design are summarized in Table 4.

Table 4

Scale Scores G-Study Results Including Percentage of Total Variance for Each Facet of Measurement in the Design p × r′

Source	*SS*	*df*	*MS*	Variance components	%
Person (*p*)	2,669,751.97	1,346	1,983.47	931.86	88.6
Rating (r′)	75.17	1	75.17	0.00	0.0
pr′, e	16,1174.3	1,346	119.74	119.74	11.4

Note. G-study = generalizability study.

As seen in Table 4, in a measurement design where scale scores are portioned into variance associated with ability (*p*) and different sets of ratings (*r′*), a high percentage of the variance in scores (88.6%) is explained by ability, minimal variance is attributable to differences between scores produced by different sets of ratings (*r′*), and a relatively smaller percentage (11.4%) is attributable to the combination of unexplained error (*e*) and differences in rank ordering of test takers across the sets of ratings (*pr′*). The generalizability coefficient associated with the G-study design was $\hat{\rho}^2 = .94$, and the D-study coefficient for the operational design using one set of ratings ($r' = 1$) was $\hat{\rho}^2 = .89$.

Discussion

This study analyzed the reliability or consistency of TOEIC speaking scores across different levels of the scoring procedure using the framework of G-theory. As expected, at the individual task level, the generalizability of scores under operational conditions varied greatly, from $\hat{\rho}^2 = .40$ to .85. The generalizability of claim-level performances based on their constituent tasks narrowed to the range of $\hat{\rho}^2 = .68$ (Claim 1) to $\hat{\rho}^2 = .76$ (Claim 2) under operational conditions, coefficients that are reasonably high but do not uniformly reflect a level of score consistency that would facilitate high-stakes decisions based on performance with respect to individual claims. The generalizability of scale scores across different sets of ratings was much higher ($\hat{\rho}^2 = .94$), and the level of consistency corresponding to operational conditions that use one set of ratings remained relatively high ($\hat{\rho}^2 = .89$)—certainly high enough according to traditional psychometric practice to justify using these scores for high-stakes decisions. Thus this study contributes backing to several of the warrants listed in Table 1 that support the claim that TOEIC speaking scores are consistent.

The results of the analysis of test-taker performance at the claim level provides support for the assertion that scores on different tasks within claims are internally consistent. The G-studies that examined the generalizability of claim scores found that very little of the total variance in scores could be attributed to the main effect of task controlling for rating (0.3%–6.1%), which suggests that the overall difficulty of tasks within a claim did not vary substantially. While this study did not conduct an analysis of internal consistency in the same manner as Liao and Wei (2010), the finding that the main effect of task at the claim level does not explain a sizable proportion of total variance is evidence to support the warrant. A larger percentage of variance was explained by the interaction between ability and task ($p \times t$), which suggests that some tasks were easier or more difficult for different test takers. This could be due to differences in the nature of the tasks performed or other contextual features of tasks; regardless, the finding that task effects were comparatively larger than rating effects is consistent with prior L2 speaking research (In'nami & Koizumi, 2016).

Analyses across all three levels of the scoring procedure suggested that differences between ratings had a minimal effect on scores, which supports the claim that scores from different ratings are consistent. Most importantly, the analysis of scale scores found that minimal variance was associated with differences between scale scores for the same test taker based on sets of ratings. The generalizability coefficient associated with operational rating conditions ($\hat{\rho}^2 = .89$) was similar in magnitude to previous research findings that used different test forms and different samples of test takers and raters to measure rater agreement using agreement rates (Liao & Wei, 2010; Qu & Ricker-Pedley, 2013) and G-theory (Liao & Wei, 2010). Although the methodological approach employed in these analyses (i.e., *rating* vs. *rater* as a facet) may lead to the underestimation of variance components associated with the rating facet, these variance component magnitudes were consistently negligible across tasks, at the claim level, and for scaled scores. Thus this series of G-studies using ratings collected under operational conditions helps strengthen the backing to support the warrant that scores from different ratings are consistent.

Thus the findings of this study provide evidence to strengthen the backing of claims about the consistency of TOEIC speaking scores. Most crucially, the generalizability and dependability of TOEIC speaking scale scores were found to be relatively high. While score consistency itself is not sufficient to facilitate high-quality decisions—score interpretations must be meaningful, impartial, generalizable, and relevant to those decisions (i.e., fair and valid)—this study contributes additional evidence that the psychometric basis for score interpretations is relatively strong.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice.* Oxford, England: Oxford University Press.

Brennan, R. L. (2001). *Generalizability theory.* New York, NY: Springer.

Brown, A. (2012). Interlocutor and rater training. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 413–425). New York, NY: Routledge.

Cardinet, J., Johnson, S., & Pini, G. (2010). *Applying generalizability theory using Edu-G*. New York, NY: Routledge.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334. https://doi.org/10.1007/BF02310555

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal & T. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 261–287). Mahwah, NJ: Lawrence Erlbaum Associates.

Everson, P., & Hines, S. (2010). How ETS scores the *TOEIC*® Speaking and Writing tests responses. In *The research foundation for the TOEIC® tests: A compendium of studies* (pp. 8.1–8.9). Princeton, NJ: Educational Testing Service.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). New York, NY: American Council on Education/Praeger.

Hines, S. (2010). Evidence-centered design: The *TOEIC*® Speaking and Writing tests. In *The research foundation for the TOEIC® tests: A compendium of studies* (pp. 7.1–7.31). Princeton, NJ: Educational Testing Service.

In'nami, Y., & Koizumi, R. (2016). Task and rater effects in L2 speaking and writing: A synthesis of generalizability studies. *Language Testing*, *33*, 341–366. https://doi.org/10.1177/0265532215587390

Knapp, T. R., & Mueller, R. O. (2010). Reliability and validity of instruments. In G. Hancock & R. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 337–341). New York, NY: Routledge.

Lee, Y.-W. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing*, *23*, 131–166. https://doi.org/10.1191/0265532206lt325oa

Lee, Y.-W., & Kantor, R. (2005). *Dependability of new ESL writing test scores: Evaluating prototype tasks and alternative rating schemes* (TOEFL Monograph No. MS-30). Princeton, NJ: Educational Testing Service.

Liao, C.-W., & Qu, Y. (2010). *Alternate test forms test-retest reliability for the TOEIC® Speaking and Writing tests*. In *The research foundation for the TOEIC® tests: A compendium of studies* (pp. 11.1–11.40). Princeton, NJ: Educational Testing Service.

Liao, C.-W., & Wei, Y. (2010). *Statistical analyses for the TOEIC® Speaking and Writing pilot study.* In *The research foundation for the TOEIC® tests: A compendium of studies* (pp. 9.1–9.25). Princeton, NJ: Educational Testing Service.

Lin, C.-K. (2013, June). *Handling sparse data in performance-based language assessments under generalizability theory framework*. Paper presented at the Language Testing Research Colloquium, Seoul, South Korea.

Qu, Y., & Ricker-Pedley, K. L. (2013). *Monitoring individual rater performance for TOEIC® Speaking and Writing tests*. In *The research foundation for the TOEIC® tests: A compendium of studies* (pp. 9.1–9.9). Princeton, NJ: Educational Testing Service.

Rajaratnam, N., Cronbach, L. J., & Gleser, G. C. (1965). Generalizability of stratified-parallel tests. *Psychometrika*, *30*, 39–56. https://doi.org/10.1007/BF02289746

Sawaki, Y. (2017, June). *Generalizability of content analytic rating scales for assessing university-level Japanese EFL learners' summarization performance*. Paper presented at Fundamental Considerations in Language Testing: An International Conference in Honor of Lyle F. Bachman, Salt Lake City, UT.

Schmidgall, J. E. (2013). *Evaluating the consistency of scores for a test of oral English within the framework of an argument for test use*. Unpublished manuscript.

Compendium Study

Evaluating the Stability of Test Score Means for the *TOEIC*® Speaking and Writing Tests

Yanxuan Qu, Yan Huo, and Eric Chan

ETS® TOEIC®

Assess to Progress.

The *TOEIC*® Speaking and Writing tests are designed to measure a person's ability to communicate in spoken and written English, respectively, in the context of daily life and the global workplace. The TOEIC Speaking test is composed of 11 constructed-response questions and takes approximately 20 minutes to complete. The TOEIC Writing test is composed of eight constructed-response questions and takes approximately 1 hour to complete. Scores are reported on a scale of 0 to 200 with increments of 10 for both the speaking and the writing tests. Test takers can choose to take either the TOEIC Speaking test or the TOEIC Writing test or both. Both tests are administered on fixed dates at secure, Internet-based test centers. The TOEIC Speaking test is currently administered much more frequently than the TOEIC Writing test.

For tests with frequent administrations, it is of paramount importance that all score means be monitored over time. Evaluating the stability of test score means over time is an important quality control procedure to prevent errors in score reporting and to maintain test score validity by ensuring that the meaning of test scores is preserved. For a test to be valid, test scores must reflect the knowledge, skills, and abilities that the test is intended to measure. Kolen and Brennan (2014, p. 333) mentioned that one useful quality control procedure is checking the consistency of score statistics (e.g., score means and score variances) over time. When score statistics fluctuate, it is important to investigate the potential causes (Allalouf, 2007; von Davier, 2012). For example, the fluctuation of score means may be due to changes in test takers' demographic factors, seasonality (the rise or fall of score means associated with specific times of the year), the result of operational errors (e.g., errors in test score reporting), or test security breaches.

To better observe and monitor the pattern and trend of the many score means across different forms or administrations, researchers at Educational Testing Service (ETS) have used ANOVA and harmonic regression to check score mean fluctuations over time (Lee & von Davier, 2013; von Davier, 2012). For example, Haberman, Guo, Liu, and Dorans (2008) used the ANOVA method (Howell, 2002) to examine the stability of *SAT*® Math and Reading score means over a 9-year period. They found that the scales of SAT Math and Reading reporting scores were stable and the fluctuations in SAT score means were mainly due to seasonal effect. The ANOVA method was particularly appropriate given that the SAT test has a small number of forms a year with fixed schedules. Harmonic regression (Bloomfield, 2000) is appropriate when there are frequent numbers of administrations across the whole year so that the seasonality pattern in score means can be modeled by a smooth sinusoidal term in a time series manner (Lee & Haberman, 2013). Lee and Haberman (2013) used the harmonic regression method to monitor score means across administrations for an international language test. They found that most of the fluctuations in the score means were explained by seasonal effect, yearly trend, and regional effect. Thus, the reporting scale for the language test was stable.

The purpose of this study was to evaluate the stability of the TOEIC Speaking and Writing test score means in an approximately 3-year period by using the harmonic regression method and the ANOVA method, respectively. Harmonic regression was chosen to monitor the stability of the TOEIC Speaking score means across forms due to the frequent administrations in Korea. Although the TOEIC Writing test was administered once a month in Korea, the number of forms in a year was sparse compared to those generated by the TOEIC Speaking test. Therefore, the ANOVA method was applied to check the stability of the TOEIC Writing score means across forms over time.

Data

The data for the TOEIC Speaking test were collected from Korean test takers who took only the TOEIC Speaking forms between February 1, 2014, and December 31, 2016. Background information was also available for each test taker (see the appendix for sample background questions). In total, 431 forms in 281 administrations were included in the analysis, with sample sizes ranging from 336 to 3,221 with an average sample size of 1,399. At the test administration level, sample sizes ranged from 336 to 11,022, with an average size of 2,135. The number of forms in each administration ranged from 1 to 5. Figure 1 shows the score means for all the 431 forms in a time series manner. The x-axis in Figure 1 is the number of days between each administration and January 1, 2014.

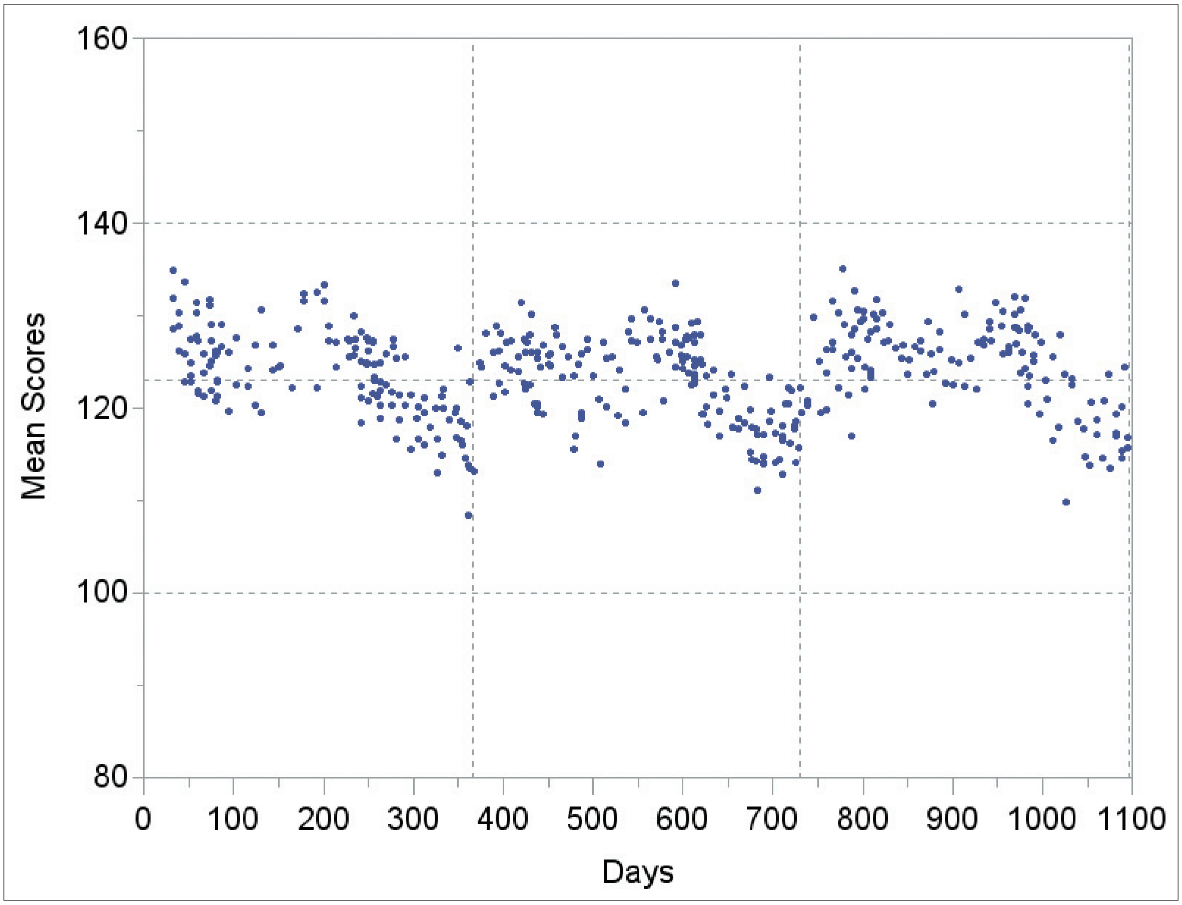


Figure 1. **Mean TOEIC Speaking scores for 431 forms over time.**

The data for the TOEIC Writing test contained writing scores and background information for Korean test takers who took forms with both speaking and writing sections between February 1, 2014, and December 31, 2016. We decided to use Korea-only data because (a) Korean test takers had the highest response rates to the background questionnaire and (b) Korean test takers regularly participated in the TOEIC Writing test (two forms each month on the same administration day) except after August 2016. In total, we had score data with background responses from 66 writing forms administered in Korea. Sample sizes per form ranged from 39 to 275, with an average sample size of 122.

Statistical Analyses

Harmonic regression is a linear regression model that contains sinusoidal terms. It can be used to check stability of score means because sinusoidal terms characterize seasonality in a time series fashion (Lee & Haberman, 2013). The harmonic regression models tried in this study are listed in Table 1.

Table 1

Models for TOEIC Speaking Mean Scores

Model	Equation
Model 0	$S_t = \mu + e_t$
Model 1	$S_t = \mu + \beta_1 y_{1t} + \beta_2 y_{2t} + e_t$
Model 2	$S_t = \mu + \beta_3 \cos(2\pi d_t / T_t) + \beta_4 \sin(2\pi d_t / T_t) + \beta_5 \sin(4\pi d_t / T_t) + \beta_6 \sin(8\pi d_t / T_t) + e_t$
Model 3	$S_t = \mu + \beta_1 y_{1t} + \beta_2 y_{2t} + \beta_3 \cos(2\pi d_t / T_t) + \beta_4 \sin(2\pi d_t / T_t) + \beta_5 \sin(4\pi d_t / T_t) + \beta_6 \sin(8\pi d_t / T_t) + e_t$
Model 4	$S_t = \mu + \beta_1 y_{1t} + \beta_2 y_{2t} + \beta_3 \cos(2\pi d_t / T_t) + \beta_4 \sin(2\pi d_t / T_t) + \beta_5 \sin(4\pi d_t / T_t) + \beta_6 \sin(8\pi d_t / T_t)$ $+\beta_7 f_{b3t} + \beta_8 f_{b6t} + \beta_9 f_{b8t} + \beta_{10} f_{b10t} + e_t$

In Table 1, where S_t is a mean score for Form t, Symbol d_t denotes the number of days elapsed since the beginning of 2014 and the time when Form t was administered. Symbol T_t is the total number of days in the year when Form t was administered. Year indicator $y_{1t} = 1$ indicates that Form t was administered in 2015, and $y_{2t} = 1$ indicates that Form t was administered in 2016. Score means in 2015 and 2016 were compared to score means in the baseline year, 2014.

Model 0 was a baseline model. Model 1 included the year effect terms. A significant year effect would indicate that the score means in year 2015 or 2016 were substantially higher (or lower) than in year 2014. Model 2 included sinusoidal terms for seasonal effect. Model 3 was a combined model with both year and seasonal effects. Model 4 is the complete model with year effect, seasonal effect and test takers' background effect. Out of 15 background questions (14 questions from the background questionnaire plus gender), Questions 3, 6, 8, and 10 showed relatively high correlations with test performance and were included in the model. Test takers' original responses to these four questions were recoded into four dummy variables. For example, Background Question 3 has four options regarding current job status. An original response of "1" (currently employed full-time) was recoded as "0," and the other three responses were recoded as "1." After the recoding for Model 4, f_{b3t} represented the fraction of test takers in each form who are not full-time employees, f_{b6t} represented the fraction of test takers in each form who have studied English for more than 10 years, f_{b8t} was the fraction of test takers in each form who used English more than 20% of the time in daily life, and f_{b10t} was the fraction of test takers in each form whose English did not always affect communication at work.

In our analyses, the year effect can be evaluated by comparing Model 1 to Model 0. The seasonality effect can be evaluated by comparing Model 2 to Model 0. The combined effect of year and seasonality can be evaluated by comparing Model 3 to Model 0. The combined effect of year, seasonality, and test takers' background can be evaluated by comparing Model 4 to Model 0. In our regression model, the seasonal terms and test takers' background variables are all indicators of test takers' performance on the test. As mentioned previously, seasonal factors are certain times of a year that are often related to business cycles within a year. Though related, seasonal factors and test takers' background factors are not necessarily identical.

To determine which harmonic regression model was the best model and which terms could be added or dropped from the regression model, we followed Lee and Haberman's (2013) example and checked if the decrease in root mean square error (RMSE) was at least 5% after including the terms and if the increase in *R* square and adjusted *R* square was noticeable. Different from *R* square, adjusted *R* square evaluates model fit by taking into account the number of predictors in a model. Additionally, the residual plot was checked for model fit. To determine whether a regression coefficient was significantly different from zero, the *p* value of each regression coefficient in the final model was compared to 0.05 divided by the total number of predictors.

In the ANOVA analyses for the TOEIC Writing test, the dependent variable was the score mean for each writing form. The independent variables included month, year, and their interaction. In the final ANOVA model (Model 1), *t* represents form (t = 1 to 66), $\alpha_{m(t)}$ shows the seasonal effect, $\beta_{y(t)}$ shows the year effect, and $\delta_{m(t)y(t)}$ shows the interaction between month and year. We also included background variables in the ANOVA analyses. In Model 2, $\beta_{2b3(t)} + \beta_{3b6(t)} + \beta_{4b8(t)} + \beta_{5b10(t)}$ represents the effect from the four recoded background questions. As for the analyses of TOEIC Speaking scores, these four background variables were recoded into dummy variables.

Model 1: $M_t = \mu + \alpha_{m(t)} + \beta_{y(t)} + \delta_{m(t)y(t)} + e$

Model 2: $M_t = \mu + \alpha_{m(t)} + \beta_{y(t)} + \delta_{m(t)y(t)} + \beta_{2b3(t)} + \beta_{3b6(t)} + \beta_{4b8(t)} + \beta_{5b10(t)} + e$

Results

Results for the *TOEIC*® Speaking Test

Table 2 shows that the *R* square value increased only slightly when the year indicator was added to the model (Model 1 vs. Model 0). However, adding the seasonal effect to the regression model increased *R* square significantly from 0.03 to 0.48. Adding test takers' background information to the regression model also increased the amount of explained variation and decreased the amount of unexplained error noticeably. From Model 3 to Model 4, *R* square increased from 0.50 to 0.56, by almost 12%, and RMSE decreased from 3.4206 to 3.2230, by 5.8%. Model 4 was chosen as the final model because no other indicators were found that could decrease RMSE by more than 5%. The fit of Model 4 was checked by a residual plot (Figure 2). Residuals are the difference between observed score means and predicted score means. All the residuals for the 431 forms appeared to be randomly and evenly distributed in Figure 2, indicating appropriate model fit.

Table 2

Model Fitting Results: Number of Predictors, Root Mean Square Errors (RMSE), R Square, and Adjusted R Square

Model	Number of predictors	RMSE	R^2	Adjusted R^2
Model 0	0	4.8134	0	0
Model 1	2	4.7533	0.0294	0.0248
Model 2	4	3.4803	0.4821	0.4772
Model 3	6	3.4206	0.5020	0.4950
Model 4	10	3.2230	0.5621	0.5517

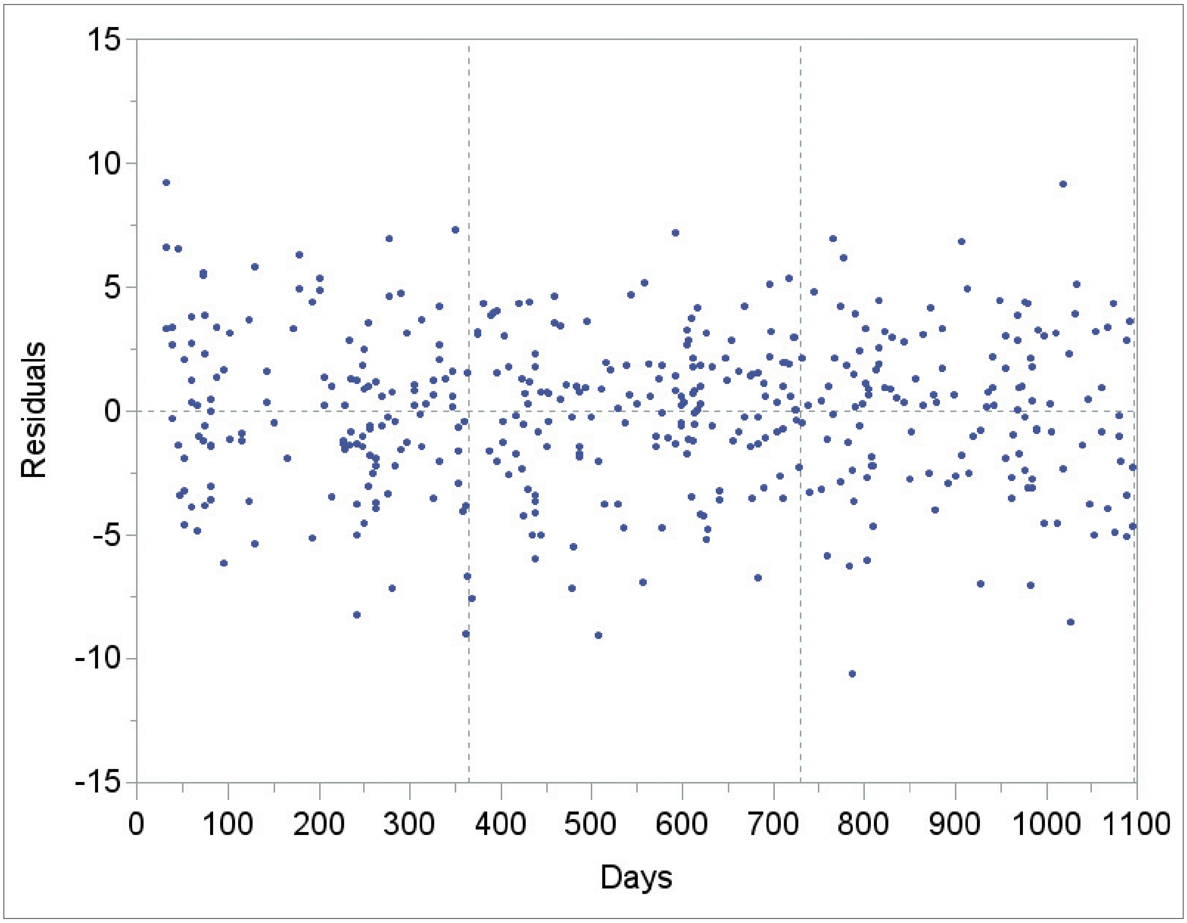


Figure 2. **Residuals for 431 TOEIC Speaking test forms over time.**

Table 3 shows the parameter estimates for the final model (Model 4). Since we conducted significance tests for 10 predictors simultaneously in the regression model, the *p* values of each predictor were compared to 0.05/10 = 0.005. A *p* value less than 0.005 indicates that the predictor is statistically significant. Therefore, the parameter estimates for the two year indicators, β_1 and β_2, were not significant, indicating very small score mean variations across 3 years. At least two seasonal parameters (β_3 and β_5) had a *p* value less than 0.005, which means the score means followed a strong periodical pattern over time. This periodical pattern can be seen clearly in Figure 1. Three background variables also had significant parameter estimates. These background variables were the fraction of test takers in each form who had studied English for more than 10 years, who used English more than 20% of the time in daily life, and whose English did not always affect communication at work.

Table 3

Estimated Parameters in Model 4 (The Final Model)

Model	Parameter	Estimate	SE	T statistic	*p* value
y_{1t}	β_1	-0.7304	0.3919	-1.8600	0.0631
y_{2t}	β_2	0.4475	0.4317	1.0400	0.3005
$\cos(2\pi d_t / T_t)$	β_3	-1.5334	0.3802	-4.0300	<.0001
$\sin(2\pi d_t / T_t)$	β_4	0.7782	0.2967	2.6200	0.0090
$\sin(4\pi d_t / T_t)$	β_5	2.1295	0.4622	4.6100	<.0001
$\sin(8\pi d_t / T_t)$	β_6	0.0353	0.2421	0.1500	0.8840
f_{b3t}	β_7	6.1553	2.9104	2.1100	0.0350
f_{b6t}	β_8	30.1261	8.1743	3.6900	0.0003
f_{b8t}	β_9	31.1538	9.1552	3.4000	0.0007
f_{b10t}	β_{10}	52.8702	12.3195	4.2900	<.0001

Figure 3 shows both observed (denoted by dots) and predicted (denoted by plus signs) mean scores for all 431 forms by the number of days elapsed between their administration date and January 1, 2014. A periodic pattern is clearly seen. In each year, the mean scores were relatively higher around the end of the first quarter and the third quarter but lower in the fourth quarter. This seasonal pattern is quite similar across the 3 years. There were multiple predicted values at an administration day in Figure 3 because there were multiple forms in one administration day.

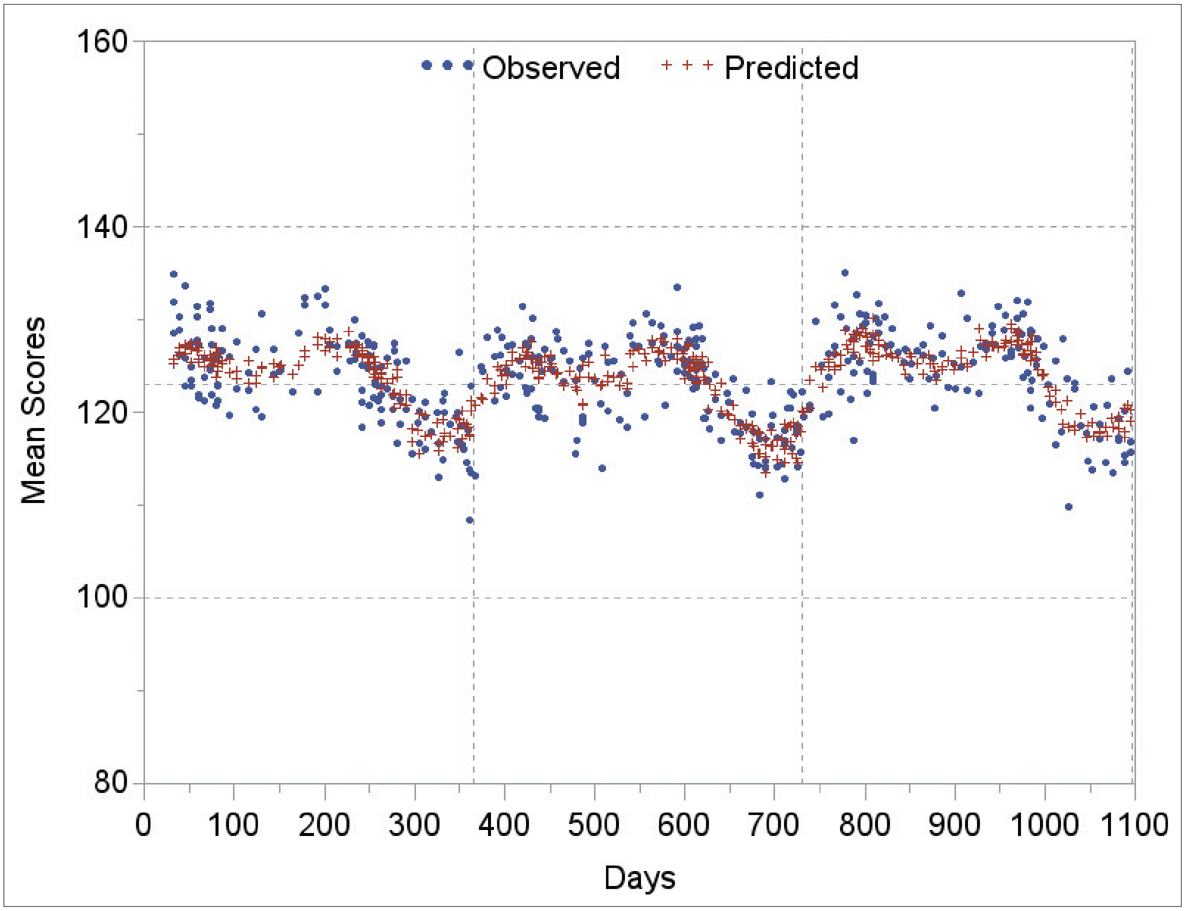


Figure 3. **Observed and predicted TOEIC Speaking score means for 431 forms over time.**

Results for the *TOEIC®* Writing Test

Tables 4 and 5 summarize the numbers of TOEIC Writing forms and the means and standard deviations of score means by month and by year. For example, Table 4 shows that there were six writing forms administered in July across 3 years. The average of these reported score means was 146.98 and the standard deviation was 3.83. Typically, two writing forms were administered each month, however, only one writing form was administered in September, October, November, and December during 2016. As a result, the total number of forms across 3 years was five instead of six in these 4 months in Table 4, and the total number of forms in 2016 was 20 instead of 24 in Table 5. Table 5 also shows that 22 instead of 24 writing forms were administered in 2014 in Korea because our data did not have forms administered in January 2014.

Table 4

Summary Statistics of TOEIC Writing Score Means by Month of Administrations

Month	*N*	Mean	SD	Min	Max
January	4	151.61	3.07	148.26	155.52
February	6	153.51	5.81	144.36	161.03
March	6	149.22	5.86	139.87	157.37
April	6	146.50	3.49	142.24	152.20
May	6	142.22	4.83	134.64	147.92
June	6	143.12	1.76	140.32	145.00
July	6	146.98	3.83	141.15	151.55
August	6	148.27	4.15	142.35	153.82
September	5	147.93	2.57	143.63	149.90
October	5	142.11	4.08	135.26	145.35
November	5	143.09	5.62	135.68	147.53
December	5	140.99	4.98	136.50	149.15
Overall	66	146.30	5.54	134.64	161.03

Table 5

Summary Statistics of TOEIC Writing Score Means by Year of Administrations

Month	*N*	Mean	SD	Min	Max
2014	22	144.49	4.98	135.26	153.25
2015	24	145.57	5.85	134.64	158.25
2016	20	149.18	4.80	139.87	161.03
Overall	66	146.30	5.54	134.64	161.03

Unlike the results for speaking, adding the four background variables did not reduce RMSE or increase *R* square substantively. In fact, RMSE increased only from 3.52 to 3.7, and *R* square increased from 0.81 to 0.82. None of the parameter estimates for the four background variables was statistically significant. Therefore, the final model did not include any background variables. Table 6 shows the ANOVA results for the final model (Model 1).

Table 6

ANOVA Results for TOEIC Writing Scaled Scores (Based on Individual Form Level: N = 66, Total R^2 = 0.81)

Component	df	Sum of squares	Mean square	*F*	*p*	R^2
Month	11	808.97	73.54	5.95	<.0001	0.41
Year	2	173.90	86.95	7.04	0.003	0.09
Interaction	21	492.10	23.43	1.9	0.05	0.25
Residual	31	383.03	12.36			0.19

Table 6 indicates that month was the major variable accounting for the score mean variations. It explained 41% of the total mean score variance. Figure 4 indicates that the average score means (connected by solid lines in Figure 4, with circles representing the score means for individual forms) tended to be higher in the first and third quarters than in the second and fourth quarters. This pattern bears some resemblance to the periodic pattern observed in the speaking results.

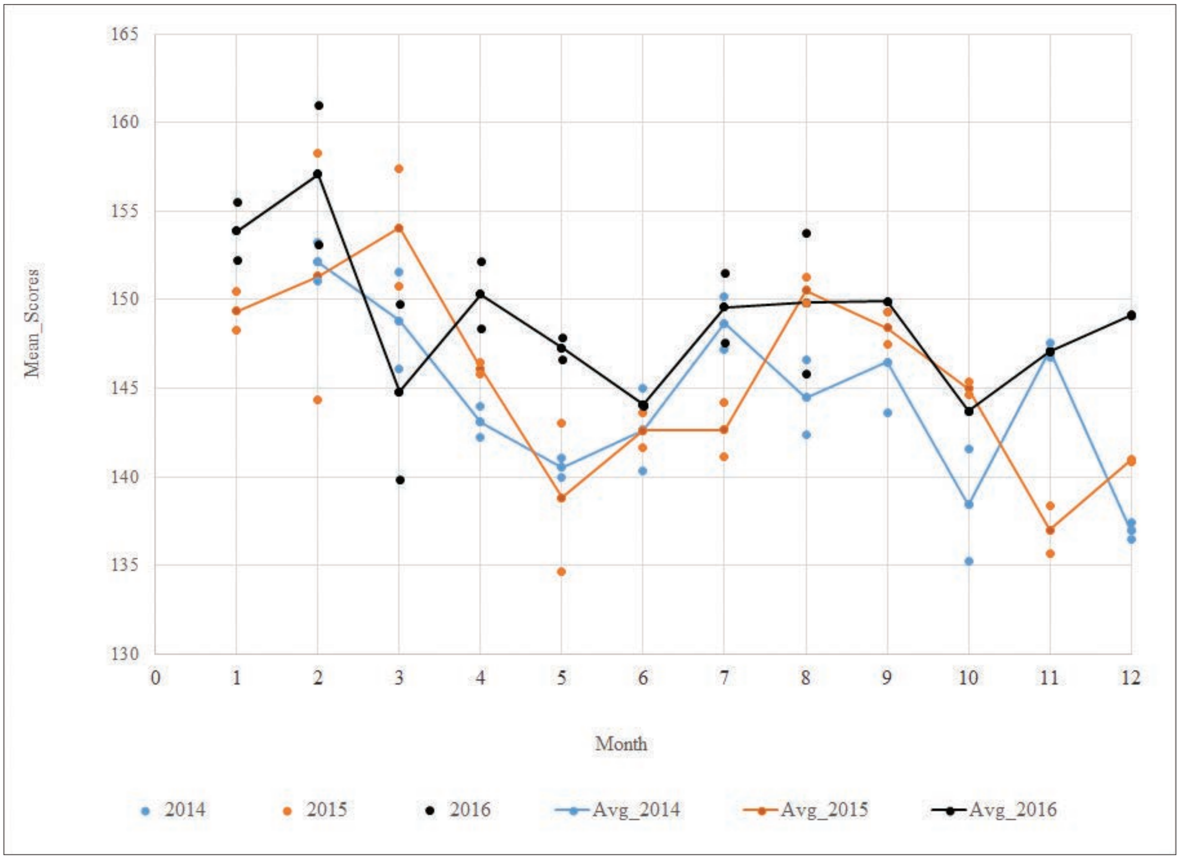


***Figure 4.* Writing score means by year and by month.**

Table 6 also shows that the year effect was significant for writing, and so was the interaction effect. Table 5 shows that the average score means in 2014 and 2015 were similar to each other, whereas the average score mean of 2016 was higher than the other 2 years, especially in February, May, and December (as seen in Figure 4). However, the score means for September 2016 through December

2016 were only based on one form. More data cumulated over a longer time period would be needed to better understand if there is indeed a year effect or an interaction effect between year and month of the administrations for writing.

Concluding Remarks

The results based on harmonic regression for speaking showed significant seasonal effect and demographic effect. In all 3 years, the TOEIC Speaking score means appeared to be higher around March and August and lower in the other months. Given the large number of forms, the large number of administrations for the TOEIC Speaking test each year, and the sample size per form, the regression model explained a reasonably high proportion (56%) of total mean score variation across forms. A large portion of the observed fluctuation in test score means was explained by the fact that test takers differ systematically in their ability and demographic characteristics according to the time of the year they choose to take the test (seasonal effects). It can be argued, therefore, that the scale of the TOEIC Speaking test is appropriately stable, after accounting for seasonal and demographic differences in test takers' overall speaking ability.

The results for writing also showed a significant seasonal effect. Within each year, the score means appeared to fluctuate in a pattern similar to the one detected for the TOEIC Speaking test. Overall, the ANOVA model explained 81% of the total variation in writing score means. The scale of the TOEIC Writing test is also stable.

References

Allalouf, A. (2007). An NCME instructional module on quality control procedures in the scoring, equating, and reporting of test scores. *Educational Measurement: Issues and Practice*, *26* (1), 36–46. https://doi.org/10.1111/j.1745-3992.2007.00087.x

Bloomfield, P. (2000). *Fourier analysis of time series: An introduction* (2nd ed.). New York, NY: Wiley. https://doi.org/10.1002/0471722235

Haberman, S. J., Guo, H., Liu, J., & Dorans, N. J. (2008). *Consistency of SAT I: Reasoning test score conversions* (Research Report No. RR-08-67). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2008.tb02153.x

Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: Duxbury/Thomson Learning.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York, NY: Springer. https://doi.org/10.1007/978-1-4939-0317-7

Lee, Y.-H., & Haberman, S. J. (2013). Harmonic regression and scale stability. *Psychometrika*, *78*, 815–829. https://doi.org/10.1007/s11336-013-9337-1

Lee, Y.-H., & von Davier, A. A. (2013). Monitoring scale scores over time via quality control charts, model-based approaches, and time series techniques. *Psychometrika*, *78*, 557–575. https://doi.org/10.1007/s11336-013-9317-5

von Davier, A. A. (2012). *The use of quality control and data mining techniques for monitoring scaled scores: An overview* (Research Report No. RR-12-20). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2012.tb02302.x

Appendix

TOEIC® Speaking and Writing Background Questionnaire

Read the choices below each question and select the one best answer. Fill in only one answer for each question.

Section I. Your Educational and/or Work-Related Background

1. Choose either the level of education in which you are currently enrolled or the highest level that you have completed.

   01. Elementary school (primary school)

   02. General secondary school (junior high school)

   03. Secondary school for university entrance qualification or equivalent (high school)

   04. Vocational/technical high school

   05. Vocational/technical school after high school

   06. Community/junior college (for associate degree)

   07. Undergraduate college or university (for bachelor's degree)

   08. Graduate or professional school (for master's or doctoral degree)

   09. Language institution

2. Choose the major that you are currently enrolled in or the major of your highest degree.

   (The majors shown in parentheses are examples only.)

   01. Liberal arts (education, fine arts, languages, literature, music, psychology)

   02. Social studies/law (international studies, law studies, political science, sociology)

   03. Accounting/business/economics/finance/marketing/trading

04. Sciences (agriculture, computer science, mathematics, physics, statistics)

05. Health (medicine, nursing, pharmacy, public health)

06. Engineering/architecture

07. Other/none

3. Which of the following best describes your current status?

01. I am employed full-time (including self-employed).

02. I am employed part-time and/or study part-time.

03. I am not employed. (Skip to Question #6.)

04. I am a full-time student. (Skip to Question #6.)

4. If you are currently employed, which industry best describes that of your current employer?

01. Agriculture/fishing/forestry/mining

02. Construction/building design

03. Manufacturing—food

04. Manufacturing—pharmaceuticals

05. Manufacturing—chemicals

06. Manufacturing—fabric/paper

07. Manufacturing—oil/petroleum/rubber

08. Manufacturing—steel/other metals

09. Manufacturing—Machinery/fine machinery

10. Manufacturing—electronic

11. Manufacturing—vehicles (includes manufacturing of all modes of transportation)

12. Manufacturing—cement/glass

13. Manufacturing—clothing

14. Manufacturing—other

15. Service—education (high school equivalent or below)

16. Service—education (college equivalent or above, assessment, research)

17. Service—court/legislative/municipal/prefecture

18. Service—foreign affairs

19. Service—armed forces

20. Service—health/hospital/medical research

21. Service—hotel/recreation/restaurant/travel

22. Service—other

23. Public utilities production/management (electricity/water supply)

24. Broadcasting/mass media

25. Telecommunication

26. Retail/wholesale

27. Trading

28. Accounting/banking/finance/security

29. Insurance

30. Real estate

31. Transportation

32. Other

5. If you are currently employed, which of the following best describes the type of job you do?

   (The jobs shown in parentheses are examples only.)

   01. Management (executive, manager, director)

   02. Scientific/technical professionals (engineer, mathematician, programmer, researcher, scientist)

   03. Teaching/training

   04. Professional specialist (accountant, broker, financial specialist, lawyer)

   05. Technician (carpenter, electrician, equipment operator, plumber)

   06. Marketing/sales (foreign exchange broker, marketing analyst, real estate agent, sales representative, travel agent)

   07. Clerical/administrative (office staff member, receptionist, secretary)

08. Services (customer service representative, human resources representative, hotel staff member, public relations representative)

09. Other

Section II. Your English-Language Experience

6. How many years have you spent studying English?

01. Less than or equal to 4 years

02. More than 4 years but less than or equal to 6 years

03. More than 6 years but less than or equal to 10 years

04. More than 10 years

7. Which of the following language skills are/were most emphasized?

01. Listening

02. Reading

03. Speaking

04. Writing

05. Listening and speaking

06. Reading and writing

07. Listening, reading, speaking, and writing

8. How much time must you use English in your daily life?

01. None at all

02. 1 to 10%

03. 11 to 20%

04. 21 to 50%

05. 51 to 100%

9. Which of the following English-language skills do you use most often?

01. Listening

02. Reading

03. Speaking

04. Writing

05. Listening and speaking

06. Reading and writing

07. Listening, reading, speaking, and writing

10. How often has difficulty with English affected your ability to communicate?

01. Almost never

02. Seldom

03. Sometimes

04. Frequently

05. Almost always

11. Have you ever lived in a country in which English is the main spoken language?

01. No (Skip to Question #13.)

02. Yes, for less than 6 months

03. Yes, for 6 to 12 months

04. Yes, for more than 1 but less than or equal to 2 years

05. Yes, for more than 2 years

12. What was your main purpose for living in a country in which English is the main spoken language?

01. To study (in other than an English-language program)

02. To participate in an English-language program

03. To travel (not work related)

04. To work

05. Other

Section III. Your Experience in Taking the *TOEIC*® Test

13. Before today, how many times have you taken the TOEIC Speaking and Writing test?

    01. Never

    02. Once

    03. Twice

    04. Three times or more

14. What is your main purpose for taking today's TOEIC Speaking and Writing test?

    01. For a job application

    02. For promotion

    03. To assess the effectiveness of an English-language program

    04. To assess future learning needs

    05. To graduate from a course of study

    06. To apply for visa

Compendium Study

Monitoring Score Change Patterns to Support *TOEIC*® Listening and Reading Test Quality

Youhua Wei and Albert Low

ETS® TOEIC®

Assess to Progress.

Quality control in educational measurement should be conducted systematically not only within individual administrations but also across administrations over time (von Davier, 2012). Across-administration test quality control may include the evaluation of the fluctuation of score summary statistics, population composition and background changes, test content evolution and difficulty shift, equating errors and scale drift, and the stability of psychometric properties such as reliability and validity. Various methods and procedures have been proposed for this purpose, such as time series analysis (Li, Li, & von Davier, 2011), harmonic regression (Lee & Haberman, 2013), linear mixed effects modeling (Lee, Liu, & von Davier, 2013), Shewhart control charts (see a brief description in von Davier, 2012), hidden Markov modeling (Lee & von Davier, 2013), and multilevel analysis (Wei, 2013; Wei & Qu, 2014).

In large-scale programs of tests that aid in making high-stakes decisions, some test takers take a test more than once, and they have been called *repeaters*. Repeater studies have been conducted to examine repeaters' score changes and explore their score growth patterns across administrations. Most studies (e.g., Kingston & Turner, 1984; Wei & Morgan, 2016; Yang, Bontya, & Moses, 2011; Zhang, 2008) have evaluated repeaters' score changes between two adjacent administrations; few studies (e.g., Nathan & Camara, 1998; Wilson, 1987) have explored repeaters' longitudinal score change patterns over multiple repetitions. On the basis of those studies, the average score changes tended to increase with the number of times tested, but the score changes were related to a number of factors, such as the number of repetitions, the interval between repetitions, initial scores, educational level, and gender.

The analyses based on the data collected at only two time points are often inadequate for investigating score growth. The longitudinal data tend to reduce quickly with the requirement of more retakes, so the results tend to be unstable. Typically, test takers may repeat a test a different number of times and at different points in time. Therefore repeaters' data tend to be unfixed and unbalanced, and advanced methods need to be used to take full advantage of the available information to provide a complete picture of repeaters' score growth trajectories, especially over a long time.

Multilevel growth modeling (e.g., Raudenbush & Bryk, 2002; Singer & Willett, 2003) is a flexible method that allows us to explore repeaters' longitudinal score change patterns when the number and spacing of time points vary across individual examinees. As for most testing programs, repeater data from the *TOEIC*® tests tend to be unbalanced and unfixed. That is, during any given period of time, test takers tend to retake the test different numbers of times and at variable intervals between repetitions. Furthermore, time intervals tend to vary both within persons and between persons. Multilevel growth modeling can handle different data sets and fully use all repeaters' information to provide a more complete picture of repeaters' growth trajectories.

The repeaters' score changes over time can be used for quality control of test performance from different perspectives. First, because repeaters are the same examinees taking a test over time, their score changes can be used to evaluate the stability of test performance across administrations. A lack of stability can signal one or more of the issues mentioned at the beginning of the report. Second, repeaters' score changes provide empirical data to evaluate the score reliability by examining test score consistency across forms, across administrations, or over time based on the same examinees, especially when the intervals between repetitions are short. Third, repeaters' score changes provide operational data to evaluate score validity by comparing the growth patterns in a testing program with patterns found in other related testing programs or with related learning theories. Fourth, and finally, a testing program can use repeaters' growth patterns to predict and monitor their performance in future administrations.

The study reported here is based on repeaters' data from the *TOEIC*® Listening and Reading test over a 4-year period from 2010 to 2014. Multilevel growth modeling was used to explore repeaters' test score change patterns. The growth modeling results were used for the quality control of test performance by evaluating the stability, reliability, and validity of test scores and the potential to monitor test performance across administrations.

Methodology

Data

The data were collected from the TOEIC Listening and Reading test in a country where English is a foreign language. The test has two sections, *Listening* and *Reading*, each consisting of 100 multiple-choice items. For each section, the raw scores range from 0 to 100 and the scale scores from 5 to 495 by increments of 5. Equating is conducted so that scale scores from different administrations or test forms are on the same scale. Therefore the longitudinal scale score data of the same test takers across administrations can be used to explore their score growth trajectories (Castellano & Ho, 2013). At each administration, a questionnaire is used to collect information on test takers' general background, English learning experience, and test-taking experience.

The test is offered in strictly scheduled monthly administrations in the country, with each administration using one unique test form. The data used in this study include *Listening* and *Reading* scale scores and background information of 19,855 test takers who had taken the test six times in 68 administrations in 4 years from 2010 to 2014. The spacing of test taking (in terms of months) varied across test takers. Table 1 shows the distribution of test takers based on the time gaps between adjacent times tested within the 4 years (e.g., between the first and second times and between the second and third times). The table shows that the number of repeaters tended to decrease when the time gap between adjacent repetitions became longer.

Table 1

Distribution of Repeaters Based on Time Gaps Between Adjacent Times Tested

Time gap (month)	First–second	Second–third	Third–fourth	Fourth–fifth	Fifth–sixth
1	3,865	4,301	4,615	4,593	4,393
2	4,536	4,986	5,044	4,909	4,356
3	1,901	2,025	1,921	1,970	2,046
4	2,759	2,420	2,330	2,108	1,886
5	1,031	833	839	842	827
6	1,288	1,314	1,370	1,376	1,396
7	486	570	502	518	529
8	816	795	735	745	711
9	265	308	331	345	378
10	446	484	489	485	569
11	268	258	219	265	307
12	752	486	469	516	764
13	173	129	118	121	180
14	239	191	164	207	275
15	95	92	75	87	130
16	177	126	162	165	180
17	119	65	56	89	115
18	135	92	81	116	174
19	56	42	29	47	75
20	101	72	65	64	103
21	29	37	36	36	68
22	71	45	45	50	88
23	33	25	26	27	54
24	61	42	35	61	66
25	26	18	22	16	27
26	28	25	20	29	42
27	13	9	6	8	22
28	24	19	19	18	34
29	14	11	6	14	16
30	10	11	5	6	14
31	4	6	6	4	5
32	10	4	5	7	5
33	2	5	0	1	3
34	10	3	2	3	3
35	3	4	3	3	2
36	4	2	3	4	2
37	1	0	1	0	2
38	2	0	1	0	3
39	1	0	0	0	2
40	1	0	0	0	3

Note. *N* = 19,855.

Data Preparation

As in a typical multilevel growth analysis (e.g., Raudenbush & Bryk, 2002; Singer & Willett, 2003), the repeated measures of each test taker in this study were considered as nested within the person. Therefore the repeaters' data had two levels, with repeated measures, including the scale scores and time-varying background, in multiple test-taking months as the Level 1 variables and unchanged person-level characteristics as the Level 2 variables.

At Level 1, the test taker's scale score in each of the multiple administrations was the dependent variable and the administration time was the independent variable. The *Listening* scale scores ranged from 105 to 495, with a mean of 334 and a standard deviation of 74. The *Reading* scale scores ranged from 85 to 495, with a mean of 279 and a standard deviation of 82. The administration time was defined as the amount of time in months that had elapsed from the first time a test taker took the test in the 4 years. The starting month and the spacing of the six test-taking months varied across test takers. For example, if one test taker took the test in January, March, May, August, November, and December in the first year, his or her administration times would be 0, 2, 4, 7, 10, and 11. If another test taker took the test in September and December in the first year, and then took the test in January, May, July, and October in the second year, his or her administration times would be 0, 3, 4, 8, 10, and 13. Therefore the possible administration times ranged from 0 to 47 in months in the 4 years.

Two types of test takers' background information tended to change across the six times of test taking and had close relations with test takers' scale scores. The first one was the test takers' *occupation status*, which was based on the survey question "Which of the following best describes your current status" (see Table 2 for the options); the second one was the test takers' *daily English use time,* which was based on the question "How much time must you use English in your daily life?" (see Table 2 for the options). These two background variables were selected and used as the time-varying independent variables or covariates at Level 1.

Table 2

Variables and Codes at Levels 1 and 2

Data level	Variable	Options	Code	Subgroup percentage	Variable name
1	Current occupation	Full-time employed	(0, 0, 0, 0)	54.95	
		Missing information	(1, 0, 0, 0)	2.19	EMPMIS
		Part-time employed	(0, 1, 0, 0)	3.54	EMPPAR
		Unemployed	(0, 0, 1, 0)	3.77	UNEMP
		Full-time student	(0, 0, 0, 1)	35.56	STUDT
	Daily English use time	None at all	1	26.37	ENGUSE
		1%–10% and missing information	2	47.54	
		11%–20%	3	14.59	
		21%–50%	4	9.37	
		51%–100%	5	2.13	
	Time	*M* = 11.55, *SD* = 10.90, min. = 0, max. = 47			TIME
	Listening score	*M* = 333.75, *SD* = 74.23, min. = 105, max. = 495			LISTEN
	Reading score	*M* = 279.30, *SD* = 82.35, min. = 85.00, max. = 495			READ
2	Education level	Vocational/technical high school	(0, 0, 0, 0, 0, 0, 0, 0)	1.56	
		Missing information/ primary school	(1, 0, 0, 0, 0, 0, 0, 0)	1.77	EDUMIS
		Junior high school	(0, 1, 0, 0, 0, 0, 0, 0)	0.11	SECOND1
		High school	(0, 0, 1, 0, 0, 0, 0, 0)	3.62	SECOND2
		Vocational/technical school	(0, 0, 0, 1, 0, 0, 0, 0)	2.08	VOTECH
		Community college	(0, 0, 0, 0, 1, 0, 0, 0)	3.00	COMMUN
		Undergraduate	(0, 0, 0, 0, 0, 1, 0, 0)	70.03	UNDERG
		Graduate	(0, 0, 0, 0, 0, 0, 1, 0)	17.63	GRADUA
		Language institute	(0, 0, 0, 0, 0, 0, 0, 1)	0.20	LANGUA
	Gender	Male	0	65.19	
		Female	1	34.81	GENDER
	Test-taking experience	Tested at least once before	0	77.65	
		Never tested before	1	22.35	PREEXP

Test takers' *gender* information remained the same, and their *educational levels* tended to be the same or very similar across the six times of test taking (see Table 2 for the specific educational levels). The examinees' *test-taking experience* before the first time tested in the 4 years of data collection period was another type of background information. It was based on test takers' responses to the question "Before today, how many times have you taken the test?" at the first time tested in the 4 years (see Table 2 for the options). These three types of background information (i.e., gender, educational level, and test-taking experience) also had close relations with test takers' scale scores, and they were used as the unchanged person-level characteristics at Level 2.

Among the five background variables, occupation status, gender, educational level, and test-taking experience were categorical, so dummy coding was conducted for these four background variables. The daily English use time was ordinal, so a Likert scale was used to quantify its values. Table 2 shows the background variables and their codes at Levels 1 and 2. The coding was mainly based on the survey questions' original response options. The test performance patterns of subgroups based on response options were also taken into account for the coding. For example, based on the survey question about test takers' occupation status, the subgroup with missing information tended to have consistent performance compared with other subgroups, so this subgroup was not removed from the sample but rather was coded as a separate subgroup. On the basis of the survey question about test takers' daily English use time, the subgroup who chose "1%–10%" and the subgroup who did not choose any option tended to have similar test performance, so these two subgroups were combined and coded as one subgroup for analysis. For the convenience of interpretation, the subgroups with lowest test performance were coded as the reference groups in most cases. For example, the subgroup of full-time employed for the occupation status background was coded as the reference group; the subgroup choosing the option of vocational/technical high school for the educational level was coded as the reference group.

Preliminary Analyses

Some descriptive analyses were conducted to explore the nature and idiosyncrasies of the repeaters' growth trajectories before the multilevel growth modeling was used. On the basis of the observation of some randomly selected repeaters' scores across repetitions, the scores tended to increase over time, but the rate of increase slowed gradually, with a substantial variation across individuals. Although the starting month and the spacing of the six test-taking months varied across test takers, each repeater had scores at six time points in the 0–47 administration months over 4 years. To show the score change trend at the group level over time, we computed repeaters' scale score means at each of the 48 time points based on the data available at each administration time, and then plotted the score means over time (i.e., months). Figure 1 shows the plots of the observed score means for *Listening* and *Reading*. The plots show that repeaters' scale score means tended to increase over time, but the increasing rate tended to decrease gradually. Therefore the preliminary analyses based on both individual and group data suggest a nonlinear growth model for repeaters' score changes for both *Listening* and *Reading*. The relationships between test takers' scale scores and their background variables were also explored in the preliminary analyses.

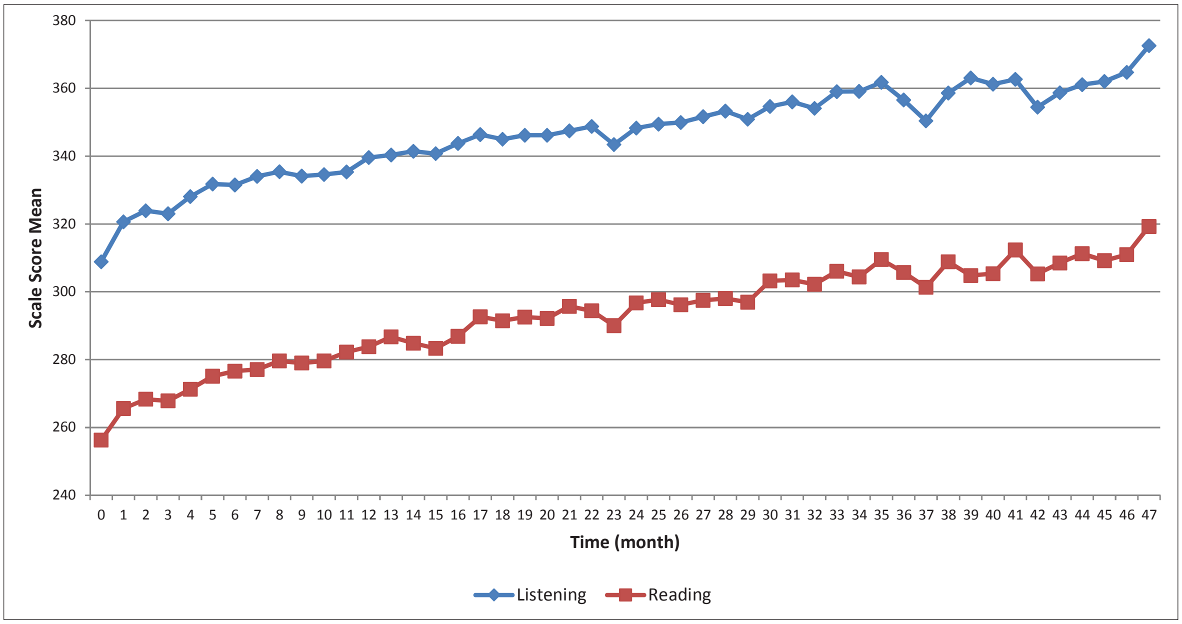


Figure 1**. Listening and Reading observed score means over time (month).**

Multilevel Growth Modeling

Multilevel growth modeling was used to explore the repeaters' score change patterns, with examinees' repeated measures at Level 1 and person-level characteristics at Level 2. On the basis of the preliminary analyses of repeaters' score changes and the relations between examinees' scores and their background information, different models were explored and results were evaluated in terms of model fit, growth parameter estimation, variance estimation, and test performance prediction.

Following the suggestions by Raudenbush and Bryk (2002) and Singer and Willett (2003) on model building, the analyses started with simple growth models and then used the "step-up" strategy to include more growth parameters and background variables based on promising submodels. Specifically, four types of models were used in this study (see the appendix for the statistical specifications of the four models).

Unconditional Means Model

As the simplest model, the unconditional means model does not include any predictors and does not describe the score change over time. However, this model partitions the total score variation into the within-person variation at Level 1 and the between-person variation at Level 2. It helps determine whether there is sufficient variation to warrant further analysis at each level.

Linear Growth Model

This model includes the linear TIME predictor in the Level 1 model. Assuming a constant rate of score change over time, this model estimates the repeaters' average score change per month. It also estimates the between-person variation in the rate of score change.

Quadratic Growth Model

Assuming the rate of score change is not constant over time, this model includes both the linear TIME predictor and the quadratic TIME^2 predictor in the Level 1 model. The linear growth parameter estimates the instantaneous or initial rate of change. The quadratic parameter estimates the acceleration in the growth trajectory.

Conditional Quadratic Growth Model

This model includes test takers' background variables in the quadratic growth model, so that the impact of examinees' background on their score growth trajectory can be examined.

The analyses first explored the repeaters' growth trajectories by evaluating different growth models without including any background variables. When necessary, polynomial models with higher degrees (e.g., cubic growth model by including the cubic TIME^3 predictor) were explored and examined. After the most appropriate growth model was identified, the time-varying background variables were added in the Level 1 model, and the person-level background variables were added in the Level 2 models, so their impacts on examinees' test scores and growth parameters could be examined.

Model Validation

The data of the other 1,861 examinees who had taken the test 12 times in the same 4 years were used to validate the repeaters' growth models, which were selected based on the original data of 19,855 examinees. The models, parameter estimates, and the impacts of background variables were compared to evaluate the validity of the models.

Results

In this section, we first summarize the results from the unconditional means model, which can provide baseline information for further analysis. Then we present the modeling results based on the linear, quadratic, and cubic growth models, followed by the results from the conditional growth model, which includes background variables in both Level 1 and Level 2 models. We close the section by evaluating the validity of the model identified from the study.

Unconditional Means Model

Listening

On the basis of the unconditional means model for *Listening* scores (see Table 3), the estimated grand mean of all repeaters' scores across the six administration times in the study was 333.75. The Level 1 variance estimate was 1,242.78, and the Level 2 variance estimate was 4,267.24, which indicates that much more score variation came from the between-person variation (77%) than the over-time within-person variation (23%). However, both the within-person variation (*SD* = 35.25) and the between-person variation (*SD* = 65.32) in the test scores were large enough to warrant further analysis. Therefore predictors at both levels would be necessary to explore the variation of the test scores.

Table 3

Results From Unconditional Means Model for Listening: $Y_{ti} = \pi_{0i} + e_{ti}$, $\pi_{0i} = \beta_{00} + r_{0i}$

Parameter	Statistic				
Fixed	Coefficient	*SE*	*t*-ratio	*df*	*p*
Grand mean	333.75	0.47	703.08	19,854	0.00
Random	Variance component	*SD*	Chi-square	*df*	*p*
Person-specific mean	4,267.24	65.32	428,882.10	19,854	0.00
Level 1 error	1,242.78	35.25			
Model fit	Deviance		Parameters		
	1,247,896.79		2		
Test of homogeneity of Level 1 variance	Chi-square		*df*	*p*	
	40,829.70		19,854	0.00	

Reading

On the basis of the unconditional means model for *Reading* scores (see Table 4), the grand mean estimate was 279.30, and 81% of the test score variation came from the between-person variation. Both the within-person variation (*SD* = 36.17) and the between-person variation (*SD* = 73.98) in *Reading* scores were large enough to warrant further analysis.

Table 4

Results From Unconditional Means Model for Reading: $Y_{ti} = \pi_{0i} + e_{ti}, \pi_{0i} = \beta_{00} + r_{0i}$

Parameter	Statistic				
Fixed	Coefficient	*SE*	*t*-ratio	*df*	*p*
Grand mean	279.30	0.54	521.66	19,854	0.00
Random	Variance component	*SD*	Chi-square	*df*	*p*
Person-specific mean	5,473.74	73.98	518,221.43	19,854	0.00
Level 1 error	1,308.38	36.17			
Model fit	Deviance		Parameters		
	1,257,781.68		2		
Test of homogeneity of Level 1 variance	Chi-square		*df*	*p*	
	36,086.43		19,854	0.00	

These results were based on the unconditional means model with the assumption of homogeneity of Level 1 variance across times. The likelihood ratio test suggests that the Level 1 variance was not homogeneous for both *Listening* and *Reading* scores. However, the estimation of fixed effects and their standard errors was robust to the violation of this assumption (Kasim & Raudenbush, 1998). A general estimate of Level 1 variance was needed in further analysis to estimate the variance explained by Level 1 predictors. Therefore results from the unconditional means model with the assumption of homogeneity of Level 1 variance were used for this study. Although both the within-person variation and the between-person variation in the test scores were sufficient to warrant further analysis, we first focused on the within-person variation by including growth parameters and time-varying background variables in the Level 1 model in the following analyses.

Linear Growth Model

Listening

On the basis of the results from the linear growth model for *Listening* scores (see Table 5), the repeaters' average initial score at the first administration time was 316.50, and their scores increased on average by 1.58 points per month in the 4 years of the data collection period. However, there were substantial between-individual variations in both the initial status and increase rate. Specifically, 95% of the repeaters' initial scores were in the range of 316.50 ± 1.96 * $\sqrt{4592.59}$ = (183.67, 449.33), and 95% of the repeaters' score growth rates were in the range of 1.58 ± 1.96 * $\sqrt{1.49}$ = (–0.81, 3.97). In addition, there was a slight negative correlation (–.26) between examinees' initial status and growth rate. Compared with the unconditional means model, the estimated Level 1 residual variance decreased by (1,242.78 – 920.26)/1,242.78 = 26%. Therefore 26% of the within-person variation in *Listening* scores was associated with the linear TIME, but a substantial amount of variance still remained unexplained at Level 1, and more predictors needed to be included in the Level 1 model.

Table 5

Results From Linear Growth Model for Listening: $Y_{ti} = \pi_{0i} + \pi_{1i}\mathrm{TIME}_{ti} + e_{ti}$, $\pi_{0i} = \beta_{00} + r_{0i}$, $\pi_{1i} = \beta_{10} + r_{1i}$

Parameter	Statistic				
Fixed	Coefficient	*SE*	*t*-ratio	*df*	*p*
Mean initial status	316.50	0.50	629.93	19,854	0.00
Mean growth rate	1.58	0.01	112.13	19,854	0.00
Random	Variance component	*SD*	Chi-square	*df*	*p*
Initial status	4,592.59	67.77	227,324.50	19,854	0.00
Growth rate	1.49	1.22	39,265.38	19,854	0.00
Level 1 error	920.26	30.34			
Model fit	Deviance		Parameters		
	1,228,466.99		4		
Test of homogeneity of Level 1 variance	Chi-square		*df*	*p*	
	33,089.66		19,844	0.00	
Tau as correlations	Initial status		Growth rate		
Initial status	1				
Growth rate	–0.26		1		

Reading

On the basis of the linear growth model for *Reading* scores (see Table 6), the repeaters' average initial score was 262.33, with 95% of the initial scores in the range of (115.76, 408.90); the repeaters' scores increased on average by 1.55 points per month, with 95% of the growth rates in the range of (–0.88, 3.98). A slight negative correlation (–.15) between examinees' initial score and growth rate was also found in *Reading* scores. Compared with the unconditional means model, about 25% of the within-person variation in *Reading* scores was associated with the linear TIME. Again, a substantial amount of variance still remained unexplained at Level 1, and more predictors needed to be included in the Level 1 model.

Table 6

Results From Linear Growth Model for Reading: $Y_{ti} = \pi_{0i} + \pi_{1i}\mathrm{TIME}_{ti} + e_{ti}$, $\pi_{0i} = \beta_{00} + r_{0i}$, $\pi_{1i} = \beta_{10} + r_{1i}$

Parameter	Statistic				
Fixed	Coefficient	*SE*	*t*-ratio	*df*	*p*
Mean initial status	262.33	0.55	475.53	19,854	0.00
Mean growth rate	1.55	0.01	106.47	19,854	0.00
Random	Variance component	*SD*	Chi-square	*df*	*p*
Initial status	5,591.54	74.78	252,925.56	19,854	0.00
Growth rate	1.54	1.24	38,419.80	19,854	0.00
Level 1 error	983.37	31.36			
Model fit	Deviance		Parameters		
	1,239,485.53		4		
Test of homogeneity of Level 1 variance	Chi-square		*df*	*p*	
	31,766.17		19,854	0.00	
Tau as correlations	Initial status		Growth rate		
Initial status	1				
Growth rate	–0.15		1		

Quadratic Growth Model

Listening

On the basis of the quadratic growth model for *Listening* scores (see Table 7), on average, the estimated initial score was 312.16, the initial growth rate was 2.83, and acceleration was –.04. The statistically significant negative mean acceleration indicates that repeaters improved their scores at a decreasing rate over time. However, there still were substantial interindividual variations in the initial score, initial growth rate, and acceleration, with 95% of the growth parameters in the ranges of (178.90, 445.42), (–1.89, 7.55), and (–.13, –.05), respectively. There was a slight negative correlation (–.20) between examinees' initial status and initial growth rate but a strong negative correlation between initial growth rate and acceleration (–.95). Compared with the linear growth model, 4% more within-person variation in *Listening* scores was associated with the addition of the quadratic parameter. However, a substantial amount of variance was unpredicted at Level 1.

Table 7

Results From Quadratic Growth Model for Listening: $Y_{ti} = \pi_{0i} + \pi_{1i}\mathrm{TIME}_{ti} + \pi_{2i}\mathrm{TIME}_{ti}^2 + e_{ti}$, $\pi_{0i} = \beta_{00} + r_{0i}, \pi_{1i} = \beta_{10} + r_{1i}, \pi_{2i} = \beta_{20} + r_{2i}$

Parameter	Statistic				
Fixed	Coefficient	*SE*	*t*-ratio	*df*	*p*
Mean initial status	312.16	0.51	611.59	19,854	0.00
Mean growth rate	2.83	0.03	89.32	19,854	0.00
Mean acceleration	–0.04	0.00	–49.43	19,854	0.00
Random	Variance component	*SD*	Chi-square[a]	*df*[a]	*p*[a]
Initial status	4,622.63	67.99	110,895.78	14,007	0.00
Initial growth rate	5.81	2.41	16,482.67	14,007	0.00
Acceleration	0.00	0.05	15,732.26	14,007	0.00
Level 1 error	874.44	29.57			
Model fit	Deviance		Parameters		
	1,225,381.39		7		
Test of homogeneity of Level 1 variance	Chi-square		*df*	*p*	
	22,600.47		13,998	0.00	
Tau as correlations	Initial status	Initial growth rate		Acceleration	
Initial status	1				
Initial growth rate	–0.20	1			
Acceleration	0.10	–0.95		1	

[a] The chi-square statistics are based on 14,008 of 19,855 units that had sufficient data for computation.

Reading

On the basis of the quadratic growth modeling results for *Reading* scores (Table 8), the estimated initial score, initial growth rate, and acceleration were 258.74, 2.59, and –.04, with 95% of the growth parameters in the ranges of (111.99, 405.49), (–2.28, 7.46), and (–.14, .06), respectively. A slight negative correlation between examinees' initial status and initial growth rate (–.12) and a strong negative correlation between initial growth rate and acceleration (–.94) were also found in *Reading* scores. Compared with the linear growth model, 3% more within-person variation in *Reading* scores was associated with the addition of the quadratic parameter. Again, there was still a substantial amount of unexplained variance at Level 1.

Table 8

Results From Quadratic Growth Model for Reading: $Y_{ti} = \pi_{0i} + \pi_{1i}\mathrm{TIME}_{ti} + \pi_{2i}\mathrm{TIME}_{ti}^{2} + e_{ti}$, $\pi_{0i} = \beta_{00} + r_{0i}, \pi_{1i} = \beta_{10} + r_{1i}, \pi_{2i} = \beta_{20} + r_{2i}$

Parameter	Statistic				
Fixed	Coefficient	*SE*	*t*-ratio	*df*	*p*
Mean initial status	258.74	0.56	463.17	19,854	0.00
Mean growth rate	2.59	0.03	78.74	19,854	0.00
Mean acceleration	–0.04	0.00	–39.46	19,854	0.00
Random	Variance component	*SD*	Chi-square[a]	*df*[a]	*p*[a]
Initial status	5,605.55	74.87	123,840.97	14,007	0.00
Initial growth rate	6.17	2.48	16,436.96	14,007	0.00
Acceleration	0.00	0.05	15,575.27	14,007	0.00
Level 1 error	939.35	30.65			
Model fit	Deviance		Parameters		
	1,237,339.16		7		
Test of homogeneity of Level 1 variance	Chi-square		*df*	*p*	
	22,349.37		14,007	0.00	
Tau as correlations	Initial status	Initial growth rate		Acceleration	
Initial status	1				
Initial growth rate	–0.12	1			
Acceleration	0.05	–0.94		1	

[a] The chi-square statistics are based on 14,008 of 19,855 units that had sufficient data for computation.

Cubic Growth Model

To explore the repeaters' score growth trajectories, the cubic growth parameter TIME^3 was added in the Level 1 quadratic growth model. The results from the cubic growth models for *Listening* and *Reading* scores are presented in Tables 9 and 10, respectively. Compared with the quadratic growth modeling results, 1% and 2% more within-person variation in *Listening* and *Reading* scores, respectively, was associated with the addition of the cubic parameter.

Table 9

Results From Cubic Growth Model for Listening: $Y_{ti} = \pi_{0i} + \pi_{1i}\mathrm{TIME}_{ti} + \pi_{2i}\mathrm{TIME}^2_{ti} + \pi_{3i}\mathrm{TIME}^3_{ti} + e_{ti}$, $\pi_{0i} = \beta_{00} + r_{0i}, \pi_{1i} = \beta_{10} + r_{1i}, \pi_{2i} = \beta_{20} + r_{2i}, \pi_{3i} = \beta_{30} + r_{3i}$

Parameter	Statistic				
Fixed	Coefficient	*SE*	*t*-ratio	*df*	*p*
Mean initial status	310.44	0.52	602.45	19,854	0.00
Mean linear	3.91	0.06	67.14	19,854	0.00
Mean quadratic	–0.13	0.00	–33.19	19,854	0.00
Mean cubic	0.00	0.00	23.32	19,854	0.00
Random	Variance component	*SD*	Chi-square[a]	*df*[a]	*p*[a]
Initial status	4,632.30	68.06	9,777.31	1,671	0.00
Linear	12.25	3.50	1,597.36	1,671	>0.50
Quadratic	0.03	0.18	1,572.82	1,671	>0.50
Cubic	0.00	0.00	1,583.87	1,671	>0.50
Level 1 error	853.96	29.22			
Model fit	Deviance		Parameters		
	1,224,738.20		11		
Test of homogeneity of Level 1 variance	Chi-square		*df*	*p*	
	2,841.46		1,671	0.000	
Tau as correlations	Initial status	Linear		Quadratic	Cubic
Initial status	1				
Linear	–0.17	1			
Quadratic	0.07	–0.90		1	
Cubic	–0.05	0.79		–0.98	1

[a]The chi-square statistics are based on 1,672 of 19,855 units that had sufficient data for computation.

Table 10

Results From Cubic Growth Model for Reading: $Y_{ti} = \pi_{0i} + \pi_{1i}\mathrm{TIME}_{ti} + \pi_{2i}\mathrm{TIME}_{ti}^{2} + \pi_{3i}\mathrm{TIME}_{ti}^{3} + e_{ti}$, $\pi_{0i} = \beta_{00} + r_{0i}, \pi_{1i} = \beta_{10} + r_{1i}, \pi_{2i} = \beta_{20} + r_{2i}, \pi_{3i} = \beta_{30} + r_{3i}$

Parameter	Statistic				
Fixed	Coefficient	*SE*	*t*-ratio	*df*	*p*
Mean initial status	257.28	0.56	456.44	19,854	0.00
Mean linear	3.51	0.06	57.70	19,854	0.00
Mean quadratic	−0.11	0.00	−27.45	19,854	0.00
Mean cubic	0.00	0.00	19.84	19,854	0.00
Random	Variance component	*SD*	Chi-square[a]	*df*[a]	*p*[a]
Initial status	5,620.28	74.97	11,601.14	1,671	0.00
Linear	14.09	3.75	1,640.27	1,671	>0.50
Quadratic	0.03	0.18	1,597.81	1,671	>0.50
Cubic	0.00	0.00	1,585.51	1,671	>0.50
Level 1 error	918.68	30.31			
Model fit	Deviance		Parameters		
	1,236,790.95		11		
Test of homogeneity of Level 1 variance	Chi-square		*df*	*p*	
	2,895.64		1,671	0.00	
Tau as correlations	Initial status	Linear		Quadratic	Cubic
Initial status	1				
Linear	−0.12	1			
Quadratic	0.06	−0.92		1	
Cubic	−0.05	0.86		−0.99	1

[a] The chi-square statistics are based on 1,672 of 19,855 units that had sufficient data for computation.

On the basis of the cubic model, all the growth parameters, except for the initial score, were fixed with no interindividual variations for both *Listening* and *Reading* scores, which is not consistent with what we found in the preliminary analyses and other models. It seems implausible that the linear, quadratic, and cubic growth parameters were invariant across individual examinees. It is very possible that there were not enough data for the cubic model to produce accurate chi-square statistics for the variances of the parameter estimates (e.g., the chi-square statistics were based on only 1,672 of 19,855 examinees). On the basis of the model fit statistics, the deviance drop was substantially smaller than the deviance drop from the linear growth model to the quadratic model and the deviance drop from the linear growth model to the unconditional means model.

To further evaluate model fit, at each of the 0–47 administration time points, we computed repeaters' fitted score means based on individual growth trajectories and then plotted the fitted score means over time based on both quadratic and cubic models. Figures 2 and 3 show the plots of observed score means and fitted score means based on the two models for *Listening* and *Reading*, respectively. On the basis of the plots, the two models had very similar model fit with the observed data, except at the last four time points, where the cubic growth trajectories were slightly better than the quadratic growth trajectories.

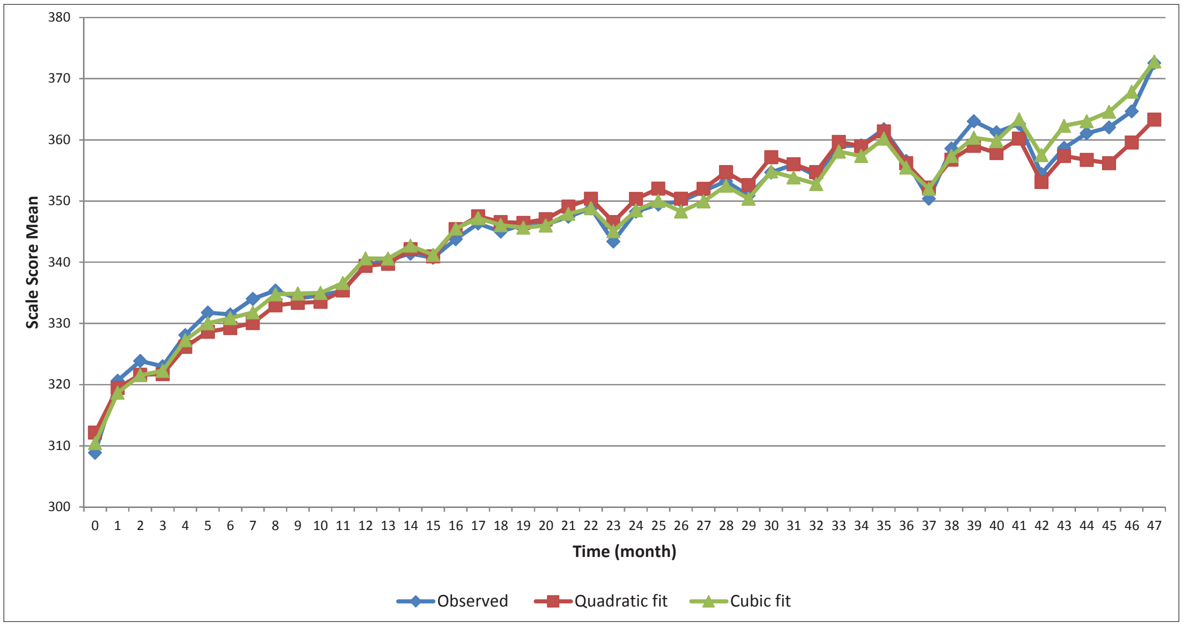


Figure 2. **Fitted and observed score means over time for Listening.**

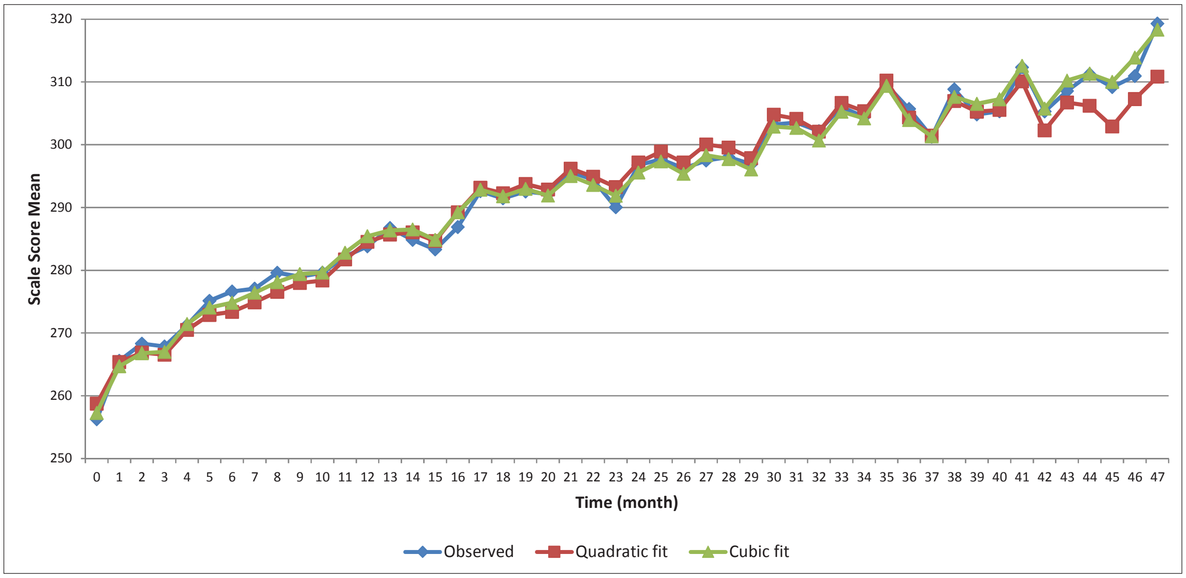


Figure 3. **Fitted and observed score means over time for Reading.**

On the basis of the growth parameter estimates, fit statistics, and the principle of parsimony in statistical modeling, the quadratic growth model was considered to be the appropriate growth model in this study.

Conditional Quadratic Growth Model

The quadratic growth model was selected to describe repeaters' score change patterns over time for both *Listening* and *Reading*. However, the examinees' background might have an impact on their score change patterns. Therefore two time-varying background variables, current occupation and daily English use time, were added in the Level 1 quadratic growth model, and three person-level background variables, gender, educational level, and test-taking experience, were added as predictors in the Level 2 models for the growth parameters. The two time-varying background variables were first separately added in the Level 1 model and their impacts on examinees' scores over time were evaluated; after the important time-varying background variables were selected to remain in the Level 1 model, the three person-level background variables were added in the Level 2 models to evaluate their impacts on the growth parameters.

Listening

On the basis of the model fit statistics, fixed effect coefficients, and the principle of parsimony in statistical modeling, the time-varying background variable daily English use time was selected to remain in the Level 1 quadratic model; the person-level background variables gender and test-taking experience remained in the Level 2 models for all three growth parameters (i.e., initial status, initial growth rate, and acceleration), and educational level remained in the Level 2 model only for the initial status.

Table 11 shows the results from the final conditional quadratic model for Listening scores. On the basis of the model, the examinees' background variables had significant impacts on their Listening score growth trajectories. For example, for the initial status, women had higher average scores than men; examinees with vocational/technical high school education had lower average score than examinees with all other educational levels; and examinees without previous test-taking experience had lower average scores than examinees with experience. The examinees' scores tended to increase with their daily English use time. For score growth rate, women had lower initial growth rates than men; examinees without test-taking experience had higher initial growth rate than examinees with experience; and both gender and test-taking experience had impacts on the acceleration of the score growth.

Table 11

Results From Conditional Quadratic Model for Listening:

$$Y_{ti} = \pi_{0i} + \pi_{1i}\text{ENGUSE}_{ti} + \pi_{2i}\text{TIME}_{ti} + \pi_{3i}\text{TIME}^2_{ti} + e_{ti},$$

$$\pi_{0i} = \beta_{00} + \beta_{01}\text{GENDER}_i + \beta_{02}\text{EDUMIS}_i + \beta_{03}\text{SECOND1}_i + \beta_{04}\text{SECOND2}_i + \beta_{05}\text{VOTECH}_i$$

$$+\beta_{06}\text{COMMUN}_i + \beta_{07}\text{UNDERGR}_i + \beta_{08}\text{GRADUA}_i + \beta_{09}\text{LANGUA}_i + \beta_{10}\text{PREEXP}_i + r_{0i},$$

$$\pi_{1i} = \beta_{10},\ \pi_{2i} = \beta_{20} + \beta_{21}\text{GENDER}_i + \beta_{22}\text{PREEXP}_i + r_{2i},$$

$$\pi_{3i} = \beta_{30} + \beta_{31}\text{GENDER}_i + \beta_{32}\text{PREEXP}_i + r_{3i}$$

Parameter	Statistic				
Fixed	Coefficient	*SE*	*t*-ratio	*df*	*p*
For initial status					
Intercept	271.91	3.40	79.91	19,844	0.00
GENDER	30.87	1.06	29.08	19,844	0.00
EDUMIS	35.68	4.87	7.33	19,844	0.00
SECOND1	74.87	13.49	5.55	19,844	0.00
SECOND2	22.94	4.39	5.23	19,844	0.00
VOTECH	11.86	4.81	2.47	19,844	0.01
COMMUN	33.13	4.51	7.35	19,844	0.00
UNDERGR	32.58	3.43	9.51	19,844	0.00
GRADUA	32.87	3.54	9.29	19,844	0.00
LANGUA	57.73	10.05	5.75	19,844	0.00
PREEXP	−34.18	1.13	−30.14	19,844	0.00
For ENGUSE slope					
Intercept	2.74	0.18	15.60	119,112	0.00
For mean linear slope					
Intercept	2.63	0.04	63.08	19,852	0.00
GENDER	−0.42	0.06	−6.47	19,852	0.00
PREEXP	1.46	0.08	17.22	19,852	0.00
For mean acceleration slope					
Intercept	−0.04	0.00	−35.43	19,852	0.00
GENDER	0.01	0.00	4.68	19,852	0.00
PREEXP	−0.02	0.00	−9.10	19,852	0.00
Random	Variance component	*SD*	Chi-square[a]	*df*[a]	*p*[a]
Initial status	4,104.09	64.06	100,977.36	13,997	0.00
Initial growth rate	5.51	2.35	16,385.22	14,005	0.00
Acceleration	0.00	0.05	15,741.04	14,005	0.00
Level 1 error	874.24	29.57			
Model fit	Deviance		Parameters		
	1,222,947.36		7		
Test of homogeneity of Level 1 variance	Chi-square		*df*	*p*	
	22,886.25		14,005	0.00	

Parameter	Statistic		
Tau as correlations	Initial status	Initial growth rate	Acceleration
Initial status	1		
Initial growth rate	−0.14	1	
Acceleration	0.03	−0.95	1

[a]The chi-square statistics are based on 14,008 of 19,855 units that had sufficient data for computation.

Reading

Similar to the modeling results for *Listening* scores, the variable daily English use time was selected to remain in the Level 1 quadratic model for *Reading* scores. However, the variable gender remained only in the models for the initial status and initial growth rate, the variable educational level remained only in the model for the initial status, and the variable test-taking experience remained in the models for all three growth parameters. Table 12 shows the results from the final conditional quadratic model for *Reading* scores. Compared with the findings for *Listening* scores, the examinees' background variables had similar impacts on their *Reading* score growth trajectories, except that gender did not have an impact on the acceleration of the *Reading* score growth.

Table 12

Results From Conditional Quadratic Model for Reading:

$$Y_{ti} = \pi_{0i} + \pi_{1i}\mathrm{ENGUSE}_{ti} + \pi_{2i}\mathrm{TIME}_{ti} + \pi_{3i}\mathrm{TIME}_{ti}^{2} + e_{ti}\,\mathbf{,}$$

$$\pi_{0i} = \beta_{00} + \beta_{01}\mathrm{GENDER}_{i} + \beta_{02}\mathrm{EDUMIS}_{i} + \beta_{03}\mathrm{SECOND2}_{i} + \beta_{04}\mathrm{COMMUN}_{i} + \beta_{05}\mathrm{UNDERGR}_{i}$$

$$+\beta_{06}\mathrm{GRADUA}_{i} + \beta_{07}\mathrm{LANGUA}_{i} + \beta_{08}\mathrm{PREEXP}_{i} + r_{0i}\,,$$

$$\pi_{1i} = \beta_{10}\,\mathbf{,}\ \pi_{2i} = \beta_{20} + \beta_{21}\mathrm{GENDER}_{i} + \beta_{22}\mathrm{PREEXP}_{i} + r_{2i}\,\mathbf{,}\ \pi_{3i} = \beta_{30} + \beta_{31}\mathrm{PREEXP}_{i} + r_{3i}$$

Parameter	Statistic				
Fixed	Coefficient	*SE*	*t*-ratio	*df*	*p*
For initial status					
Intercept	202.73	2.72	74.61	19,846	0.00
GENDER	6.48	1.15	5.63	19,846	0.00
EDUMIS	56.72	4.78	11.88	19,846	0.00
SECOND2	11.44	3.95	2.89	19,846	0.00
COMMUN	39.57	4.25	9.32	19,846	0.00
UNDERGR	57.19	2.71	21.08	19,846	0.00
GRADUA	65.52	2.93	22.37	19,846	0.00
LANGUA	47.84	11.68	4.10	19,846	0.00
PREEXP	−27.46	1.27	−21.65	19,846	
For ENGUSE slope					
Intercept	2.72	0.18	14.97	119,115	0.00

Parameter	Statistic				
For mean linear slope					
Intercept	2.31	0.04	62.49	19,852	0.00
GENDER	−0.15	0.03	−5.34	19,852	0.00
PREEXP	1.36	0.09	15.71	19,852	0.00
For mean acceleration slope					
Intercept	−0.03	0.00	−31.15	19,853	0.00
PREEXP	−0.02	0.00	−8.70	19,853	0.00
Random	Variance component	*SD*	Chi-square[a]	*df*[a]	*p*[a]
Initial status	5,176.85	71.95	115,746.18	13,999	0.00
Initial growth rate	5.73	2.39	16,283.00	14,005	0.00
Acceleration	0.00	0.05	15,518.41	14,006	0.00
Level 1 error	940.11	30.66			
Model fit	Deviance		Parameters		
	1,235,544.37		7		
Test of homogeneity of Level 1 variance	Chi-square		*df*	*p*	
	22,346.17		14,007	0.00	
Tau as correlations	Initial status	Initial growth rate		Acceleration	
Initial status	1				
Initial growth rate	−0.08	1			
Acceleration	0.02	−0.94		1	

[a] The chi-square statistics are based on 14,008 of 19,855 units that had sufficient data for computation.

For both *Listening* and *Reading* scores, the conditional quadratic growth model fit better than the quadratic growth model based on the deviance statistics. However, the examinees' initial growth rates still had very strong negative relations with acceleration over time (−.95 for *Listening* and −.94 for *Reading*). In addition, there were still substantial between-individual variations in the growth trajectories, which include examinees' initial status, initial growth rate, and acceleration over time.

Model Validation

The data of another group of 1,861 examinees who had taken the test 12 times in the same 4 years were used to examine the validity of the selected quadratic growth models (results are not presented in this report). A comparison of the linear, quadratic, and cubic growth models for the new group's *Listening* and *Reading* scores found that the quadratic model was the most appropriate model. The average initial score, initial growth rate, and acceleration based on 1,861 examinees were consistent with the growth parameters based on the 19,855 examinees. The strong negative correlation between initial growth rate and acceleration also remained consistent between the two samples.

When the conditional quadratic growth models based on the data of the 19,855 examinees were applied to the new group's *Listening* and *Reading* scores, the association of examinees' background variables with their growth trajectories remained similar.

Discussion

It is an important part of quality control for a testing program to monitor test performance across administrations, and various methods and procedures have been proposed for this purpose (von Davier, 2012). The existence of the same examinees who repeat the test in different administrations provides data to evaluate test performance over time. A testing program can use repeaters' data across administrations to examine score change patterns and then use these patterns to monitor test performance over time. This study used multilevel growth modeling to analyze a balanced but unfixed data set in which all examinees repeated the same number of test administrations but with variable intervals between test takings. The definition of TIME as the number of months that had elapsed from the first time tested and the use of equated scores from different administrations and forms put all examinees in the same framework for growth modeling analysis.

On the basis of the unconditional means model, the test scores varied much more among different examinees than they varied over time within persons. The within-person score variation in the 4 years was close to the standard error of score difference (i.e., 35; see Educational Testing Service [ETS], 2013) for each of the *Listening* and *Reading* sections in the TOEIC Listening and Reading test, which indicates the stability of test performance over time.

On the basis of the linear growth modeling results, the constant growth rate over time for both *Listening* and *Reading* was small, with about a 1.6 score point increase per month, which suggests the stability of repeaters' scores over time. However, as expected, the between-person variations in both initial status and growth rate were large: Individuals began at different proficiency levels, and they made progress at different rates. Although the linear growth model fit much better than the unconditional means model, a substantial proportion of within-person score variation still remained unrelated to the linear TIME predictor. In addition, a closer look at the plots of observed score means and fitted score means suggested that repeaters' scores did not increase at a constant rate, especially at the very beginning (i.e., from the first to the second times) and later times of testing. Therefore the linear growth model might describe repeaters' growth trajectories in the earlier times of testing (except for the first repetition) but may not account for the changing growth rate in their long-term score change patterns.

The quadratic growth model uses a linear parameter to estimate the initial growth rate at the very beginning and a quadratic parameter to estimate acceleration over time. The quadratic modeling results indicate that the repeaters' scores tended to increase more in their earlier repetitions, but the increase rate declined gradually over time. The repeater's growth trajectories based on the data in this study were consistent with the repeater score change patterns found in other testing programs, such as the *TOEFL*® test (Wilson, 1987), the *SAT*® I test (Nathan & Camara, 1998), and the *GRE*® General test (Rock & Werts, 1979).

The quadratic growth modeling yielded slight negative correlations between examinees' initial status and initial growth rate, which means that examinees with lower initial scores tended to have somewhat higher growth rates in their early times of testing. This finding was consistent with previous studies (Nathan & Camara, 1998; Wei & Morgan, 2016; Wilson, 1987; Yang et al., 2011). The relatively low negative correlations may be related to the repeater group composition in this study. Only 22.35% of the repeaters had never taken the test before. In other words, the majority of repeaters had taken the test at least once before the data collection period for this study, and the initial scores were not really their first-time scores in their test-taking experience. Accordingly, their initial scores in the data collection period did not show strong relations with their score changes. However, the quadratic growth modeling produced a very strong negative correlation between repeaters' initial growth rate and acceleration, which means that the examinees with lower initial growth rate tended to have higher acceleration over the growth trajectory. The quadratic growth modeling can easily find this repeater score change pattern, but descriptive and simple analyses often ignored the pattern.

The negative acceleration parameter estimate in the quadratic growth model suggests that the increase rate would decline over time. At what point in time would scores no longer exhibit any significant increase? The repeaters' fitted mean score plots based on individual growth trajectories did not show how the score increase rate changed over time. Figures 4 and 5 show the average quadratic and cubic growth functions (i.e., group growth trajectories) and observed score means over time for *Listening* and *Reading*, respectively. The plots based on the quadratic growth function showed that the mean score growth rate changed from positive to negative in the 34th month for *Listening* and in the 37th month for *Reading*. It is probably unreasonable to believe that repeaters' scores would increase in the earlier times but decrease in the later times in 4 years. The cubic growth function showed that repeaters' scores increased faster in the earlier time, then increased slowly in the middle, and finally increased faster again in the later time. It seems that such a cubic growth trajectory is consistent with the general learning curve with plateau phase for many skills. Comparing the average quadratic and cubic growth trajectories with the observed means plots found that both models worked equally well for most of the time points, but the cubic model fit better with the observed data in the last few administration months for both *Listening* and *Reading*. Additional longitudinal data with more times of testing may provide stronger empirical evidence for the cubic growth model. However, the quadratic model was selected in this study based on the principle of parsimony in statistical modeling and the convenience in interpretation of growth parameters. Compared with the cubic model, the quadratic growth parameters are much easier to interpret for repeaters' score change patterns in the testing program. This is particularly true when examinees' background variables were included in the growth models.

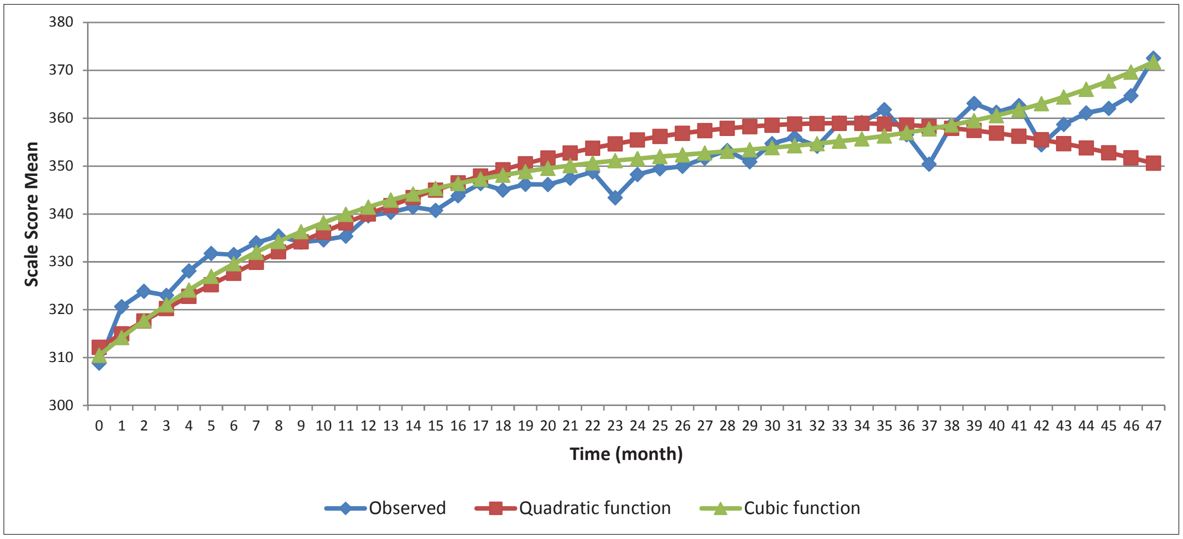


Figure 4. **Growth functions and observed score means over time for Listening.**

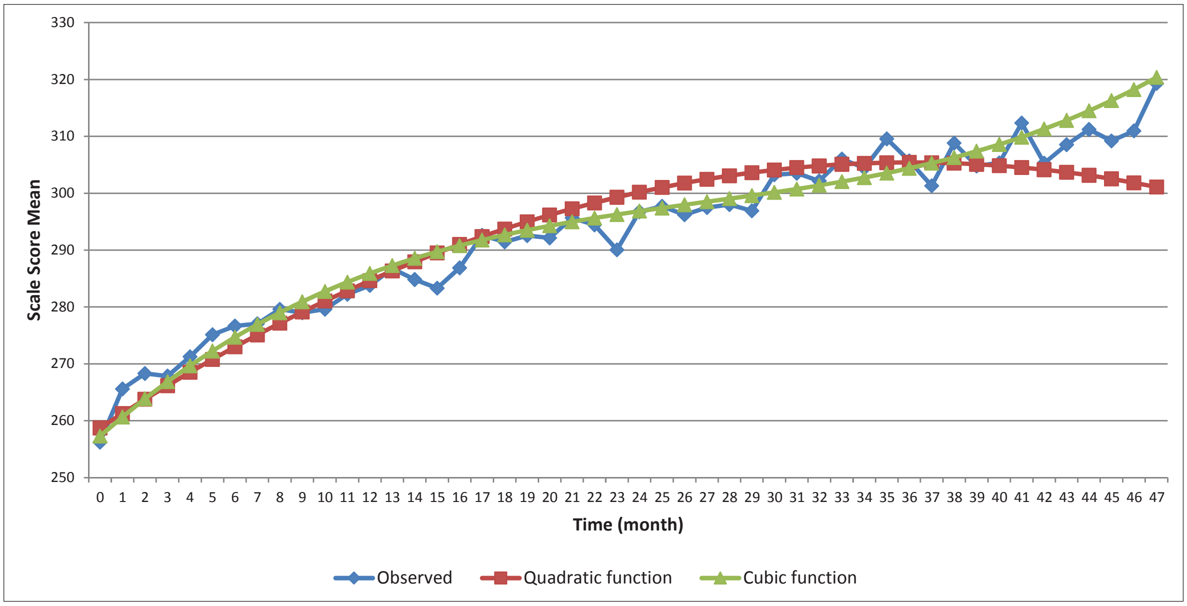


Figure 5. **Growth functions and observed score means over time for Reading.**

The growth modeling results have important implications for the testing program. For the quality control of test performance, the repeaters' score change patterns can be used for the evaluation of the TOEIC Listening and Reading test scores from different perspectives. From a reliability perspective, the stability of repeaters' scores was a strong indicator of a high reliability of test scores across administrations, across forms, and over time. Reliability refers to the extent to which test scores are consistent across forms or occasions of testing. Therefore a testing program can evaluate test score reliability by examining form-to-form differences or differences in performance over time (ETS, 2014). The existence of many repeaters across administrations in the TOEIC Listening and Reading test provides empirical data to evaluate the consistency of test scores across forms and over time. The linear, quadratic, and cubic models in this study found very small monthly score increases. After taking account of the skill improvement due to learning or maturation, the test scores can be considered as consistent across forms and testing occasions.

From a validity perspective, the growth modeling results provided empirical evidence based on the relations of test scores to other variables. For validity evidence, the patterns of association between test scores and other variables should be consistent with theoretical expectations (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). This study included TIME and examinees' backgrounds as "other variables." The relations of TOEIC Listening and Reading test scores to these variables were consistent with related experience, theory, and previous studies. For example, (a) the repeaters' score growth trajectories (i.e., the relations of test scores to TIME) on the TOEIC tests were consistent with the repeaters' score change patterns found in other popular testing programs (e.g., TOEFL, GRE, and SAT), (b) the cubic growth trajectories (i.e., the relations of test scores to TIME) were consistent with the general learning curve for many skills, (c) the examinees' score increases with their daily English use time was consistent with language learning experience, and (d) the impact of examinees' educational level on their scores was consistent with English education experience.

From the test users' perspective, the gradual and stable score increase patterns supported the claim that TOEIC Listening and Reading test scores are suitable for measurement of progress in English proficiency over time. The small monthly growth rate reflected the stability of test scores, which supports the validity of the test scores for the intended use over time (von Davier, 2012). The increasing trend in test scores over time reflected repeaters' performance growth due to maturation and learning. Therefore test users can use TOEIC Listening and Reading test scores to evaluate English learning and training progress. In addition, the growth modeling results may help test takers or test users make decisions about retesting and learning strategies. For example, examinees with no previous test-taking experience tended to obtain lower scores than examinees with experience, but their scores increased more at the next testing. This may indicate that test takers can improve test scores by being familiar with the test and by retaking the test. Also, because examinees' scores increased with their daily English use time, one effective way to improve test scores is to use English more often in daily life. Furthermore, it is not unusual for English learners to make rapid progress at the beginning, then make slow or even little progress, and finally make apparent progress again if they keep learning. It may be helpful to know this learning curve for English learning and training.

The testing program can also use repeaters' growth modeling results to predict their performance on future administrations. The growth modeling results indicate that (a) the repeaters' fitted score means based on the models were consistent with their observed score means and (b) the growth parameters based on two different samples were very close to each other. This suggests that growth modeling is very promising for predicting repeaters' score means at the group level, which is consistent with findings from other studies (Wei, 2013; Wei & Qu, 2014). Therefore we can monitor test performance by comparing repeaters' observed and predicted score means. If repeaters' observed growth patterns are very different from the expected growth patterns, the testing program needs to investigate the inconsistency and find underlying reasons for it, such as population changes, scoring mistakes, or security breaches.

However, the testing program should be careful when making a judgment for an individual test taker. At the individual level, this study found that about 30% of score variance could be predicted based on the quadratic models. There were substantial individual variations in the growth trajectories, which is consistent with the findings of other studies (e.g., Wei & Morgan, 2016; Yang et al., 2011; Zhang, 2008). Therefore it is difficult to accurately predict individual repeaters' test scores. It is misleading to use the average growth trajectories to make a judgment about an individual's score change pattern.

To explore the individual variations in the group's growth trajectories, we have at least two ways to distinguish different growth patterns. One way is to use observed covariates to identify different patterns based on observed subgroups, as we did in this study. For example, adding the covariate test-taking experience helped us distinguish growth patterns for examinees without any test-taking experience and examinees with experience. More observed covariates can be included in the models to distinguish different patterns in future studies. The other way is to let the data distinguish growth patterns based on latent or underlying subgroups. Future studies can use latent class or mixture modeling methods to explore different latent score change patterns in the observed repeaters' data (e.g., Wei, 2016).

Conclusions

On the basis of the multilevel growth modeling analysis of the TOEIC Listening and Reading test scores of 19,855 examinees who had taken the test six times in 4 years, this study found that (a) examinees' scores increased with repeated testing; (b) examinees' score increase rates were higher in the early repetitions, then gradually dropped over time; (c) examinees without previous test-taking experience tended to have lower initial scores and higher initial increase rates, and their score increase rates tended to drop faster over time; (d) examinees' educational background had a significant relationship with their initial scores but had little association with their score increase rates; and (e) examinees' gender had some relationship to their initial scores and increase rates. The results suggest that multilevel growth modeling analysis can be used to evaluate test performance across administrations by exploring repeaters' score change patterns over time. Furthermore, growth modeling results support the reliability and validity of the TOEIC scores. The results also indicate that TOEIC scores can be used to evaluate English learning or training progress.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Castellano, K. E., & Ho, A. D. (2013). *A practitioner's guide to growth models*. Washington, DC: Council of Chief State School Officers.

Educational Testing Service. (2013). *TOEIC® user's guide*. Princeton, NJ: Author.

Educational Testing Service. (2014). *ETS standards for quality and fairness*. Princeton, NJ: Author.

Kasim, R., & Raudenbush, S. (1998). Application of Gibbs sampling to nested variance components models with heterogeneous within-group variance. *Journal of Educational and Behavioral Statistics, 20*, 93–116. https://doi.org/10.3102/10769986023002093

Kingston, N., & Turner N. (1984). *Analysis of score change patterns of examinees repeating the Graduate Record Examinations® General Test* (Research Report No. RR-84-22). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1984.tb00062.x

Lee, Y.-H., & Haberman, S. J. (2013). Harmonic regression and scale stability. *Psychometrika*, *78*, 815–829. https://doi.org/10.1007/s11336-013-9337-1

Lee, Y.-H., Liu, M., & von Davier, A. A. (2013). Detection of unusual test administrations using a linear mixed effects model. In R. Millsap, L. van der Ark, D. Bolt, & C. Woods (Eds.), *New developments in quantitative psychology: Proceedings of the 77th international meeting of the Psychometric Society* (Vol. 66, pp. 133–149). New York, NY: Springer. https://doi.org/10.1007/978-1-4614-9348-8_9

Lee, Y.-H., & von Davier, A. A. (2013). Monitoring scale scores over time via quality control charts, model-based approaches, and time series techniques. *Psychometrika*, *78*, 557–575. https://doi.org/10.1007/s11336-013-9317-5

Li, D., Li, S., & von Davier, A. A. (2011). Applying time-serious analysis to detect scale drift. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 327–346). New York, NY: Springer.

Nathan, J. S., & Camara, W. J. (1998). *Score change when retaking the SAT I: Reasoning Test* (Research Note No. RN-05). New York, NY: College Board.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Application and data analysis methods*. Thousand Oaks, CA: Sage.

Rock, D. R., & Werts, C. (1979). *An analysis of time-related increments and/or decrements for GRE® repeaters across ability and sex groups* (GRE Board Research Report No. 77-9R). Princeton, NJ: Educational Testing Service.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* New York, NY: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195152968.001.0001

von Davier, A. A. (2012). *The use of quality control and data mining techniques for monitoring scaled scores: An overview* (Research Report No. RR-12-20). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2012.tb02302.x

Wei, Y. (2013). Monitoring *TOEIC®* Listening and Reading test performance across administrations using examinees' background information. In D. E. Powers (Ed.), *The research foundation for the TOEIC® tests: A compendium of studies* (2nd ed., pp. 11.0–11.28). Princeton, NJ: Educational Testing Service.

Wei, Y. (2016, April). *Using growth mixture modeling to explore test takers' score change patterns.* Paper presented at the annual meeting of National Council on Measurement in Education, Washington, DC.

Wei, Y., & Morgan, R. (2016). *An evaluation of the single-group growth model (SGGM) as an alternative to common-item equating* (Research Report No. RR-16-01). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12087

Wei, Y., & Qu, Y. (2014). *Using multilevel analysis to monitor test performance across administrations* (Research Report No. RR-14-29). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12029

Wilson, K. M. (1987). *Patterns of test taking and score change for examinees who repeat the Test of English as a Foreign Language* (Research Report No. RR-87-03). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1987.tb00207.x

Yang, W., Bontya, A. M., & Moses, T. M. (2011). *Repeater effects on score equating for a graduate admissions exam* (Research Report No. RR-11-17). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2011.tb02253.x

Zhang, Y. (2008). *Repeater analyses for TOEFL iBT®* (Research Memorandum No. RM-08-05). Princeton, NJ: Educational Testing Service.

Appendix

The unconditional means model is

$$Y_{ti} = \pi_{0i} + e_{ti}$$

$$\pi_{0i} = \beta_{00} + r_{0i},$$

where Y_{ti} is the test score of examinee i at administration time t; π_{0i} is the score mean of examinee i across administration times; e_{ti} is the residual or unique effect associated with examinee i at administration time t (i.e., within-person deviation) and is assumed to be normally distributed with N (0, σ^2); β_{00} is the grand score mean (i.e., the average of all test takers' scores over time) of the population of test takers; and r_{0i} is the random effect associated with the examinee i (i.e., between-person deviation) and is assumed to be normally distributed with N (0, τ_{00}).

The linear growth model is

$$Y_{ti} = \pi_{0i} + \pi_{1i}\mathrm{TIME}_{ti} + e_{ti,}$$

$$\pi_{0i} = \beta_{00} + r_{0i,}$$

$$\pi_{1i} = \beta_{10} + r_{1i,}$$

where TIME_{ti} is the amount of time that had elapsed in months from the first time the examinee i took the test to administration time t; π_{1i} is the growth rate for examinee i over the 4 years of data collection and represents the expected change during a fixed unit of time (i.e., a month); π_{0i}, the intercept, is the initial status or the true score of examinee i at the first administration time (i.e., $\mathrm{TIME}_{ti} = 0$); e_{ti} is the deviation (i.e., residual) of examinee i at administration time t from his or her true linear growth trajectory; β_{00} and β_{10} are the mean intercept and mean linear growth rate that represent the mean growth trajectory of the population; and r_{0i} and r_{1i} are the deviations of examinee i's trajectory from the mean growth trajectory of the population in terms of initial status and linear growth rate, with a variance–covariance matrix:

$$\begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{pmatrix},$$

where τ_{00} is the unconditional variance in the Level 1 intercepts, τ_{11} is the unconditional variance in the Level 1 growth rates, and τ_{01} or τ_{10} is the unconditional covariance between the Level 1 intercepts and linear growth rates.

The quadratic growth model is

$$Y_{ti} = \pi_{0i} + \pi_{1i}\mathrm{TIME}_{ti} + \pi_{2i}\mathrm{TIME}_{ti}^{2} + e_{ti,}$$

$$\pi_{0i} = \beta_{00} + r_{0i,}$$

$$\pi_{1i} = \beta_{10} + r_{1i,}$$

$$\pi_{2i} = \beta_{20} + r_{2i,}$$

where the linear component, π_{1i}, is the instantaneous growth rate for examinee i at the first administration time; the quadratic component, π_{2i}, is the acceleration in the growth trajectory; β_{00}, β_{10}, and β_{20} are the mean intercept, mean instantaneous growth rate, and mean acceleration of the population, respectively; and r_{0i}, r_{1i}, and r_{2i} are the deviations of examinee i's trajectory from the mean growth trajectory of the population in terms of initial status, instantaneous growth rate, and acceleration, respectively, with a variance–covariance matrix:

$$\begin{pmatrix} \tau_{00} & \tau_{01} & \tau_{02} \\ \tau_{10} & \tau_{11} & \tau_{12} \\ \tau_{20} & \tau_{21} & \tau_{22} \end{pmatrix},$$

where the variance and covariance have similar interpretations in the linear growth model, with an addition of the acceleration component.

The conditional quadratic growth model is

$$Y_{ti} = \pi_{0i} + \pi_{1i}\mathrm{COV}_{ti} + \pi_{2i}\mathrm{TIME}_{ti} + \pi_{3i}\mathrm{TIME}_{ti}^{2} + e_{ti,}$$

$$\pi_{0i} = \beta_{00} + \beta_{01}X_{i} + r_{0i,}$$

$$\pi_{1i} = \beta_{10} + r_{1i,}$$

$$\pi_{2i} = \beta_{20} + \beta_{21}X_{i} + r_{2i,}$$

$$\pi_{3i} = \beta_{30} + \beta_{31}X_{i} + r_{3i,}$$

where the time-varying background covariate COV_{ti} was added in the Level 1 model and the person-level background variable X_i was added in the Level 2 models.

Compendium Study

Linking *TOEIC*® Speaking Test Scores Using *TOEIC*® Listening Test Scores

Sooyeon Kim

Assess to Progress.

In developing multiple forms of a test, test developers use test specifications to ensure that the alternate forms are similar in content and statistical characteristics. As well specified as the test development process may be, typically, slight differences may occur in the statistical difficulty of the alternate forms. For tests containing constructed-response (CR) items that require test takers to construct responses (instead of selecting them from multiple choices), the specifications must also include a scoring rubric for each item, which must be consistently applied by the raters when the CR items are employed in different test forms or administrations. Even so, CR items bring certain complications in that rater standards may shift slightly from one administration to another, even if the scoring rubric has not changed. Thus, one form can be more difficult than another due to either (a) the inclusion of more difficult items, (b) more stringent scoring by raters, or (c) both. Under these circumstances, scores on one form would not indicate the same level of ability as the same scores on another. Test equating is a statistical method for adjusting for difference in difficulty among forms that are built to the same specifications. Various equating designs and methods have been discussed thoroughly in the literature (Kolen & Brennan, 2004). Perhaps most often, equating occurs in the context of the nonequivalent groups with anchor test (NEAT) design, in which a set of items common to both the new and reference forms is used to place both forms on the same scale.

In using a NEAT design, a major drawback with tests comprising CR items is the difficulty of identifying a satisfactory anchor test. In many cases, for example, CR items are not reused across different test forms because of ease of memorization (Muraki, Hombo, & Lee, 2000), so that there are no common CR items available for equating. Even if CR items were reused, the CR anchor items may not behave in the same way in both testing groups over time, because raters might change their scoring standards from one time to the next. Thus use of common CR items, which are not strictly equivalent, would lead to erroneous results (Kim, Walker, & McHale, 2010b; Tate, 1999). Some practitioners have suggested using MC items as anchors to adjust for differences in difficulty among test forms containing CR items (e.g., Baghi, Bent, DeLain, & Hennings, 1995; Ercikan et al., 1998). Evidence suggests, however, that using an all-MC anchor with tests made up of CR items will lead to biased equating results (Kim & Kolen, 2006; Kim, Walker, & McHale, 2010a; Li, Lissitz, & Yang, 1999), possibly because the MC and CR items may measure somewhat different constructs (Bennett, Rock, & Wang, 1991; Sykes, Hou, Hanson, & Wang, 2002). For those limitations, many testing programs carry out routine statistical procedures (e.g., monitor raters' scoring behaviors or item difficulty) instead of equating in an attempt to maintain score comparability over forms and administrations.

The *TOEIC®* Speaking Test

The *TOEIC®* tests are English language proficiency tests for people whose native language is not English. The TOEIC Speaking test is intended to measure the test taker's ability to communicate in spoken English in daily life and in the workplace. The test consists of 11 items, representing six types of speaking tasks, requiring about 20 minutes to complete. The type of task and rating scale are presented in Table 1. For security reasons, all of the TOEIC Speaking test forms include newly developed items only, and thus no common CR items exist across any forms.

Table 1

Test Specifications of TOEIC Speaking Test

Item	Task	Rating scale
1–2	Read a text aloud	Intonation: 0–3; Pronunciation: 0–3
3	Describe a picture	0–3
4–6	Respond to questions	0–3
7–9	Respond to questions using information provided	0–3
10	Propose a solution	0–5
11	Express an opinion	0–5

The scaled scores of the TOEIC Speaking test range from 0 to 200 in increments of 10. The comparability of the scores across forms of the TOEIC Speaking test is mainly controlled through consistent item development and scoring. Because it is often difficult to achieve these conditions constantly in practice, however, the TOEIC program routinely exercises additional statistical checks to enhance the score comparability across forms.

Purpose

The major purpose of this study is to assess the effectiveness of the current practice of applying a single scale score conversion to all new editions of the test. To that end, a comparison of the scores resulting from a linking design and the current practice was made. Score conversions based on the NEAT design through TOEIC Listening test scores (external MC anchor) were derived from 30 operationally administered forms of the TOEIC Speaking test. The conversions resulting from a conventional linking procedure were then compared to the operational conversion resulting from the current practice to compute score differences resulting from two different procedures. In practice, the true relationship between any two TOEIC Speaking test forms is unknown, and thus this comparison cannot lead to a definitive conclusion as to which procedure is a better choice for the TOEIC Speaking test. The magnitude of the score differences between the two procedures could be used as a gauge to assess the effectiveness of the current practice.

Method

Data

For this study, test takers' records were gathered from the TOEIC Speaking test forms that had been administered between February 2014 and November 2015. The 60 forms taken by a large number of test takers (e.g., more than 1,000) were designated as either a new form (30 forms) or a reference form (30 forms). None of these forms shared items with another, so the choice regarding which new form to link to which reference form was somewhat arbitrary. However, I attempted to mimic the real world by linking more recently administered forms to older forms. In general, the gap between the new and reference form administrations was 3–6 months.

Table 2 presents the descriptive statistics of the TOEIC Speaking test scores and the TOEIC Listening test anchor scores in both new and reference form groups at each administration. The correlations between the TOEIC Speaking test (CR) scores and TOEIC Listening test (MC) anchor scores are also included in Table 2. As the anchor standardized mean differences (SMDs) indicate, the reference group was more able than the new group on 22 forms out of 30 (e.g., SMD ≤ −.1). The size of the difference between the MC anchor means of the new and reference form groups varied from −.36 to .24 in standard deviation units. The correlations between the CR score and MC anchor scores ranged from .57 to .71 ($M = .63$, $SD = .03$). As expected, the anchor correlations are not as high as the anchor correlations usually observed in the MC-only test equating using an internal anchor ($r = .80$ or higher).

Table 2

Means and Standard Deviations of the TOEIC Speaking Test and Listening Test Scores in the New and Reference Form Groups

Speaking form	NF *N*	NF Speaking, *M* (*SD*)	NF anchor, *M* (*SD*)	NF *r*	RF *N*	RF Speaking, *M* (*SD*)	RF anchor, *M* (*SD*)	RF *r*	SMD (new-ref)
1	1,321	130 (20.7)	387 (69.3)	0.66	1,666	124 (21.3)	402 (68.9)	0.62	−0.22
2	1,482	126 (21.0)	393 (66.9)	0.63	2,053	129 (22.4)	396 (68.0)	0.64	−0.03
3	1,350	130 (21.0)	389 (66.1)	0.62	1,503	126 (21.2)	404 (66.5)	0.63	−0.22
4	1,695	129 (22.9)	395 (68.9)	0.62	1,651	126 (22.8)	404 (69.7)	0.66	−0.13
5	1,868	125 (21.3)	395 (67.0)	0.61	1,731	124(22.6)	411 (64.0)	0.62	−0.24
6	1,337	129 (20.0)	396 (68.8)	0.60	1,719	132 (20.2)	408 (66.8)	0.61	−0.18
7	1,569	127 (21.0)	396 (69.7)	0.63	1,546	121 (22.6)	404 (70.4)	0.62	−0.10
8	1,959	128 (22.1)	396 (67.8)	0.65	1,540	130 (21.4)	406 (68.9)	0.61	−0.15
9	1,986	128 (20.0)	398 (68.0)	0.60	1,427	119 (21.9)	396 (72.7)	0.63	0.03
10	1,596	120 (21.3)	394 (69.5)	0.63	1,334	129 (22.1)	409 (67.7)	0.64	−0.22
11	1,588	129 (20.7)	395 (67.5)	0.62	1,814	123 (21.5)	406 (68.8)	0.63	−0.17
12	1,353	124 (22.4)	396 (68.2)	0.67	1,384	127 (21.4)	409 (66.3)	0.65	−0.19
13	1,938	126 (20.8)	393 (70.4)	0.63	1,306	125 (20.8)	408 (65.7)	0.60	−0.22
14	2,057	126 (21.3)	395 (66.8)	0.62	1,328	121 (23.4)	405 (69.2)	0.66	−0.14
15	1,939	128 (21.1)	391 (70.0)	0.66	1,472	127 (22.8)	398 (70.9)	0.63	−0.11
16	1,671	128 (19.7)	396 (68.7)	0.60	1,345	121 (21.3)	402 (67.1)	0.57	−0.09
17	1,805	127 (21.3)	396 (68.0)	0.60	1,419	127 (22.0)	397 (70.6)	0.62	−0.01
18	1,203	124 (20.6)	386 (69.9)	0.58	1,674	121 (21.9)	396 (70.4)	0.65	−0.13
19	1,371	125 (21.4)	385 (71.0)	0.62	1,202	130 (21.0)	410 (65.5)	0.65	−0.36
20	1,362	124 (22.5)	388 (72.7)	0.62	1,205	121 (23.6)	396 (75.1)	0.64	−0.10
21	1,203	130 (20.9)	408 (65.8)	0.57	1,325	128 (22.3)	401 (74.6)	0.66	0.09
22	1,139	127 (19.8)	387 (67.4)	0.61	1,223	130 (21.9)	405 (68.0)	0.62	−0.27
23	1,276	127 (21.3)	398 (67.0)	0.64	1,213	125 (20.4)	397 (67.9)	0.62	0.02
24	1,307	121 (20.7)	394 (71.0)	0.63	1,320	135 (18.7)	411 (64.7)	0.56	−0.26
25	1,419	129 (21.6)	392 (68.8)	0.64	1,178	117 (24.5)	393 (72.8)	0.66	−0.01
26	1,297	126 (20.2)	394 (67.7)	0.59	1,286	127 (22.9)	401 (70.8)	0.62	−0.11
27	1,332	121 (22.2)	392 (69.7)	0.63	1,430	127 (21.3)	404 (66.1)	0.59	−0.18
28	1,256	124 (21.8)	390 (73.2)	0.63	1,412	125 (21.7)	398 (69.2)	0.64	−0.10
29	1,331	130 (21.7)	398 (71.8)	0.63	1,327	128 (22.6)	407 (68.6)	0.59	−0.13
30	1,498	127 (22.2)	403 (69.8)	0.63	1,282	125 (25.0)	386 (75.6)	0.71	0.24

Note. r = correlation between TOEIC Speaking and Listening test (anchor) scores. Not all test takers' anchor scores were available at the time of linking. NF = new form linking group, RF = reference form linking group, SMD = standardized mean difference between the MC anchor scores (new group minus reference group).

Procedure

For each form, score linking through the TOEIC Listening test scores was conducted using the test takers whose TOEIC Listening test scores were available at the time of linking.[1] On average, TOEIC Listening test scores were available for approximately two-thirds of test takers (*M* = 67%, range = 50%–78%).[2] In the NEAT design with external MC anchor, the chained equipercentile (Kolen & Brennan, 2004) method was used to produce the scaled sore conversion.[3] The resulting conversion was then applied to every test taker in the new form group to obtain his or her TOEIC Speaking test scaled score. This scaled score is the scaled score the test taker would have received if the linking design with an external MC anchor had been implemented in the operational setting. I computed the difference between the new scaled score based on the MC linking design and the test taker's operational scaled score based on the current practice. Then I computed the percentage of test takers whose scaled scores were categorized as follows: no difference, a 10-point difference, a 20-point difference, and so on. In addition, the means and standard deviations from all test takers in the new form group were calculated based on the two sets of scaled scores, along with the SMDs (linking minus operational in the new group).

Results

Table 3 presents the scaled score difference (external linking conversion minus operational conversion) results associated with the 30 TOEIC Speaking test forms investigated in this study. As shown, the results are highly consistent across all the 30 forms, indicating similar differences. The differences were primarily within the range of −10 to +10. For 14 of the forms (Forms 2, 6, 8, 12, 15, 17, 19, 21–23, 26–28, and 30), the scaled scores remained unchanged for more than two-thirds of the test takers. On 10 forms (Forms 1, 3, 5, 7, 9, 11, 14, 16, 18, and 25), more than 85% of the test takers' scaled scores decreased by 10 points when the linking conversion was applied. The linking conversion, however, led to a 10-point increment for 95% of the test takers who took Form 24. The differences on the remaining forms were rather evenly distributed across two adjacent categories, either −10 to 0 (Forms 4, 13, 20, and 29) or 0 to +10 (Form 10). For those five forms, approximately 33%–50% of the test takers retained the same scaled scores, whereas approximately 50%–63% of the test takers' scaled scores changed by 10 points. Very few test takers' score differences were greater than 10 points.

Table 3

Differences Between External Anchor Linking Scaled Scores and Operational Scaled Scores for the 30 Test Forms

Form	Below –20	–20	–10	0	10	20	Above 20
1	–	0.7	96.7	2.2	0.4	–	–
2	–	–	–	98.8	1.2	–	–
3	–	0.1	86.2	13.7	–	–	–
4	–	–	49.2	50.8	–	–	–
5	–	–	97.5	2.5	–	–	–
6	–	–	0.1	99.9	–	–	–
7	–	0.1	85.1	14.8	–	–	–
8	–	–	–	99.1	0.9	–	–
9	0.2	–	99.6	0.3	–	–	–
10	–	–	–	41.4	58.6	–	–
11	0.3	–	94.0	5.7	–	–	–
12	–	0.4	7.0	90.8	1.7	0.1	–
13	–	–	45.9	53.2	0.7	0.1	0.1
14	–	–	98.8	1.1	0.2	–	–
15	0.1	–	29.0	70.9	–	–	–
16	0.5	–	99.1	0.4	–	–	–
17	–	0.3	2.0	97.7	–	–	–
18	–	–	89.6	10.3	0.1	–	–
19	–	–	0.7	99.3	–	–	–
20	–	–	56.2	43.7	0.1	–	–
21	0.3	0.2	0.4	99.2	–	–	–
22	–	–	–	99.1	0.8	0.1	–
23	–	–	–	99.7	0.3	0.1	–
24	–	–	–	3.8	95.5	0.6	0.2
25	–	–	99.4	0.6	–	–	–
26	–	–	–	95.3	4.7	0.1	–
27	–	–	–	98.8	1.2	–	–
28	–	–	7.4	92.6	–	–	–
29	–	–	62.7	33.0	4.1	0.2	–
30	–	–	–	100.0	–	–	–

Table 4 presents the means and standard deviations of the operational and linked scaled scores computed from all test takers in the new form group. The total group also includes the test takers who were excluded from the linking process because their anchor scores were not available at the time of linking. The direction of mean differences between the two conversions was matched with the difference direction shown in Table 3. As expected, the SMDs of the score differences between the current practice and MC anchor linking was close to half of a standard deviation (around 10 points) under the nine form cases (Forms 1, 5, 9, 11, 14, 16, 18, 24, and 25). For many forms used in this study, the MC–anchor linking produced lower means than did the current practice. The largest mean difference between the two conditions was 10 points, which is slightly lower than one-half of a standard deviation.

Table 4

Means and Standard Deviations of the Scaled Scores Derived From Chained Equipercentile Linking and Current Operational Practice

Speaking form	*N*	Current practice *M*	Current practice *SD*	Chained equipercentile *M*	Chained equipercentile *SD*	SMD
1	2,621	130	22.9	120	22.0	−0.44
2	2,419	125	22.1	125	22.4	0.00
3	2,310	130	23.3	121	21.3	−0.39
4	2,629	128	24.0	123	22.8	−0.21
5	2,887	125	22.3	115	21.8	−0.44
6	1,885	128	21.2	128	21.2	0.00
7	2,258	126	21.7	118	22.4	−0.39
8	2,794	128	22.9	128	22.7	0.00
9	2,859	127	20.9	117	21.4	−0.47
10	2,394	119	22.5	125	20.9	0.27
11	2,403	129	22.0	119	21.5	−0.44
12	2,031	123	23.7	122	21.8	−0.03
13	2,641	125	21.9	121	23.5	−0.20
14	2,730	125	22.1	115	22.0	−0.45
15	2,717	128	22.3	125	24.6	−0.13
16	2,226	127	20.9	117	21.5	−0.48
17	2,432	127	22.3	126	22.2	−0.01
18	1,660	123	21.9	114	22.7	−0.40
19	1,916	125	22.6	125	22.4	0.00
20	1,902	124	23.2	118	21.0	−0.25
21	1,545	128	22.6	128	23.3	−0.01
22	2,312	127	22.2	127	21.9	0.00
23	2,016	125	22.3	125	22.2	0.00
24	1,708	120	21.8	130	20.8	0.46
25	2,466	128	23.7	118	23.9	−0.42
26	1,982	126	22.0	126	22.6	0.02
27	1,973	122	23.1	122	22.8	0.01
28	1,990	122	23.9	121	24.5	−0.03
29	2,047	128	23.8	122	23.9	−0.24
30	2,013	126	23.2	126	23.2	0.00

Note. SMD = standardized mean difference.

To display the score region where most test takers were located, the scaled score distributions of the new form group accumulated over the 30 forms/administrations are presented in Figure 1. In the figure, one distribution was associated with the relative frequency from the current practice, and another was associated with the relative frequency from the MC–anchor linking. Figure 2 plots the differences from the operational conversion across the scaled score region from the 1st percentile to the 99th percentile in the new form group. There were 30 difference lines associated with the 30 forms, and the dotted lines at ±10 indicate half of a standard deviation. The differences were generally smaller than 10 points across the score region where most test takers were located. However, the difference line associated with Form 25 (solid blue line) was beyond the ±10 band. In Form 25, the new form group was as able as the reference group, as the SMD of the TOEIC Listening test score indicates (SMD = −.01). However, the TOEIC Speaking test score of the new form group (*M* = 129) was much higher than that of the reference form group (*M* = 117), leading to the SMD of .53. Because the TOEIC Speaking test form difference in difficulty was adjusted through the TOEIC Listening test scores under the MC–anchor linking design and the reference form group did as well on the TOEIC Listening test, the new TOEIC Speaking test form appeared much easier in the process of score linking.

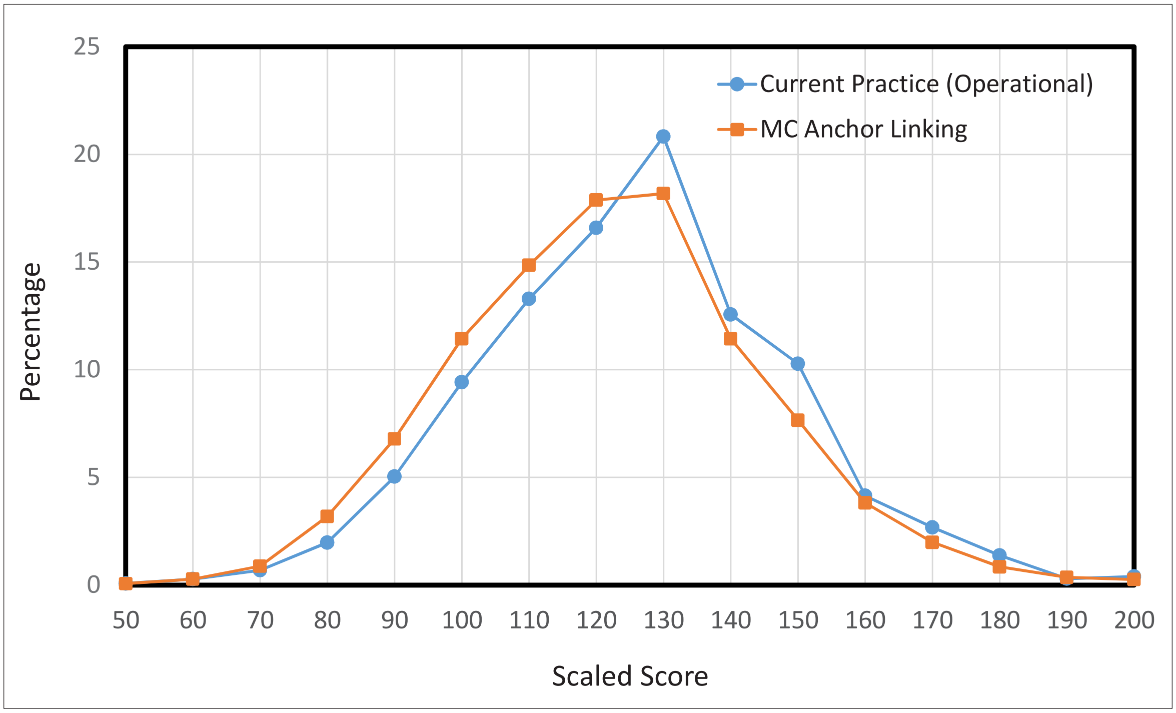


Figure 1**. Percentage distribution of the scaled scores on the entire new form group.**

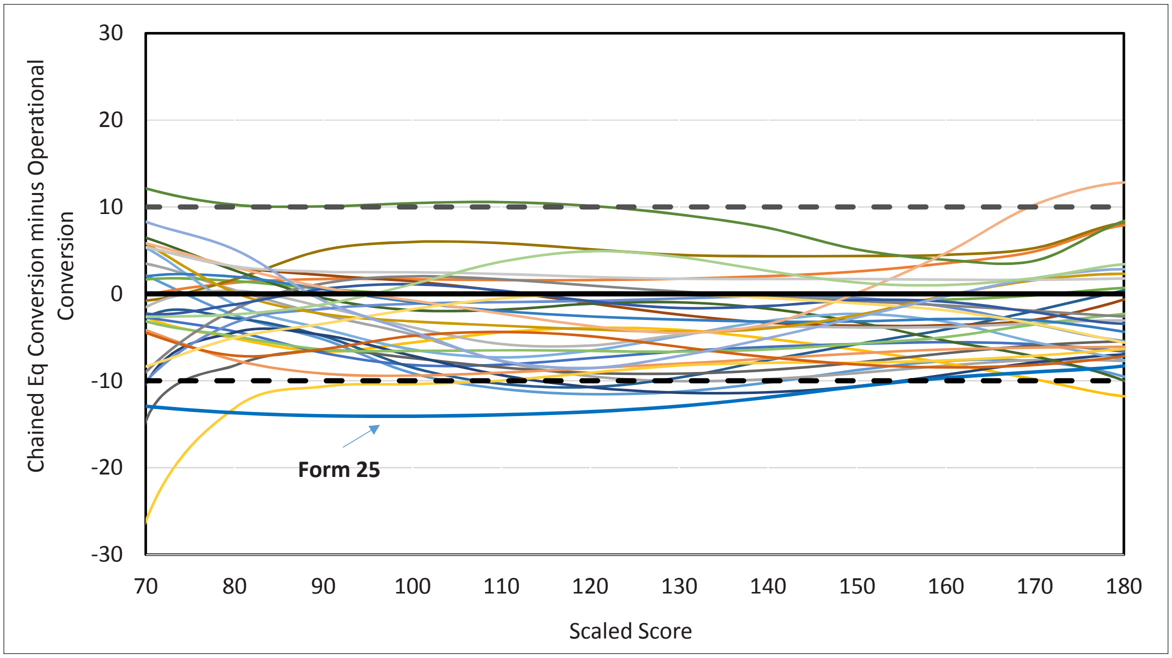


***Figure 2*. Difference plots between chained equipercentile conversion and operational conversion.**

Discussion

Owing to security concerns, the TOEIC program uses new editions of the TOEIC Speaking test (which include only newly developed CR items) for every test administration. To ensure score comparability over different forms, some form of equating is desirable. When test forms consist of CR items only, however, score equating through the conventional design (e.g., NEAT) is not always feasible because of a lack of proper common items. Although the *TOEIC*® Listening and Reading test can be used as an external MC anchor to link the TOEIC Speaking test scores, using an external MC anchor can potentially be problematic in that anchors consisting of external MC items alone may not adequately represent the CR test content and thus may not produce satisfactory links. In addition, because not all test takers' external MC anchor scores are available at the time of linking, it is often questionable how well a linking sample represents the entire group of test takers. Owing to various practical limitations (e.g., no common CR items, low volume, operational demands for reporting scores in a short time), the current practice of the TOEIC Speaking test is based on the assumptions that forms are sufficiently similar in difficulty and that raters use the same scoring standard, as is intended.

The purpose of the study was to compare the current practice to a procedure by which scores are derived from an MC external anchor linking design. The external MC anchor linking may not be optimal unless the correlation between MC and CR is substantially high. Even so, some testing programs use this approach operationally as a method to produce comparable CR scores over the forms. Because the external anchor scores were available for many of the TOEIC Speaking test takers and the correlations between MC and CR were moderate, the conversions derived from the external MC anchor linking were used to assess the effectiveness of the current practice in this study.

The most interesting feature of the study's results was the comparability of scores derived from different procedures. The scores derived through the external MC anchor design were generally comparable to the current reported scores based on the consistency of both item difficulty and scoring. Although some test takers' scaled scores changed by as much as 10 points, this change is comparable to measurement error, as the standard error of measurement of the TOEIC Speaking test is approximately 10–11. The present findings indicate that adopting external MC anchor design in the operational setting would have little practical impact and may therefore be unnecessary. This study suggests that the psychometric benefits that may be achieved by replacing the current practice with an external MC anchor linking may be negligible. Given the moderate correlation between TOEIC Speaking test and TOEIC Listening test scores, the linkage between the two sets of scores will be weak, thus yielding minimal benefit to improved equivalence across forms.

The TOEIC program uses several strategies in an attempt to maintain score comparability over different forms and administrations. Test developers exercise their expertise to assemble the TOEIC Speaking test forms to be as parallel as possible. Because difficulty levels of CR items are determined as a function of item–rater combinations, however, CR scoring, either stringent or lenient, may change the level of item difficulty as well. Often slight scoring shifts over time are unavoidable. In the TOEIC program, all raters are thoroughly trained in the use of the rating rubrics to enhance the consistency of the ratings and, therefore, the reliability of the TOEIC Speaking test scores. Raters are required to pass a certification test, consisting of a number of benchmark responses for which consensus ratings exist, prior to starting the rating of operational responses. Expert raters provide ongoing monitoring of their ratings and are also available to provide support and feedback as needed.

The TOEIC program conducts comprehensive postadministration analyses on every administration to (a) evaluate the quality of the ratings, (b) assess the statistical/psychometric properties of each item, and (c) monitor test takers' performance over time. For example, rating consistency per item, indicated by the correlation between the two ratings and the weighted kappa coefficient (Fleiss, Cohen, & Everitt, 1969), is calculated based on the 10%–15% of double scoring data. Using the double scoring data, rating agreement per item, indicated by the percentage of same rating (no difference), the percentage of adjacent ratings (1 point difference), and the percentage of discrepant ratings (more than 1 point difference), is also examined to ensure raters' scoring consistency. Such analyses are particularly helpful in evaluating the need for additional rater training. Furthermore, descriptive statistics of each item and psychometric properties of each item type are assessed against the historical data accumulated over several years. Empirical evidence, such as historical charts and data, help to inform judgment regarding the current forms' performance. Such monitoring and analyses provide relevant empirical evidence for scoring stability and test fairness.

References

Baghi, H., Bent, P., DeLain, M., & Hennings, S. (1995, April). *A comparison of the results from two equatings for performance-based student assessments.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Bennett, R. E., Rock, D. A., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement*, *28*, 77–92. https://doi.org/10.1111/j.1745-3984.1991.tb00345.x

Ercikan, K., Schwarz, R., Julian, M. W., Burket, G. R., Weber, M. W., & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response test item type. *Journal of Educational Measurement*, *35*, 137–154. https://doi.org/10.1111/j.1745-3984.1998.tb00531.x

Fleiss, J. L., Cohen, J., & Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin, 72*, 323–327. https://doi.org/10.1037/h0028106

Kim, S., & Kolen, M. J. (2006). Robustness to format effects of IRT linking methods for mixed-format tests. *Applied Measurement in Education, 19*, 357–381. https://doi.org/10.1207/s15324818ame1904_7

Kim, S., Walker, M. E., & McHale, F. (2010a). Comparisons among designs for equating mixed-format tests in large scale assessments. *Journal of Educational Measurement, 47*, 36–53. https://doi.org/10.1111/j.1745-3984.2009.00098.x

Kim, S., Walker, M. E., & McHale, F. (2010b). Investigation the effectiveness of equating designs for constructed response tests in large scale assessment. *Journal of Educational Measurement, 47*, 186–201. https://doi.org/10.1111/j.1745-3984.2010.00108.x

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer. https://doi.org/10.1007/978-1-4757-4310-4

Li, Y. H., Lissitz, R. W., & Yang, Y. N. (1999, April). *Estimating IRT equating coefficients for tests with polytomously and dichotomously scored items.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Muraki, E., Hombo, C. M., & Lee, Y.-W. (2000). Equating and linking of performance assessments. *Applied Psychological Measurement, 24*, 325–337. https://doi.org/10.1177/01466210022031787

Sykes, R. C., Hou, L., Hanson, B., & Wang, Z. (2002, April). *Multidimensionality and the equating of a mixed-format math examination.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Tate, R. L. (1999). A cautionary note on IRT-based linking of tests with polytomous items. *Journal of Educational Measurement*, *36*, 336–346. https://doi.org/10.1111/j.1745-3984.1999.tb00560.x

Notes

[1] The TOEIC Listening and Reading Comprehension (LC & RC) MC test scores can be used as an external MC anchor to link the Speaking score. Because the LC section scores showed slightly higher correlations with the Speaking scores than did either RC or LC & RC combined, the LC section score was used as an anchor in the study. However, the same trend appeared in both (RC only and LC & RC combined) anchor conditions.

[2] There exists substantial overlap between the TOEIC Speaking test population and the TOEIC LC & RC test population. However, not all TOEIC Speaking test takers take the TOEIC LC & RC test, and vice versa.

[3] Frequency estimation equipercentile (often called poststratification equipercentile [PSE]; Kolen & Brennan, 2004, pp. 135–143) was also used to produce the conversion table for each of the 30 forms. Because both chained equipercentile and PSE produced very similar results, the PSE results were not presented in this report for simplicity. The frequency estimation equipercentile results are available on request.

Compendium Study

Articulating and Evaluating Validity Arguments for the *TOEIC*® Tests

Jonathan Schmidgall

How can you determine whether a test is suitable for the purpose for which it was designed? This fundamental question of validity has preoccupied test developers, researchers, and score users for decades. In the first TOEIC® program compendium, Powers (2010) provided a clear, accessible overview of validity that focused on two critical aspects: whether scores mean what they are supposed to mean and whether a test fulfills its designated purpose. Subsequently, consensus-based professional standards have come to embrace the view that test developers must convince stakeholders (i.e., anyone affected by the test) that the intended use of a test is appropriately supported or justified (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014; Educational Testing Service [ETS], 2015; Newton, 2012). This view is formalized in the argument-based approach to justifying test use.

The argument-based approach to justifying test use consists of a comprehensive set of claims made by the test developer. These claims are supported or undermined by evidence, which may include documentation from the test development process and ongoing research. Through an examination of the test developer's claims and the evidence to support them, various stakeholders may arrive at a global evaluation of whether the intended use of the test has been adequately justified. Different stakeholders may value different types of evidence; for example, teachers may be primarily concerned about evidence that the test has a positive impact on teaching and learning, whereas score users may be more concerned about the outcomes of decisions based on the test.

The purpose of this report is to provide an accessible introduction to the argument-based approach, its implementation for TOEIC tests, and the perceived benefits for stakeholders. I begin this report with a brief overview of the assessment use argument (AUA), a prominent argument-based approach to validation (Chapelle & Voss, 2014). Next, I detail the approach that has been used to articulate fully specified validation arguments for TOEIC tests. This approach incorporates evidence from a variety of sources, including test documentation, monitoring activities, and research. Finally, I provide an overview of the two primary ways in which the validation arguments are used: to influence the research agenda and to communicate with stakeholders. The report concludes with a brief discussion of the benefits of this approach, as well as several suggestions for extending it.

The Assessment Use Argument

An AUA is "a conceptual framework for guiding the development and use of a particular language assessment, including the interpretations and uses we make on the basis of the assessment" (Bachman & Palmer, 2010, p. 99). The framework is structured as a hierarchical set of claims made by the test developer that specifies how test scores should be interpreted and used to make decisions. The AUA draws upon well-established conceptualizations of validity, including Messick's (1989) progressive matrix, and a formalized argument structure in which transformations of data represent inferences that support claims (Toulmin, 2003).

Although this approach may seem abstract, the use of argument structure requires test developers to make statements about the expected measurement quality and uses of a test explicitly and systematically. In the AUA framework, these statements take the form of four high-level claims. These four claims explicitly link test-taker performance to test scores, scores to interpretations about test-taker ability, interpretations about ability to decisions, and finally, decisions to the consequences that follow. Thus, the AUA framework takes the general form shown in Figure 1.

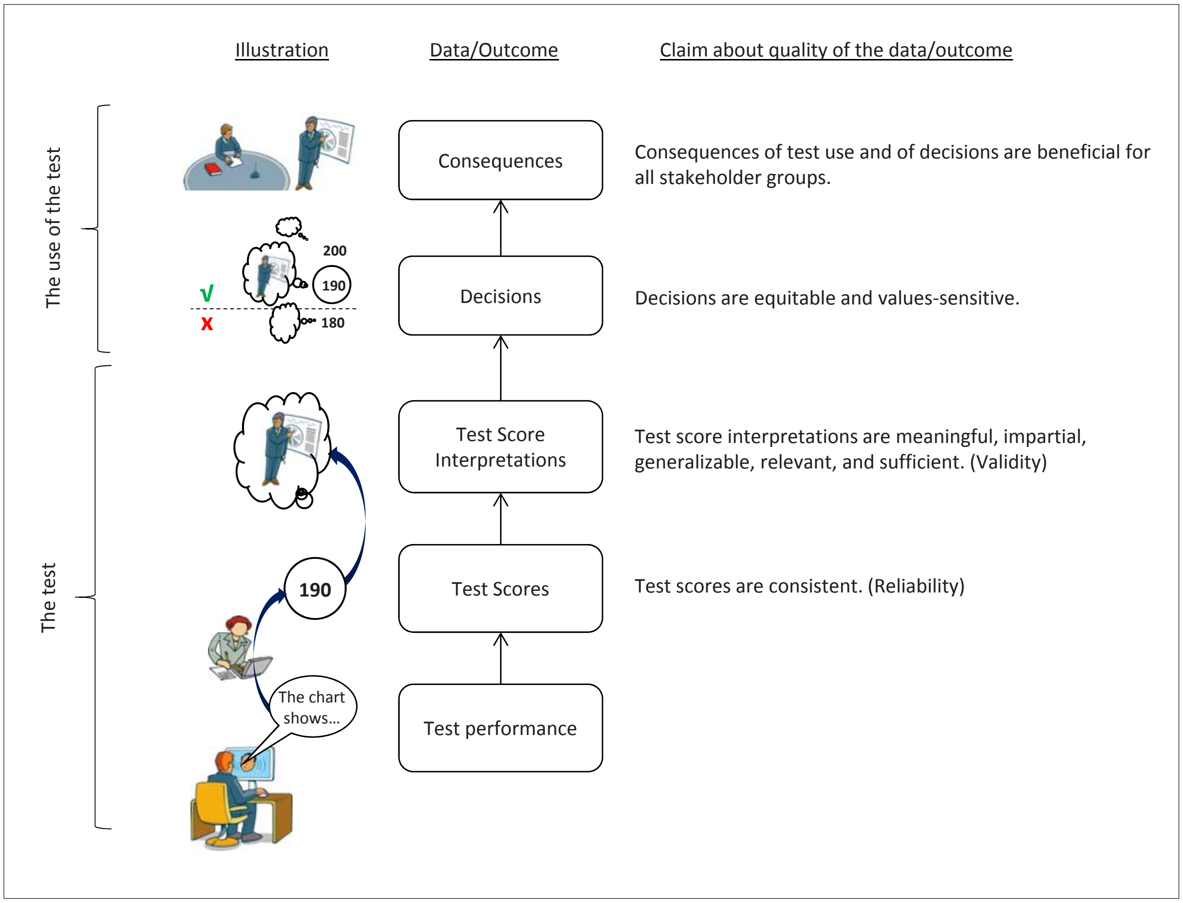


Illustration
Data/Outcome
Claim about quality of the data/outcome
The use of the test
The test
200
190
√
X
180
190
The chart shows…
Consequences
Decisions
Test Score Interpretations
Test Scores
Test performance
Consequences of test use and of decisions are beneficial for all stakeholder groups.
Decisions are equitable and values-sensitive.
Test score interpretations are meaningful, impartial, generalizable, relevant, and sufficient. (Validity)
Test scores are consistent. (Reliability)

***Figure 1*. Data and claims in an assessment use argument.**

Each box in Figure 1 represents data related to a test or its use. Data are transformed into an outcome through an inference, represented by arrows. An outcome is expected to have particular qualities, as specified in the corresponding claim. The outcome also serves as data for the next step up the inferential ladder, linking test performances to consequences in the real world.

As seen in Figure 1, the foundational data for all subsequent inferences are the test performance. The illustration shows what this might look like for a computer-delivered speaking test: a test taker speaking in response to questions. The test taker's speaking responses are the test performance, which is transformed into a test score (e.g., 190) through a rating procedure. Figure 1 shows that a claim should be made about the qualities of test scores: They are consistent. Aspects of the testing procedure that are unrelated to the test taker's ability (e.g., test forms, test administrations, and raters) should not unduly influence scores, and so, scores are expected to be consistent across these aspects.

Test scores are then transformed into interpretations about test takers' abilities. These interpretations should be meaningful, impartial, generalizable, and appropriate (relevant and sufficient) for the decisions to be made. In the case of our illustrative speaking test, a score of 190 may, for instance, imply that the test taker has a high level of speaking proficiency and would be expected to present familiar information orally with a high degree of comprehensibility.

Test score interpretations are typically used to make decisions. Consequently, the score interpretation is transformed into a specific decision category or contributes in some way to a broader decision-making process. As an outcome, decisions should be fair and sensitive to the values of the decision maker. In the illustrative example, the test taker's score of 190 exceeds the benchmark set by the decision maker, which contributes to a hiring decision. The benchmark set by the decision maker (190 or higher) is relatively high, which reflects the decision maker's need for a high level of speaking proficiency and the desire to minimize false-positive decision errors.

Ultimately, the use of a test and the decisions that ensue produce an outcome: real-world consequences. Consequences should be beneficial for stakeholder groups; otherwise, the effectiveness of the test (or decision-making process) may be in question. In the illustrative example, the hypothetical test taker who was hired is able to give effective oral presentations on the job in the real world, benefitting himself and his employer.

This approach provides a coherent way to relate the traditional measurement concepts of reliability and validity, and it treats validation as the act of providing evidence to support claims rather than a specific quality of a test (e.g., criterion-related validity, construct validity, etc.). This approach is also comprehensive in that it captures a wide range of desirable aspects of a test that have historically been collapsed into imprecisely defined categories (e.g., consequential validity, response validity, construct validity, content validity). Schmidgall and Choi (2011) reviewed language-testing research articles, categorizing their research questions using traditional categories of validity (e.g., construct validity) and claims and warrants in an AUA. They found that individual research questions in the publications reviewed could usually be linked to specific claims in an AUA and that traditional categories of validity may be scattered across levels in an AUA. This observation suggests that the use of argument-based approaches such as the AUA—as opposed to traditional categories of validity—may help clarify the implications of validity research.

Both decision makers and test developers share responsibility for justifying assessment use. Test developers are expected to provide evidence to support the claim that test scores are consistent and that scores may be used to make interpretations about test-taker abilities. Decision makers need evidence that decisions are values-sensitive and equitable and that consequences of decisions are beneficial. Unfortunately, decision makers may lack the expertise or resources needed to provide adequate backing for these claims, such as the ability to conduct a benchmarking or standard-setting study. Test developers often have this expertise but may not be aware of features of a specific decision-making context that may influence how decisions are made, such as the level of language proficiency required or the decision maker's tolerance for different types of decision errors. Consequently, an AUA may be enhanced through collaboration between decision makers and test developers. Decision makers can utilize a test developer's expertise to support benchmarking studies that promote values-sensitive and equitable decision-making, whereas test developers can receive feedback on the test's effectiveness.

The structure of an AUA provides a basis for a comprehensive justification of test use that links real-world concerns about decisions and their consequences with the traditional concerns of test developers: reliability and validity. As a comprehensive list of claims and evidence, an AUA can be used to identify weaknesses in the overall argument for test use and prioritize research or test development projects. For example, Wang, Choi, Schmidgall, and Bachman (2012) reviewed the Pearson Test of English Academic and, based on documents obtained from the test developer, produced a detailed AUA that explicitly specified claims regarding test use. The review generated a number of specific recommendations that could be used to inform research.

As a simple hierarchical set of claims, an AUA can be used as a communication tool that illustrates the key issues that determine important qualities of the usefulness of a test, including fairness, impact, reliability, and validity. The concerns of individuals and stakeholder groups vary, and one of the challenges for research is addressing these concerns in a coherent manner while enhancing the "assessment literacy" of stakeholders. For example, stakeholders may be concerned about the following issues:

- Score consistency (How can you make sure that all raters follow the scoring guides?)
- Interpretation of scores (When we calculate criterion-validity, who or what is the criterion?)
- Decisions based on these interpretations (What are the cut scores in other institutions?)
- Consequences of test use (How have TOEIC tests been helpful for job-seekers?)
- Test use that relates to a number of these issues (How can recruiters know that TOEIC scores meets the needs of the market?)

By delivering versions of an AUA oriented toward specific stakeholder groups, a test developer with a strong research program may be able to help stakeholders answer their questions and become more sophisticated consumers of assessment products.

Constructing Assessment Use Arguments for *TOEIC*® Tests

The previous section discussed how the AUA can be used to specify four high-level claims about the measurement quality and intended use of a test. A fully specified AUA also contains a large number of warrants: statements made in support of each high-level claim. Bachman and Palmer (2010) elaborated a reasonably exhaustive list of potential warrants in their book-length description of the AUA. This general, idealized version of an AUA was designed to incorporate the accumulated knowledge of testing professionals as reflected in influential publications on validity and validation (e.g., AERA, APA, & NCME, 1999; Kane, 2006; Messick, 1989). When constructing an AUA, this is a logical place to start: adapting the generalized AUA to a specific context. Figure 2 illustrates the process the TOEIC research program used to create fully specified AUAs for TOEIC tests, beginning with the elaboration of idealized AUAs.

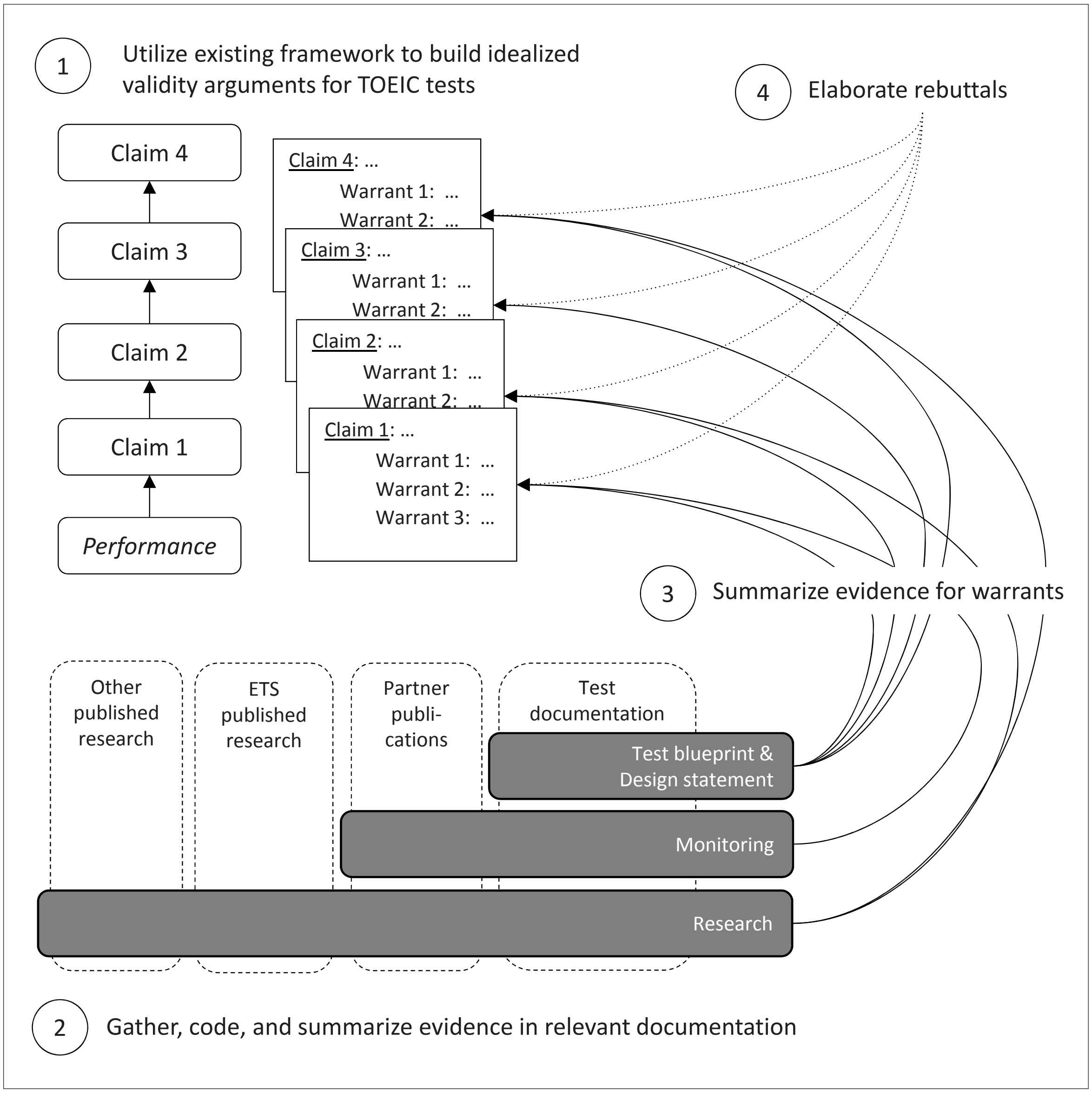


***Figure 2.* Overview of the process for creating TOEIC test assessment use arguments.**

Step 1: Articulate Claims and Warrants

As shown in Figure 2, the first step in the process was to utilize the existing AUA framework as proposed by Bachman and Palmer (2010) to build idealized validation arguments for the TOEIC tests. This idealized version contained all of the claims (and supporting warrants) that test developers and score users might want to see supported. Essentially, it represented a best-case scenario for a test developer interested in adhering to best practices in measurement with multiple warrants supporting each of the four high-level claims summarized earlier (see Figure 1). Although testing occurs in the real world in which there are important trade-offs between reliability, validity, and practicality (i.e., cost and convenience) that must be carefully considered, the idealized AUA represents an aspirational set of claims and warrants for a test developer to aim to support.

One of the challenges of creating AUAs for TOEIC tests is the fact that the tests are used for multiple purposes. For example, TOEIC tests are intended to facilitate hiring, placement, promotion, and progress decisions (e.g., ETS, 2013, p. 27). Although an AUA should be articulated for each intended use of the test, we initially constructed AUAs aligned with one particular use or decision in mind: hiring. This use was chosen based on feedback from key stakeholder groups and is an example of a particularly high-stakes use of TOEIC tests.

Idealized AUAs were constructed for TOEIC Speaking, Writing, Listening and Reading, and *TOEIC Bridge*^{TM} tests, with adaptations to Bachman and Palmer's (2010) generalized AUA structure based on the particular design of each test and the intended use. For example, the TOEIC Listening and Reading tests do not use human raters, so warrants about inter- or intrarater consistency are not relevant to support the claim that TOEIC Listening and Reading scores are consistent. As another example, TOEIC tests are intended to be used with other criteria to facilitate hiring decisions (see ETS, 2013, p. 26), so a warrant that "TOEIC score interpretations provide sufficient information to facilitate hiring decisions" was not included in the AUAs constructed for this particular use.

Steps 2 and 3: Collect Evidence and Relate It to Claims and Warrants

The second step in the process illustrated in Figure 2 involved the collection and synthesis of evidence from the test design process, ongoing statistical and procedural monitoring, and research activity. Evidence from the test design process included documentation that was produced as part of the evidence-centered design process (see Hines, 2010; Schedl, 2010) and documentation that described test administration and scoring procedures. This included the initial justification for the definitions of the abilities measured by each test, as well as the item and test specifications. This documentation was synthesized and used to support a variety of warrants in the AUA, including warrants to support claims about the consistency of scores, plus the meaningfulness, impartiality, and generalizability of score interpretations. This documentation was produced entirely by the test developer (ETS), and much of it is confidential (e.g., item and test specifications).

The second type of evidence synthesized and summarized in each AUA derived from ongoing statistical and procedural monitoring. Most of this documentation was produced by ETS and includes statistical monitoring such as the stability of scores across test forms and administrations and potential changes in the demographic characteristics of the test-taker population. Procedural monitoring occurred as well and potentially included feedback provided to the test developer by test takers, score users, and local partners on test administration, security, the use of scores.

The final type of evidence included in each AUA derived from research and review articles published by ETS, its partners, trade journals, or individual researchers. The first two edited volumes in the TOEIC program compendium included more than 20 papers, each of which contributed evidence to support various warrants in TOEIC tests AUAs. Additional research and practitioner publications were identified periodically through manual searches of journals in language assessment and keyword searches using Google Scholar. For example, periodic searches between June 2014 and June 2017 identified 113 publications that explicitly mentioned TOEIC tests, of which 76 were reviewed and coded for their relevance to TOEIC test AUAs. Publications were excluded when their mention of TOEIC tests was cursory or without consequence for an AUA; for example, Lawn and Lawn (2015) mentioned TOEIC tests as an example of an English language assessment.

The 76 publications identified as relevant to TOEIC test AUAs were reviewed and coded, and their findings or claims were incorporated as evidence (backing) or criticism (rebuttal) to a relevant warrant. Publications were coded based on the TOEIC program assessments to which they applied (TOEIC Reading, Listening, Speaking, Writing tests, TOEIC Bridge test, unidentified), their substantive focus (reliability, validity, test use, test review), and local context (e.g., Japan, Korea). The vast majority of publications pertained to the TOEIC Listening (78%) and TOEIC Reading (75%) tests; less than 10% pertained to the TOEIC Speaking, TOEIC Writing, or TOEIC Bridge tests. Publications varied in their substantive focal points, although many focused on issues pertaining to test use (51%) and validity (41%). Very few publications focused on reliability or score consistency (5%), which is not surprising; ideally, this quality of test scores should be examined under operational conditions and, thus, is primarily the responsibility of the test developer (ETS) and its local partners. The publications included in the review varied from unpublished graduate student papers to publications in international peer-reviewed journals, and around 13% of the publications were TOEIC test reviews. Most of the publications were published by researchers or practitioners in Japan (70%). Other local contexts included Korea (14%), Taiwan (8%), and China, Costa Rica, Indonesia, Thailand, and Vietnam (each less than 5%).

Step 4: Elaborate Rebuttals and Evaluate the Overall Plausibility of Assessment Use Argument

The final step in the process was to critically examine the existing evidence for each warrant and evaluate the overall plausibility of each AUA. Prior to this critical exercise, some potential rebuttals had already been documented based on the review of research and practitioner publications. For example, in a small-scale study of the impact of TOEIC Listening and Reading test use at a business school in Thailand, Apichatronajanakul (2011) found evidence of both positive and negative washback.

Evidence of negative washback constituted a potential rebuttal to the warrant that the consequences of using TOEIC Listening and Reading test would be beneficial to test takers, which could potentially undermine the overall claim that the consequences of TOEIC Listening and Reading test use are beneficial. When evidence for a particular warrant was mixed or mostly lacking, it underscored the need to consider the seriousness of existing or potential rebuttals and their impact on a broader claim about the measurement quality or use of the test.

Uses of Validity Arguments for *TOEIC*® Tests

The fully specified AUAs reflect a broad consideration of evidence to support uses of TOEIC tests and have been developed with two primary applications in mind. First, they have been used to help inform a research agenda for the TOEIC program. Research is critical for supporting claims about the measurement quality and intended use of tests, but all test developers have limited resources. Based on critical evaluations of fully specified TOEIC test AUAs, one area of research that the TOEIC program has pursued over the last several years has been focused on the uses of TOEIC tests and their potential impact on various stakeholder groups; several chapters in the present volume address these focal issues (e.g., "The Case of Taiwan: Perceptions of College Students About the Use of the *TOEIC*® Tests as a Condition of Graduation" by Hsieh, and "Insights Into Using *TOEIC*® Test Scores to Inform Human Resource Management Decisions" by Oliveri and Tannenbaum).

The other application of the fully specified AUAs has been the creation of simplified versions aligned with the needs of different stakeholders. Although the wording of claims and warrants in an AUA are designed to be accessible to nonexperts, reviewing and evaluating a fully specified AUA requires a significant investment on the part of the reader. Simplified versions allow the fully specified AUA to be condensed and adapted with a particular readership in mind. For example, all test programs at ETS are periodically audited to ensure their compliance with professional standards. This compliance includes adhering to the *ETS Standards for Quality and Fairness* (ETS, 2015). Given that test auditors are familiar with these standards, evidence presented in the simplified AUA is directly related to various ETS standards.

An extremely condensed version of the AUA is presented on the TOEIC research website (https://www.ets.org/toeic/research). Here, only the high-level claims (illustrated in Figure 1) are presented under the assumption that many readers will have an extremely limited assessment literacy. By focusing on the four fundamental, high-level claims, TOEIC research intends to communicate the key elements of an argument for test use to a broad audience. The four descriptive categories corresponding to the four high-level AUA claims on the TOEIC research website are shown in Figure 3.

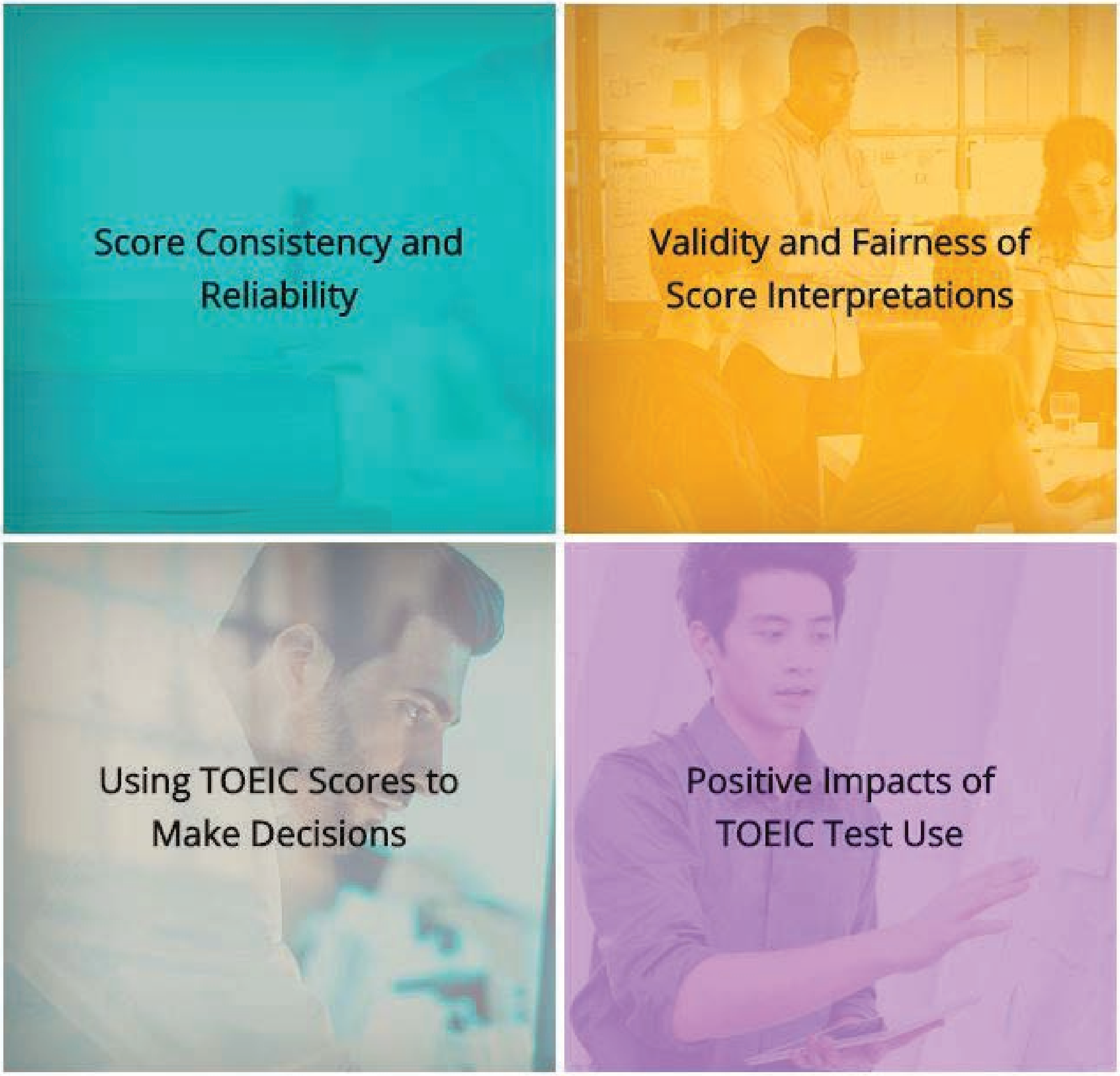
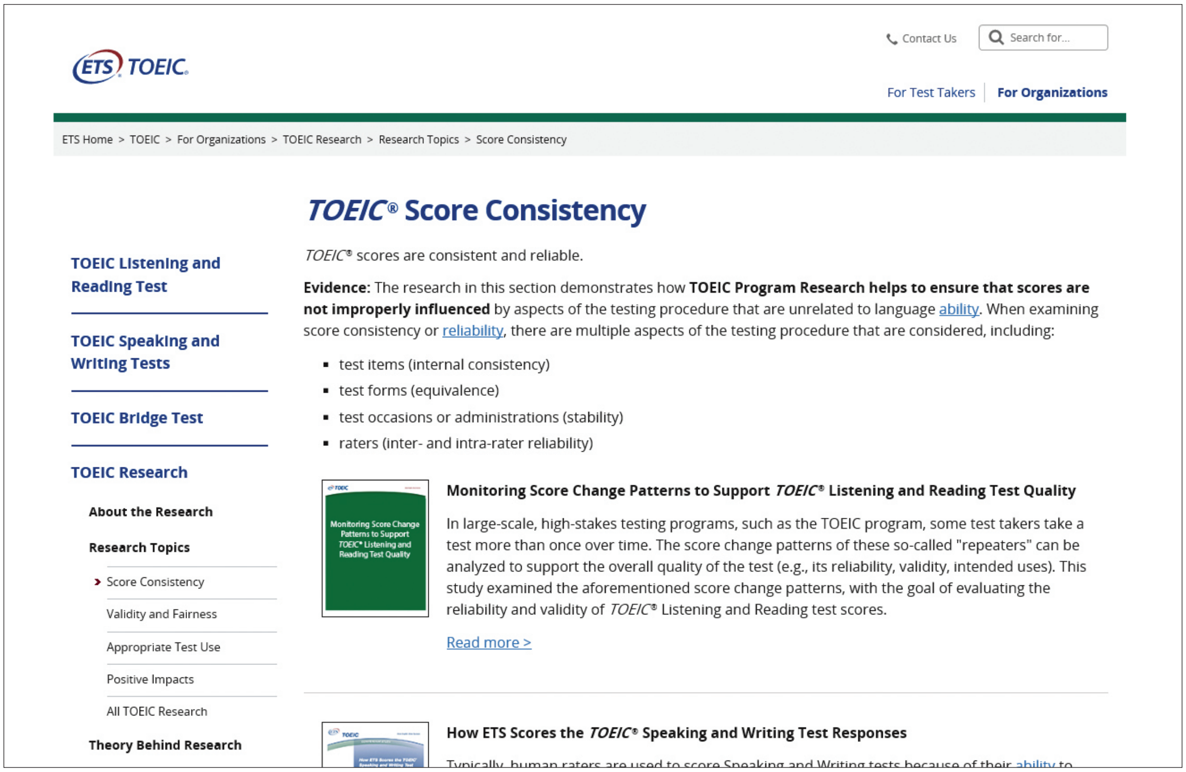

Score Consistency and Reliability
Validity and Fairness of Score Interpretations
Using TOEIC Scores to Make Decisions
Positive Impacts of TOEIC Test Use

***Figure 3.* Website categories corresponding to the four high-level assessment use argument claims.**

The four boxes shown in Figure 3 roughly correspond to each of the four overall claims in an AUA. If a website user places the mouse on one of the boxes, text appears to summarize the corresponding claim. For example, the following text corresponds to the score consistency and reliability category: "TOEIC scores are consistent and reliable, and are not improperly influenced by factors unrelated to language ability."

Website users who are interested in more information may click on one of the four categories or explanatory text (e.g., “TOEIC scores are consistent and reliable, and are not improperly influenced by factors unrelated to language ability”) to read a brief and accessible summary of the types of warrants that support the overall claim (e.g., scores are consistent across test items, test forms, test administrations, raters) and review some of the evidence that is available to support claims and warrants. Figure 4 below illustrates how this has been implemented for the score consistency topic.


ETS TOEIC
Contact Us
Search for...
For Test Takers
For Organizations
Monitoring Score Change Patterns to Support TOEIC® Listening and Reading Test Quality
How ETS Scores the TOEIC Speaking and Writing Test

ETS Home > TOEIC > For Organizations > TOEIC Research > Research Topics > Score Consistency

TOEIC Listening and Reading Test

TOEIC Speaking and Writing Tests

TOEIC Bridge Test

TOEIC Research

About the Research

Research Topics

› Score Consistency

Validity and Fairness

Appropriate Test Use

Positive Impacts

All TOEIC Research

Theory Behind Research

TOEIC® Score Consistency

TOEIC® scores are consistent and reliable.

Evidence: The research in this section demonstrates how **TOEIC Program Research helps to ensure that scores are not improperly influenced** by aspects of the testing procedure that are unrelated to language ability. When examining score consistency or reliability, there are multiple aspects of the testing procedure that are considered, including:

- test items (internal consistency)
- test forms (equivalence)
- test occasions or administrations (stability)
- raters (inter- and intra-rater reliability)

Monitoring Score Change Patterns to Support *TOEIC*® Listening and Reading Test Quality

In large-scale, high-stakes testing programs, such as the TOEIC program, some test takers take a test more than once over time. The score change patterns of these so-called "repeaters" can be analyzed to support the overall quality of the test (e.g., its reliability, validity, intended uses). This study examined the aforementioned score change patterns, with the goal of evaluating the reliability and validity of *TOEIC*® Listening and Reading test scores.

Read more >

How ETS Scores the *TOEIC*® Speaking and Writing Test Responses

Typically, human raters are used to score Speaking and Writing tests because of their ability to

***Figure 4.* Descriptive text and information on the website to support the claim that scores are consistent and reliable.**

As illustrated in Figure 4, each category includes a restatement of the overall claim, an outline of warrants that support the claim, and a list of relevant research evidence. An executive summary is provided for each research paper. This summary includes the purpose of the study, the evidence it produced, and the implications of the evidence for relevant claims. A link is provided to an electronic copy of the research publication for those that are interested.

Discussion

This paper provided a rationale for the argument-based approach to validation and an overview of the AUA, one such approach. It highlighted how this approach has been implemented in a novel way to produce fully specified AUAs for the TOEIC tests, which are then applied to guide TOEIC test research and disseminate the argument for TOEIC test use to various stakeholders. One of the purposes of providing simplified AUAs is to increase the assessment literacy of different stakeholder groups, including test takers and score users. The current design of the TOEIC research website reflects this intention. In the future, researchers at ETS hope to report on the effectiveness of the simplified AUA for promoting assessment literacy.

For tests that have multiple uses—such as the TOEIC tests—one of the challenges of using an argument-based approach to justifying test use is that it may require a number of individual AUAs. This implies a lot of documentation that could be difficult to create, maintain, and adapt for the purpose of communicating with different stakeholders. However, there is a potential solution to this challenge: a theory of action.

A theory of action is a logical model of how components of a test (e.g., test scores) can facilitate actions (i.e., decisions) that have intermediate and long-term outcomes (i.e., consequences). As exemplified by Bennett (2010), it includes a visualization that functions as a high-level summary of all of the intended uses of an assessment and their expected consequences. In a single figure, it could provide an accessible summary of the supported uses of a TOEIC test and the expected consequences of test use. Such a figure would also indicate the relationship between test components (e.g., scores), decisions, and consequences. Supporting documentation is expected to summarize the evidence to support each hypothesized relationship in the logic model, including potential rebuttals. Thus, the future publication of a theory of action for TOEIC tests may be a beneficial tool to communicate claims and supporting evidence about TOEIC test use in an accessible manner.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Apichatronajanakul, P. (2011). The washback effects of the *TOEIC®* examination on the teachers and students of a Thai business school. *Language Testing in Asia*, *1*(1), 62–75. https://doi.org/10.1186/2229-0443-1-1-62

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford, United Kingdom: Oxford University Press.

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning (CBAL): A preliminary theory of action for summative and formative assessment. *Measurement*, *8*, 70–91. https://doi.org/10.1080/15366367.2010.508686

Chapelle, C. A., & Voss, E. (2014). Evaluation of language tests through validation research. In A. Kunnan (Ed.), *The companion to language assessment*. New York, NY: Wiley.

Educational Testing Service. (2013). *TOEIC® user guide: Speaking and Writing*. Princeton, NJ: Author.

Educational Testing Service. (2015). *ETS standards for quality and fairness*. Princeton, NJ: Author.

Hines, S. (2010). Evidence-centered design: The *TOEIC®* Speaking and Writing tests. In *The research foundation for the TOEIC® tests: A compendium of studies* (pp. 7.1–7.31). Princeton, NJ: Educational Testing Service.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.). New York, NY: American Council on Education and Praeger.

Lawn, M. J., & Lawn, E. (2015). Increasing English communicative competence through online English conversation blended e-learning. *International Journal of Information and Education Technology, 5* (2), 105–112. https://doi.org/10.7763/IJIET.2015.V5.485

Messick, S. J. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

Newton, P. E. (2012). Clarifying the consensus definition of validity. *Measurement: Interdisciplinary Research and Perspectives, 10* (1–2), 1–29. https://doi.org/10.1080/15366367.2012.669666

Powers, D. E. (2010). Validity: What does it mean for the *TOEIC*® tests? In *The research foundation for the TOEIC® tests: A compendium of studies* (pp. 1.1–1.11). Princeton, NJ: Educational Testing Service.

Schedl, M. (2010). Background and goals of the *TOEIC*® Listening and Reading test redesign project. In *The research foundation for the TOEIC® tests: A compendium of studies* (pp. 2.1–2.18). Princeton, NJ: Educational Testing Service.

Schmidgall, J. E., & Choi, I. K. (2011, May). *Frameworks for validity: A comparison of traditional and argument-based approaches for reviewing research.* Paper presented at the 14th annual conference of the Southern California Association for Language Assessment Researchers, Los Angeles, CA.

Toulmin, S. E. (2003). *The uses of argument* (updated ed.). Cambridge, United Kingdom: Cambridge University Press. https://doi.org/10.1017/CBO9780511840005

Wang, H., Choi, I., Schmidgall, J., & Bachman, L. F. (2012). Review of Pearson Test of English Academic: Building an assessment use argument. *Language Testing*, *29*, 603–619. https://doi.org/10.1177/0265532212448619

Compendium Study

The Case of Taiwan: Perceptions of College Students About the Use of the *TOEIC*® Tests as a Condition of Graduation

Ching-Ni Hsieh

ETS® TOEIC®

Assess to Progress.

Research has shown that high-stakes tests can have a critical impact on test takers and other stakeholders (e.g., Alderson & Hamp-Lyons, 1996; Cheng, Andrews, & Yu, 2010; Hamp-Lyons, 1997, 1998; Murray, Riazi, & Cross, 2012; Qi, 2005, 2007; Wall, 2000). High-stakes tests are those that are used for making important educational and professional decisions about individuals, such as admissions, graduation, employment, or promotion. One of the high-stakes testing situations that has important consequences for students is the use of an exit test as a condition of graduation. Exit tests are generally considered to represent one of the mechanisms to ensure accountability in education (Berry & Lewkowicz, 2000). In the field of language testing and assessment thus far, relatively few studies have examined the use of standardized language proficiency tests as exit tests and their impact on teaching and learning (e.g., Berry & Lewkowicz, 2000; Nhan, 2013; Tasi & Tsou, 2009). The current study adds to this line of research with an examination of college students' perceptions of the English-language graduation requirement policy implemented by Taiwan's higher education (HE) institutions and the appropriateness of the *TOEIC*® tests as an exit test in this context. The use of the TOEIC assessments as an exit test has a direct impact on students who take the test to meet the graduation requirement, and it has important educational and economic implications for students and society at large. Language learners or students whose education and language learning are directly influenced by the implementation of the language graduation requirement policy are perhaps the most important stakeholders in this testing situation. However, their views on the policy implementation are rarely consulted, if at all (Shih, 2010). Previous research has suggested that learners' attitudes toward and perception of a test and its use can affect their motivation and performance on the test (Bachman & Palmer, 1996). Given that the test-taker perceptions can have wide-ranging consequences, opinions about the use of TOEIC tests as an exit test warrant further investigation to ensure the tests' validity.

Research Context

Since 2003, the Ministry of Education (MOE) in Taiwan has encouraged HE institutions to set English proficiency thresholds for undergraduates and to implement an English-language requirement policy for graduation. The objective is to raise students' English-language proficiency (ELP) and to better prepare students to cope with global competition and meet the ELP requirements of the workplace (Shih, 2009, 2010, 2012). To this end, the MOE has advocated the use of standardized English proficiency tests as exit requirements. The *TOEIC*® tests, along with other standardized tests such as the *TOEFL*® test, International English Language Testing System (IELTS®), and the locally developed General English Proficiency Test (GEPT) have been recommended by the MOE to meet this objective. As a result, the majority of colleges and universities in Taiwan have set a requirement for students to achieve a satisfactory score on one of the recommended English proficiency tests before graduation.

Taiwan has two major types of 4-year HE institutions: general universities and technical colleges. General universities usually have more rigorous admissions standards with respect to test scores compared to technical colleges that focus on vocational education or training (although there are different tiers of general universities and technical colleges). Taiwan's MOE has advised HE institutions that adopt the exit requirement policy to set varied criteria according to students' language learning and communication needs and levels of proficiency. The MOE has recommended that general universities set an English-language graduation criterion at the B1 or above level on the Common European Framework of Reference (CEFR) as put forth by the Council of Europe (2001). Furthermore, the MOE has suggested that technical colleges adopt the A2 level as their benchmark for graduation. Apart from recommending the CEFR levels as the English-language graduation benchmark, the MOE does not prescribe what English-language skills should be tested or what test scores on which tests should be required. University authorities could formulate their own policies and decide upon the specific ELP requirements for their students (Shih, 2010, 2012).

The TOEIC program, one of the MOE's recommended proficiency tests for college graduation, is an ELP test for those whose native language is not English (Educational Testing Service [ETS], 2015). It measures the everyday English skills of people working in an international environment. Test scores indicate how well test takers can communicate with others in English in business, commerce, and industry. The test does not require specialized knowledge or vocabulary beyond that of a person who uses English in everyday work activities. The primary uses of the TOEIC tests are to allow test takers to verify their current level of English proficiency, qualify for a new position or promotion in a company, enhance their professional credentials, monitor their progress in English, set their own learning goals, and involve their employer in advancing their English ability. In the past decade, the TOEIC has gained increasingly wide recognition by test takers and score users in Taiwan, and the test scores are now being used extensively for recruitment and promotion by both domestic and multinational corporations and organizations (Pan & Roever, 2016).

In the mid-2000s, many of Taiwan's HE institutions began to accept the TOEIC program as one of the language proficiency tests suitable for exit purposes. The majority of those that accept the TOEIC tests require students to take only the *TOEIC*® Listening and Reading test and earn test scores that meet the minimum requirements for graduation. The TOEIC *Listening* section assesses how well test takers understand spoken English, and the TOEIC *Reading* section tests how well test takers understand written English. The TOEIC Listening and Reading test is designed to enable test takers to demonstrate their English listening and reading skills and thereby qualify for better employment opportunities and gain a competitive edge in the global workplace.

As of 2015, a total of 133 (79%) of Taiwan's 169 HE institutions had complied with the MOE directive and had implemented the English-language graduation requirement policy; all of these accept TOEIC Listening and Reading test scores (Nichols, 2016). TOEIC Speaking and Writing test scores, in contrast, are less commonly required for college graduation, except for certain academic disciplines such as English or business, primarily because of the concerns about testing fees and the possible low passing rates.

Scores from the TOEIC Listening and Reading test have been mapped onto the CEFR (Tannenbaum & Wylie, 2013). The minimum cut score for the B1 level is a total TOEIC Listening and Reading score of 550, and the minimum cut score for A2 is 225. The majority of the universities and colleges that have adopted the TOEIC tests have set a cut score for graduation between 450 and 550. Some top-tier institutions have set a higher bar; others have set different criteria for different majors (Shih, 2012). For example, English and business majors are usually required to obtain a higher passing score than their non-English or nonbusiness major counterparts (Pan & Roever, 2016).

A growing body of empirical studies has investigated the impact of the English-language graduation requirement policy on language learning and teaching in Taiwan. Most studies have examined the viewpoints of students and teachers from technical colleges (e.g., Chu, 2009; Hsu, 2009; Pan, 2014; Pan & Newfields, 2011; Shih, 2009, 2010; Tasi & Tsou, 2009). Few studies have included perceptions of stakeholders from general universities (e.g., Chen, 2008; Vongpumivitch, 2006; Wu, 2012). These studies have found conflicting views regarding whether students are in support of or against the language requirement policy. In addition, students in technical colleges have broadly reported that the policy causes pressure and anxiety. These feelings were especially strong for low-proficiency students and for those with little interest or motivation to learn English, who felt that they were forced to "study for the test" (Hsu, 2009; Tasi & Tsou, 2009). Most of the studies (e.g., Pan, 2014; Shih, 2010) also show that the policy has resulted in limited or no washback (positive or negative effects in the classroom). In contrast, Chen (2008) reported that students from a top-tier, general university thought that the exit requirement policy helped students improve their ELP, even though the policy did not play a significant role in their learning motivation.

Interestingly, but perhaps unsurprisingly, most of the publicly available empirical studies conducted in the Taiwanese context have examined the use of the GEPT, the locally developed test, which was once the most popular exit test (Roever & Pan, 2008; Shih, 2007). Little research exists that examines the use of the TOEIC assessments as an exit test in the Taiwanese HE context and how the Taiwanese college students perceive the utility of the test. As Shohamy, Donitsa-Schmidt, and Ferman (1996) suggested, the perceived status of a test, such as whether a test is locally or internationally known, is directly linked to students' motivation, time devoted, and effort exerted to prepare for the test. It remains unclear whether the use of the TOEIC tests—tests with an international reputation and recognition for validity—is more likely to be perceived positively as a college exit test by test takers.

Within the context of the exit requirement policy in Taiwan, previous studies have suggested that students' views about the policy are often not considered when university authorities set the policy (Shih, 2010). Students are arguably the most important stakeholders in this testing situation, and their opinions about the implementation of the policy and its impact warrant further investigation. Previous research has also suggested that students of varying proficiency could have differing views about the policy and its impact on learning. Investigations into these issues could provide the MOE and policymakers with important insights into the appropriateness of the graduation benchmarks or cut scores. This study examined the opinions of students about the language requirement policy and the efficacy of the TOEIC assessments as an exit test. The research questions guiding this study were as follows:

1. How do Taiwanese college students perceive the English-language graduation requirement policy? Are their perceptions related to their ELP?

2. How do Taiwanese college students who take the TOEIC tests for graduation perceive the efficacy of the TOEIC as an exit test? Do students of high and low proficiency differ in their perceptions?

Methodology

Instruments

The instruments used in this study included an online student survey and a set of semistructured interview questions. The survey questions asked about students' biographic, educational, and language learning backgrounds; TOEIC test-taking experiences; and perceptions about the English-language graduation requirement policy and the use of the TOEIC tests as an exit test. The survey was created using the online survey software SurveyMonkey® (http://www.surveymonkey.com). The interview further explored factors that influenced students' perceptions.

Participants

Student Survey

The ETS representative in Taiwan served as the site coordinator for the project, sending out the link to the survey to more than 22,000 college students who had received valid TOEIC Listening and Reading scores (i.e., TOEIC Listening and Reading scores obtained within the past 2 years) at the time of data collection. A total of 1,527 surveys with valid responses were returned (response rate = 7.1%). Information about the respondents' latest TOEIC Listening and Reading scores and the number of times test takers had taken the TOEIC tests were extracted from the official TOEIC test database. There were 361 males and 1,145 females; 21 respondents preferred not to respond to the gender question. The students' ages ranged between 19 and 24 ($M = 21.6$, $SD = 0.9$), and they had taken the TOEIC Listening and Reading test between one and 11 times ($M = 3.11$, $SD = 1.33$). The students' most recent TOEIC Listening and Reading scores ranged between 175 and 980 ($M = 610$, $SD = 169.4$). The survey respondents represented 123 Taiwan HE institutions that used TOEIC tests as an exit test, with a total of 75 general universities and 48 technical colleges. The students came from a wide range of academic disciplines, broadly representing arts and humanities (24%), social sciences (6%), natural sciences and engineering (20%), education (2%), and business (34%), with 14% missing data.

All survey respondents were asked about their reasons for taking the TOEIC tests and their test-preparation activities. Both questions allowed students to include all relevant answers. The most commonly reported reason was "to meet the graduation requirement" (71.1%), followed by "to improve English-language proficiency" (59.7%), "to prepare for the job market" (51.4%), "to get an internationally recognized English proficiency test certificate" (37%), "to qualify for a better paying job" (31.6%), and "to communicate with English speakers" (22.7%). When asked how they prepared for the TOEIC tests, 69.0% reported using practice tests, 38.6% enrolled in language courses offered by their university or college, and 8.1% took courses at private language schools. Few students (12.9%) took TOEIC test-preparation courses at test-preparation training schools. The majority of the respondents (71.4%) also reported watching movies in English, reading magazines in English, listening to radio programs in English, and the like to prepare for the test. Very few students (2.8%) hired personal tutors or made foreign friends to help improve their proficiency in English.

Student Interviews

Twenty-six randomly selected survey respondents who expressed interest in participating in the follow-up interviews were interviewed by the researcher. The interviewees included 12 males and 14 females between the ages of 19 and 23 (M = 21.38, SD = 1.06). They had taken the TOEIC tests between two and five times (M = 2.73, SD = 0.87). Their most recent TOEIC Listening and Reading scores ranged from 405 to 945, and the mean score was 681.15 (SD = 157.18). Although the interviewees' demographics largely mirrored those of the survey respondents in terms of age and gender, their TOEIC Listening and Reading mean score was higher than the mean score of the total pool of survey respondents, suggesting that the interviewees were overall more proficient language users. There were 21 students from general universities and five from technical colleges. The students represented 10 different institutions.

Procedures

Student Survey

The link to the online survey was sent to the site coordinator, who distributed it to TOEIC test takers who met the selection criteria (i.e., college students who had taken the TOEIC Listening and Reading test at least once, had valid TOEIC Listening and Reading scores, and whose university or college used TOEIC Listening and Reading test as one of its English-language exit tests). The survey stayed live for roughly 1 month to allow the students sufficient time to respond.

Student Interviews

The 26 students who agreed to participate in the follow-up interviews were contacted by the site coordinator to set up the interview sessions. The interviews were conducted one-on-one between the researcher and the students through phone calls. All interviews were conducted in Mandarin Chinese, the researcher's and the students' first language. Each interview session lasted between 15 and 20 minutes and was audio-recorded.

Data Analysis

Student Survey

The responses to the online survey were extracted from the SurveyMonkey website. Frequency counts, descriptive statistics, and cross-tab analyses were performed to answer the research questions.

Student Interviews

The audio recordings of the interviews were translated from Mandarin Chinese to English and then transcribed into English semantically by the researcher. The transcripts were used as data for analysis. First, the researcher read and reread the transcripts in their entirety several times to obtain a holistic impression of the responses and to determine a preliminary list of eight analytical categories that had the potential to provide answers to the research questions. For example, one of the categories was policy goals, intended to examine the students' viewpoints on the goals of the language exit requirement policy. Further analysis was conducted to identify different variants under each analytical category. For example, under the analytical category, policy's impact on ELP, three variants emerged from the data: *positive*, *negative*, and *no impact*. Each variant was defined and described in the coding scheme following the procedures for analyzing semistructured interviews proposed by Schmidt (2004). The coding scheme was subsequently revised to incorporate new analytical categories or variants and to remove those that were redundant or irrelevant. Once the coding scheme was finalized, six of the transcripts were double-coded by the researcher and a research assistant to establish coder reliability. Interrater agreement was reached at 95%. Discrepant cases were resolved through discussion. The researcher then coded the entire dataset using the qualitative data analysis software NVivo 11 (http://www.qsrinternational.com/nvivo-product/nvivo11-for-windows).

Results and Discussion

This section presents the results by research question and discusses their implications.

Research Question 1

How do Taiwanese college students perceive the English-language graduation requirement policy? Are their perceptions related to their ELP?

Student Survey Results

The students were asked to respond to statements about their viewpoints regarding the necessity of the ELP requirement policy and the policy's impact on their language learning (see Table 1). The majority (82.3%) supported the policy and thought that HE institutions should require students to achieve a certain level of English proficiency before graduation. Interestingly, more than 90% of the students indicated that they would be motivated to study English even if the policy were not in place, implying that the policy is not a primary motivator for many students.

Although previous case studies found that Taiwanese college students felt that they were forced to take standardized language proficiency tests due to the exit requirement (e.g., Chu, 2009; Hsu, 2009), 81.9% of the respondents indicated that they would still be motivated to prepare and take standardized language proficiency tests, regardless of the presence of the policy. Contrary to previous findings, this result suggests that in the test-driven educational environment in Taiwan, students would take standardized ELP tests for reasons other than the graduation requirement. It can be argued that because there was a relatively high percentage of respondents who took the TOEIC tests as a graduation requirement, the use of the TOEIC tests as an exit test could have impacted how respondents perceived the necessity and implementation of the policy. These students may perceive added value in obtaining an internationally recognized ELP certificate such as the TOEIC certificate, insofar as it is beneficial to have a policy that encourages them to obtain a satisfactory score on the TOEIC tests in preparation for their future employment. The students' future job prospects and impact on credentials needed for job applications may serve as a strong impetus for them to take standardized ELP tests.

Table 1

Students' Opinions About the Exit Requirement Policy

Statement	Agree (%)	Disagree (%)
It is necessary for universities to require their students to obtain a certain score on an English proficiency test before graduation.	83.2	16.7
I would still be motivated to study English in college even if there were no English graduation requirement.	91.7	8.3
I would still be motivated to prepare to take standardized English-language proficiency tests even if there were no English graduation requirement.	81.9	18.1

Note. Total sample *N* = 1,527.

To explore whether students' opinions about the exit requirement policy are related to their language proficiency, the survey respondents were first split into two proficiency groups, *high* and *low* (i.e., falling above and below respondents' median TOEIC Listening and Reading score of 615). It should be noted that these were relative proficiency groups only for the convenience of the analysis and should not be interpreted as representing different proficiency levels based on the TOEIC Listening and Reading scores of a more representative sample. The mean difference in the total TOEIC Listening and Reading scores between the high- and low-proficiency groups was 278.3 total score points, and the difference was statistically significant, $t(1483) = 56.24$, $P < .001$, $d = 2.87$.

Cross-tabulations were performed to determine if language proficiency could predict students' viewpoints for or against the language graduation requirement policy (see Table 2). The analysis was conducted within each type of HE institution. Significant differences were found between the high- and low-proficiency groups in their opinions about the necessity of the policy, both for students of general universities, χ^2 (1) = 10.45, $P < .01$, and for students of technical colleges, χ^2 (1) = 7.61, $P < .01$. Based on the odds ratio effect size, the high-proficiency students of general universities were 1.69 times more likely to support the policy than their low-proficiency counterparts; similarly, the high-proficiency students of technical colleges were 2.38 times more likely to support the policy than their low-proficiency counterparts. The results suggest that more proficient students tend to have a more positive attitude toward the policy, perhaps in part because the policy is less likely to pose a threat for their graduation. It is also possible that they believe that the policy could serve as an incentive for them to obtain an English-language certificate that could be beneficial for their future.

Table 2

Opinions About the Exit Requirement Policy by Proficiency

Type of higher education institution	TOEIC scores		Opinion about exit requirement policy		*N*	χ^2
	Mean	*SD*	Necessary	Not necessary	Total	Sig. (2-sided)
General university						.001
High	754.57	86.47	507	89	596	
Low	506.46	80.49	312	93	405	
Technical college						.006
High	733.87	91.39	152	13	165	
Low	433.15	114.95	300	61	361	

Student Interview Results

The interview data extended the survey results and helped illuminate potential factors that influence students' perceptions of the English-language graduation requirement policy. The students' responses centered on three major themes: policy goals, policy's impact on ELP, and policy impact on motivation. Table 3 indicates the number of students who commented on each of the subcategories within each main category and the number of comments made.

Table 3

Student Interviewees' Perceptions About the English-Language Graduation Requirement

Main category	Subcategory	Number of students	Number of comments
Policy goals	To ensure students have minimum ELP for communication	18	23
	To help improve students' ELP	18	28
	To motivate students to study English	12	17
	To prepare students for further study	5	6
	To prepare students for future employment	12	17
Policy impact on ELP	Positive	22	39
	Negative	0	0
	No impact	3	4
Policy impact on motivation	Positive	18	30
	Negative	3	4
	No impact	9	12
	Short-term impact	2	2

Note. ELP = English-language proficiency.

The students made numerous comments on what they thought the language policy was designed for. Collectively, the data showed that the students had a fair understanding of the policy goals and were generally in support of their implementation. Eighteen students considered that the policy was meant to ensure that college graduates have the minimum level of English proficiency for communication and that the policy could help students improve their ELP. Twelve students believed that the policy was put in place to motivate students and to prepare them for future employment; five students thought that the policy could help prepare students for their future studies.

When asked about the policy's impact on learning, the interviewees largely commented positively about its influence on their ELP. "I think it's a great policy, because when you go on the job market, English is always required. So it's a good motivator," commented Student 26 (high proficiency, technical college, sophomore). In contrast, several students felt that the policy had limited or no impact on their ELP or motivation. A few pointed out that the policy's impact was short term in nature: "I'm a senior student now, so the policy pushed me to study English for a while. After I passed the test, I lost my motivation to study English," said Student 13 (low proficiency, general university, senior). Another factor that influenced the students' views about the policy pertained to their perception that an appropriate cut score for graduation was lacking. Student 1 (high proficiency, general university, senior) reported:

> I don't think the policy is meaningful or has any impact on students' learning or motivation because the requirement is so low. In my school, the requirement is 450 on the TOEIC Listening and Reading test. Even if you don't prepare for the test and you just guess randomly, you might be able to get 200 or 300 points. The requirement set by my university was just the English level required for middle-school students. If the requirement were higher, I think it would help students improve their proficiency better.

Similar concerns about the inadequate cut scores were raised in the responses of seven students who generally supported the policy but were skeptical about the positive washback the policy could induce. For instance, Student 2 (low proficiency, technical college, junior) stated:

> Well, I've discussed this issue [low cut score] with many of my friends, and we all agree that a higher graduation requirement would really motivate us to study English. Since the requirement in my school is so low, most students think that it's a piece of cake, and they don't need to work hard and they'll be able to meet the requirement anyways. Although I think it's a good policy, it doesn't really make a dent in motivating students.

Interestingly, Student 2, whose most recent TOEIC Listening and Reading score was 455 and was one of the least proficient students in the entire interviewee pool, seemed confident that students at his college, which was a second-tier technical college, would not have much difficulty meeting the requirement. Several universities in Taiwan had in fact set a cut score of 450 or lower on the TOEIC LR to avoid having a small passing rate, which could impact the ratings they receive from routine program evaluations conducted by the MOE and the funding they obtain from the government (Shih, 2012). The students' comments above, however, reveal that an inadequate cut score could undermine the goals of the policy and adversely cause students to lose the incentive to work hard, knowing that the threat of failing was minimal.

Research Question 2

How do Taiwanese college students who take the TOEIC tests for graduation perceive the efficacy of the TOEIC tests as an exit test? Do students of high and low proficiency differ in their perceptions?

Student Survey Results

Given that college students have multiple standardized tests to choose from for meeting the exit requirement, it is logical to assume that students who choose to take the TOEIC tests would perceive the use of the TOEIC tests as an exit test in a more positive light. To determine whether this assumption holds, the following analysis included only responses of students (*N* = 1,086) who indicated that they took the TOEIC Listening and Reading test in order to pass the exit requirement, among other reasons.

Table 4 shows the frequency and percentage of students who agreed or disagreed with each statement related to their perceptions about the use of the TOEIC test. The majority of the students (86.4%) reported that preparing for the TOEIC tests had a positive impact on their ELP. Students' views about the validity of the test scores, however, were somewhat mixed. A slight majority of respondents (54.7%) indicated that the test scores accurately reflected their level of ELP, and the other half disagreed. The respondents' differing views on score validity may have been influenced by the exit requirement policy that required only TOEIC Listening and Reading test scores. Students who did not show confidence in the validity of the scores may feel that, as a measure of receptive (listening and reading) and not productive (speaking and writing) skills, the TOEIC Listening and Reading test could not fully reveal their language competency. The result could also reflect an issue of how the statement was phrased. Had the students been asked about whether the scores demonstrated their English listening and reading abilities in the workplace, the results could have been different.

Students' perceptions regarding the requirement of TOEIC Speaking and Writing tests for graduation was divided, with slightly less than half (48.3%) admitting that productive skills were important and should be required. It was interesting to note that the students' mean TOEIC Listening and Reading scores differed significantly between the two groups of students who were in favor or not in favor of the requirement of productive skills, $t(1084) = 6.53, p < .001, d = 0.39$ for listening and $t(1084) = 5.62, p < .001, d = 0.34$ for reading. The results indicated that students with better listening and reading skills as measured by the TOEIC test had a significantly stronger preference for the requirement of an adequate level of speaking and writing proficiency before graduation. Students with lower receptive skills are less likely to pass the graduation requirement, and thus they might have a fear of adding another obstacle if productive skills are also required.

When asked about their test-taking motivation, almost 75% of the students responded that they would be motivated to prepare to take the TOEIC tests even if they were not required to pass an English proficiency test before graduation. This high level of test-taking motivation once again demonstrated that the students believe in the utility of the TOEIC test scores for purposes other than passing the exit requirement.

Table 4

Survey Respondents' Perceptions About the Use of the TOEIC Tests as an Exit Test

Statement	Agree (%)	Disagree (%)
Preparing for the TOEIC helps me improve my ELP.	86.4	13.6
TOEIC LR scores accurately reflect my ELP.	54.7	45.3
TOEIC speaking and writing tests should be required for graduation.	48.3	51.7
If there were no English exit requirement, I would still be motivated to prepare for and take the TOEIC.	74.6	25.4

Note. Total sample $N = 1{,}086$.

The second part of Research Question 2 addressed whether students of high and low ELP view the efficacy of the TOEIC tests as an exit test differently. Four separate cross-tab analyses were performed to determine if the perception differences exist (see Table 5). The analysis showed that the high- and low-proficiency students did not differ in their views on the use of the TOEIC tests to improve their ELP, χ^2 (1) = .904, $P = .342$. In general, the majority were positive about its impact on language learning. A nonsignificant difference was also found for students' perceptions on the validity of the TOEIC scores, χ^2 (1) = .197, $P = .657$.

With regard to the requirement of the TOEIC Speaking and Writing test scores for graduation, results of the cross-tab analysis yielded a significant difference between proficiency groups, χ^2 (1) = 34.25, $P < .001$. The odds ratio effect size indicated that the high-proficiency group was 2.06 times more supportive of requiring speaking and writing tests compared to the low-proficiency one. A significant difference between proficiency groups was also found for test-taking motivation, χ^2 (1) = 17.97, $P < .001$. The odds ratio effect size indicated that high-proficiency students were 1.85 times more likely to prepare to take the TOEIC regardless of the presence of the policy, compared to the low-proficiency ones.

Table 5

Survey Respondents' Perceptions About the Efficacy of TOEIC Tests by Proficiency

Survey question	Proficiency group		*N*	χ^2
	High	Low	Total	Sig. (2-sided)
TOEIC helps improve my proficiency				
Agree	392	546	938	.342
Disagree	68	80	148	
TOEIC scores are valid				
Agree	248	346	594	.657
Disagree	212	280	492	
TOEIC speaking and writing should be required				
Agree	270	255	525	.000
Disagree	190	371	561	
I would take TOEIC without requirement				
Agree	373	437	810	.000
Disagree	87	189	276	

Note. Total sample *N* = 1,086.

Student Interview Results

The interviewees were asked to discuss their experiences and opinions about preparing for and taking the TOEIC tests and the use of the TOEIC tests as an exit test. To help answer the research question, the discussion focused on students' opinions about the use of the TOEIC tests as an exit test. Students' perceptions involved three major themes: test design, test purpose, and score validity. Table 6 summarizes the students' responses.

Table 6

Survey Respondents' Perceptions About the Use of the TOEIC Tests as an Exit Test

Main category	Subcategory	Number of students	Number of comments
Test design	Content is authentic	9	14
	Measure workplace English	8	9
	Include various native accents	4	6
	A well-designed test	4	6
	Reading section too long	5	5
Test purpose	Future employment	15	9
	Further study	2	3
	Scholarship application	1	1
	Measure student's ELP	11	20
Score validity	Scores reflect actual ELP	18	22
	Scores do not reflect actual ELP	2	3

Note. ELP = English-language proficiency.

When responding to the question about the utility of the TOEIC tests as an exit test, the student interviewees generally commented positively, specifying that the test has good design features. Specifically, several students believed that the test content is authentic and the TOEIC Listening and Reading test is a good measure of workplace English. Others stated that the listening test reflects real-world English-language use because it includes various native accents. A few students reported that the TOEIC test is a much better and preferred test for graduation requirements. Five students (19%) felt that the reading test is too long and that they had difficulty finishing the test in time.

The second major theme pertains to the students' purposes for taking the test. The interviewees reported that obtaining a satisfactory score on the TOEIC test before graduation could simultaneously serve multiple purposes, such as qualifying for certain jobs, applying for graduate schools and scholarships, or simply assessing one's ELP. One sample quote follows (Student 3, high proficiency, general university, senior):

> The TOEIC test is very popular among companies and organizations and it has a good reputation for its good discrimination, so I decided to give it a try for my graduation requirement. Also since I'm about to graduate, the test certificate is very helpful when I apply for jobs.

Eleven students added that the TOEIC test assesses not only workplace English but also language use in real-world contexts. These students were sensitive to what is assessed in the test and considered that the test could also measure their general ELP. Student 5 explained in detail (low proficiency, technical college, freshman):

> I think the TOEIC test scores are very helpful and have practical utility because the test focuses on workplace English. . . . After I prepared for the TOEIC test, I also realized that the TOEIC test content involves a lot of real-life situations and you have to pay special attention to all the "wh" questions, like why, how, what, etc. The test content is a good reflection of the language skills you need in daily life and it can measure your general English-language proficiency as well.

The third theme that emerged from the data was related to students' perceptions about TOEIC score validity. Eighteen students made comments about their perceptions of score validity and were positive about the accuracy of the TOEIC Listening and Reading scores they obtained, suggesting that the scores were perceived as very reliable and reflected their actual English-language abilities. "I took the test two times, and my scores were very close both times. So I think the scores were very consistent and reflected my English abilities," said Student 19 (high-proficiency, general university, senior). When asked why he thought that the test score was an accurate reflection of his English skills, Student 3 commented:

> I think the scores were quite accurate because when I was responding to the questions, I knew which items I answered correctly and which ones I didn't know the answers. I mean, I know what my level is, and the scores reflected that.

Student 5 made a comment about score interpretations that helped explain the differing views on score validity seen from the study survey. She stated:

> I think speaking and writing abilities are different from reading and listening abilities. So, yes, I think the TOEIC Listening and Reading scores are good measures of my listening and reading abilities. But they can't reflect my speaking and writing abilities. . . . The university should not use listening and reading tests only to judge my overall English proficiency.

Two students did not regard the TOEIC Listening and Reading scores as valid and felt that it is easy to achieve a high score on the TOEIC tests by intensive test-preparation training. The following quote exemplifies the key opinion expressed (Student 10, high proficiency, general university, senior):

> I think it's possible to get very high scores on the TOEIC test if you just prepare for the test in a relatively short period of time. You know, preparing for the TOEIC test is very easy. All the items are multiple-choice items. The only thing is that there are many items on the test. If you can read faster and respond faster to the reading items, you can obtain high scores. So the scores cannot really reflect your English ability.

Although this comment about the possible effect of test preparation on score gains points to a concern of students with the use of the TOEIC tests as an exit test, the fact that the majority of the students believed that the TOEIC test scores have good reliability and validity suggests that the impact of test preparation might not be a serious problem within the context of this study. To better understand the role of test preparation and its impact, more empirical studies should be conducted to examine the nature of test-preparation activities and how these activities might, on the one hand, assist students in validly responding to test questions or, on the other hand, artificially inflate their test scores. Results of such investigations can help strengthen the validity argument for the TOEIC test, justify its use, and bring about more positive perceptions in the local context.

Conclusion and Implications

This study examined TOEIC test takers' perceptions about the English-language graduation requirement policy and the use of the TOEIC tests as an exit test within Taiwan's HE institutions. The survey results and the interview data collectively showed that the majority of students were in favor of the policy and were positive about the use of the TOEIC tests as an exit test. Findings of the study suggest that students' levels of ELP are related to their perceptions about the policy and the appropriateness of the test use. The results also revealed that the cut scores set by some institutions might be too low to bring about positive washback, corroborating the findings of Shih (2007). In light of the results, we recommend that Taiwan's MOE policy makers and university decision makers periodically review the cut scores to ensure that the requirements are appropriate for their students and the intended purpose of the exam.

The survey respondents generally believed that preparing to take the TOEIC tests was helpful for improving their ELP and acknowledged that there could be added value in requiring students to take the TOEIC Speaking and Writing tests. To promote positive washback, future research should explore the social, educational, and economic impact of requiring the TOEIC Speaking and Writing tests and issues that could arise as a result of such a requirement. It is likely that additional teaching and learning resources will need to be provided to help students increase their speaking and writing proficiency in order to succeed.

In general, students' perceptions of the purposes of the TOEIC tests are in line with the intended uses of the TOEIC tests for preparing test takers to gain a competitive edge in the job market. The results of the study have provided an important piece of empirical evidence in support of the use of the TOEIC test in Taiwan's HE context, and they have implications for the use of the TOEIC tests as a college exit test in other Asian contexts, such as Korea (Choi, 2008) and Vietnam (Nhan, 2013) and for exit tests in general (e.g., Berry & Lewkowicz, 2000; Spolsky, 1997). With the growing trend of using the TOEIC tests as a college gate-keeping test, more research that investigates this high-stakes test use across different contexts is called for. In light of the study results, future research will benefit from exploring individual learner factors (e.g., motivation, test anxiety, language proficiency) and educational backgrounds (e.g., academic major, year in school) and how these variables interact with students' perceptions about the use of the test scores for graduation within and across the testing contexts.

Findings of this study have implications for the creation and implementation of language testing policies. The study results suggest that test users, such as university staff or MOE policy makers who interpret and use language test scores, may not necessarily have a full understanding of what proficiency test cut scores mean and how to best use the information provided by language tests to make decisions. Given the high-stakes nature of their work, it is critical that these score users exercise their roles in an informed and ethical manner in the interest of valid test interpretation and use. To this end, the TOEIC program and the local partner could produce educational materials to help build the assessment literacy of test users. Specifically, score users and language educators need to be much better informed about the TOEIC test, the test processes, and the principles and concepts that guide testing practices. An appropriate level of assessment literacy among score users can help mitigate the risk of misuse of test scores in making decisions for students (O'Loughlin, 2013).

In high-stakes testing situations such as college exit tests, there is also a need to inform students about the test design, intended use, and score interpretations. When students are well informed of the test practices, they can better establish a link between their learning goals and the assessment tasks, and positive washback can be promoted. Policymakers should keep in mind that students' perceptions about test use can play an important mediating role in policy implementation. As the study results revealed, students' beliefs may not always be congruent with the intentions underlying the language policy required of them. Future research should continue to examine test-taker perceptions of the impact of exit tests on language learning, focusing on individual students, their learning goals, and their understanding of test scores. Reasons for possible inconsistencies between the intended policy objectives and student perceptions should be identified, discussed, and adequately addressed to successfully achieve the intended purposes of a language testing policy.

References

Alderson, C., & Hamp-Lyons, L. (1996). *TOEFL®* preparation courses: A study of washback. *Language Testing*, *13*, 280–297. https://doi.org/10.1177/026553229601300304

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice.* Oxford, United Kingdom: Oxford University Press.

Berry, V., & Lewkowicz, J. (2000). Exit-tests: Is there an alternative? *Hong Kong Journal of Applied Linguistics*, *5* (1), 19–49.

Chen, T.-H. (2008). *Impacts of compulsory standardized exams on college students' L2 learning motivation* (Unpublished master's thesis). National Chiao Tung University, Hsinchu, Taiwan.

Cheng, L., Andrews, S., & Yu, Y. (2010). Impact and consequences of school-based assessment (SBA): Students' and parents' views of SBA in Hong Kong. *Language Testing*, *28*, 221–249. https://doi.org/10.1177/0265532210384253

Choi, I.-C. (2008). The impact of EFL testing on EFL education in Korea. *Language Testing*, *25*, 39–62. https://doi.org/10.1177/0265532207083744

Chu, H.-Y. (2009). *Stakes, needs and washback: An investigation of the English benchmark policy for graduation and EFL education at two universities of technology in Taiwan* (Unpublished doctoral dissertation). National Taiwan Normal University, Taipei, Taiwan.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge, England: Cambridge University Press.

Educational Testing Service. (2015). *TOEIC® examinee handbook: Listening and Reading*. Princeton, NJ: Author.

Hamp-Lyons, L. (1997). Washback, impact and validity: Ethical concerns. *Language Testing*, *14*, 295–303. https://doi.org/10.1177/026553229701400306

Hamp-Lyons, L. (1998). Ethical test preparation practice: The case of the *TOEFL®*. *TESOL Quarterly*, *32*, 295–303. https://doi.org/10.2307/3587587

Hsu, H.-F. (2009). *The impact of implementing English proficiency tests as a graduation requirement at Taiwanese universities of technology* (Unpublished doctoral dissertation). University of York, York, Canada.

Murray, J. C., Riazi, A. M., & Cross, J. L. (2012). Test candidates' attitudes and their relationship to demographic and experiential variables: The case of overseas trained teachers in NSW, *Australia*. *Language Testing*, *29*, 577–595. https://doi.org/10.1177/0265532212440690

Nhan, T. (2013). The *TOEIC®* test as an exit requirement in universities and colleges in Danang City, Vietnam: Challenges and impacts. *International Journal of Innovative Interdisciplinary Research*, *2*, 33–50.

Nichols, J. (2016). Do high-stakes English proficiency tests motivate Taiwanese university students to learn English? *American Journal of Educational Research*, *4*, 927–930.

O'Loughlin, K. (2013). Developing the assessment literacy of university proficiency test users. *Language Testing*, *30*, 363–380. https://doi.org/10.1177/0265532213480336

Pan, Y.-C. (2014). Learner washback variability in standardized exit tests. *TESL-EJ*, *18* (2), 1–30.

Pan, Y.-C., & Newfields, T. (2011). Teacher and student washback on test preparation evidenced from Taiwan's English certification exit requirements. *International Journal of Pedagogies and Learning*, *6*, 260–272. https://doi.org/10.5172/ijpl.2011.6.3.260

Pan, Y.-C., & Roever, C. (2016). Consequences of test use: A case study of employers' voice on the social impact of English certification exit requirements in Taiwan. *Language Testing in Asia*, *6*, 1–22. https://doi.org/10.1186/s40468-016-0029-5

Qi, L. (2005). Stakeholders' conflicting aims undermine the washback function of a high-stakes test. *Language Testing*, *22*, 142–173. https://doi.org/10.1191/0265532205lt300oa

Qi, L. (2007). Is testing an efficient agent for pedagogical change? Examining the intended washback of the writing task in a high-stakes English test in China. *Assessment in Education*, *14*, 51–74. https://doi.org/10.1080/09695940701272856

Roever, C., & Pan, Y.-C. (2008). Test review: GEPT: General English Proficiency Test. *Language Testing*, *25*, 403–408. https://doi.org/10.1177/0265532208090159

Schmidt, C. (2004). The analysis of semi-structured interviews. In U. Flick, E. v. Kardorff, & I. Steinke (Eds.), *A companion to qualitative research* (pp. 253–258). London, England: Sage.

Shih, C.-M. (2007). A new washback model of students' learning. *The Canadian Modern Language Review*, *64*, 135–162. https://doi.org/10.3138/cmlr.64.1.135

Shih, C.-M. (2009). How tests change teaching: A model for reference. *English Teaching: Practice and Critique*, *8*, 188–206.

Shih, C.-M. (2010). The washback of the General English Proficiency Test on university policies: A Taiwan case study. *Language Assessment Quarterly,* 7, 234–254. https://doi.org/10.1080/15434301003664196

Shih, C.-M. (2012). Policy analysis of the English graduation benchmark in Taiwan. *Perspectives in Education*, *30* (3), 60–68.

Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Test impact revised: Washback effect over time. *Language Testing*, *13*, 298–317. https://doi.org/10.1177/026553229601300305

Spolsky, B. (1997). The ethics of gatekeeping tests: What have we learned in a hundred years? *Language Testing*, *14*, 242–247. https://doi.org/10.1177/026553229701400302

Tannenbaum, R. J., & Wylie, E. C. (2013). Mapping *TOEIC*® and *TOEIC Bridge*™ test scores to the Common European Framework of Reference. In D. E. Powers (Ed.), *The research foundation for the TOEIC® tests: A compendium of studies: Volume II* (pp. 6.1– 6.10). Princeton, NJ: Educational Testing Service.

Tasi, Y., & Tsou, C.-H. (2009). A standardized English language proficiency test as the graduation benchmark: Student perspectives on its application in higher education. *Assessment in Education: Principles, Policy & Practice*, *16*, 319–330. https://doi.org/10.1080/09695940903319711

Vongpumivitch, V. (2006). *An impact study of Taiwan's General English Proficiency Test (GEPT)*. Paper presented at the annual meeting of the Language Testing Research Colloquium, Melbourne, Australia.

Wall, D. (2000). The impact of high-stakes testing on teaching and learning: Can this be predicted or controlled? *System*, *28*, 499–509. https://doi.org/10.1016/S0346-251X(00)00035-X

Wu, J. (2012). GEPT and English language teaching and testing in Taiwan. *Language Assessment Quarterly*, *9*, 11–25. https://doi.org/10.1080/15434303.2011.553251

Compendium Study

Insights Into Using *TOEIC*® Test Scores to Inform Human Resource Management Decisions

María Elena Oliveri and Richard J. Tannenbaum

ETS® TOEIC®

Assess to Progress.

As has been noted (e.g., Harzing, Köster, & Magner, 2011), the use of English in the international workplace continues to rise, with corporate literature being increasingly published exclusively in English. Kim (2013) remarked that international businesses often indicate that knowledge of English is perceived as a requirement for employability, and without a working knowledge of English, otherwise fully qualified individuals may be disqualified from the applicant pool (Peltokorpi, 2010). Consistent with these assertions are the results of an Educational Testing Service (ETS) survey of large, multinational companies in 13 countries. Responses from 749 human resource (HR) managers across professional, scientific, and technical sectors revealed that English proficiency is considered central to workplace success (Educational Testing Service [ETS], 2014c).

The ability to communicate in English can have a positive effect on employability, but the lack of it can have negative consequences in the workplace. Piekkari (2006) provided examples of such consequences, which include (a) difficulty in communicating with external clients or vendors; (b) possible restrictions in the range of customers, suppliers, and other business partners; (c) a reduced ability to transfer knowledge across organizational units; and (d) difficulties collaborating in team projects and expanding international networks. Additionally, employees may feel disconnected with the employing company, leading to increased employee turnover (Ojanperä, 2014; Park, 2013; Peltokorpi, 2010).

To identify English-proficient candidates, international businesses have often used tests of English as a way to inform HR decision making related to hiring, promotion, and employee training (Newton, 2010). Moritoshi (2001) cautioned, however, that although assessments can be an objective and standardized tool to help inform HR decisions fairly and equitably, they need to be employed judiciously. To this end, this report explores the ways in which HR managers use scores from an English proficiency test (the *TOEIC*® test) that is designed to inform HR decisions in an international workplace. The TOEIC tests are widely administered tests that are used by more than 9,000 organizations worldwide across diverse industries, such as aviation, automobile, engineering, tourism, and banking (ETS, 2015, 2016).

To facilitate the appropriate use of TOEIC scores, the TOEIC program provides a guide (ETS, 2013) to help test score users use scores appropriately. For instance, the guide suggests that a score should not be the only source of evidence to inform decisions; rather, multiple sources (e.g., graduate or undergraduate grade point averages, years of experience in the targeted position, and letters of recommendations from past supervisors and colleagues) should be used to balance the limitations of any single measure of language proficiency. The use of multiple sources of data to inform decisions, rather than reliance on a single test score, is considered best practice *(Standards for Educational and Psychological Testing*; American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). That said, best practice is not always strictly observed. Hence, our goal here is to document how international organizations currently report using TOEIC Listening and Reading test scores. We anticipated that our search might uncover both exemplary uses and also, possibly, unfortunate misuses.

Gathering and Analyzing Test Users' Responses on Test Score Use

To explore the use of TOEIC scores, we utilized two data sources: (a) previously collected users' testimonials that described their use of TOEIC scores to inform HR decisions and (b) test-use examples collected from HR managers and the ETS Preferred Network (EPN) for the TOEIC program specifically for this project. These sources provided descriptive information about how HR managers use TOEIC scores.

We collected test-use examples in three steps. First, we developed an initial set of literature-based examples that served as a frame of reference for developing our own examples, which we administered to HR managers to elicit test-use responses from them. The literature-based examples were informed by our review of the workplace literature, which focused on articles that illustrated positive and negative consequences possibly arising in international businesses owing to employees' low English proficiency (for examples, see Marra, 2012; Ojanperä, 2014). Next, we organized the collected examples by type of HR decision. Third, we compiled the list of examples, which are provided in the appendix.

We then e-mailed the examples given in the appendix to the EPN members for the TOEIC program in the following 14 countries: Brazil, Mexico, Chile, Colombia, Taiwan, Vietnam, Spain, France, Poland, Germany, Italy, Japan, Korea, and Thailand. Five countries responded: Korea, Japan, Taiwan, Spain, and Brazil. Responses from Brazil and Taiwan were from the EPN members themselves and were informed by general conversations with HR managers with whom they worked. Responses from Japan, Korea, and Spain were completed by HR managers from a total of 16 companies: in Japan, five companies, including IHI Corporation, Casio Computer, Fuji Xerox, Honda Motor, Motorcycle Operations Department, and Nihon Spansion Limited; in Spain, four companies, including Bank of Spain, CNMV, Deloitte, and Acciona; and in Korea, six companies, which elected to remain anonymous. Respondents were asked to provide three to five examples of their use of TOEIC scores to inform HR decisions related to hiring and selection, training and professional development, on-the-job performance, and promotion. HR managers and EPN members were selected as our source of information because of their firsthand experience in how English is implemented in company operations. We then reviewed and organized the test-use examples along HR decisions such as hiring, promotion, and training.

Note that our analysis extends only to TOEIC score uses for the workplace. We note that TOEIC scores are also sometimes used by colleges, universities, and language training institutes, for example, to measure progress in English-language programs, to certify language competency skills, and to make decisions on eligibility of scholarships. Although such uses are important, we will not discuss them in this report, given our workplace focus.

Insights Into Using *TOEIC®* Scores to Inform Human Resources Decisions

Using Scores to Inform Hiring Decisions

According to the Institute for International Business Communication (2011), many international businesses require a minimum TOEIC score to be hired, often between 500 and 850 points, with the minimum required score varying by job type (Peltokorpi & Vaara, 2014). For instance, electronic companies (e.g., Packard Bell and Kenwood Electronics Technologies) use different minimum scores to hire employees for particular positions: Technicians require scores over 640, whereas buyers require scores greater than 850 (ETS, 2007d).

The specific ways in which TOEIC scores are used to inform hiring also vary. One often reported use is to help in screening. To illustrate, the testimonial from Minera Los Pelambres, a Chilean copper-mining company, exemplifies the use of TOEIC scores as a cost-savings approach to help narrow down a field of prospective employees from 50 to 2-4 candidates who advance to the interview stage (ETS, 2007a). In so doing, the TOEIC program acts as a filter to assist in reducing a large applicant pool to a more manageable, smaller pool of applicants. The testimonials from the 2007 Qingdao International Regatta also revealed the use of TOEIC tests as a filter for HR managers to identify applicants who could communicate well with staff in hotels, airports, security, guest reception, clinics, hospitals, and the media (ETS, 2008). Moreover, at the Shanghai Expo, recruiters used TOEIC scores to identify qualified volunteers who were talented professionals possessing global perspectives and cross-cultural communication skills (ETS, 2010).

In the airline industry, Air France and International Thai Airways use TOEIC scores to screen staff (e.g., flight attendants and ground staff) on nontechnical English skills to help facilitate communication between staff and passengers and to supplement the technical skills required by the International Civil Aviation Organization (ETS, 2007c, 2011a).

Our analysis of test-use examples collected for this project provided additional insight into the use of TOEIC scores for hiring. For instance, two HR managers from Korea suggested scores, along with grade point average, provide them with a "yardstick to measure job applicants' readiness. "The EPN member from Brazil suggested that because of the TOEIC program's international recognition, its inclusion in hiring attracts talent, as it helps enhance the companies' credibility with international trading partners and employment offers. Moreover, responses from two HR managers from Spain (CNMV and Deloitte) suggested that TOEIC scores help increase confidence that prospective employees will be fluent in English and possess the needed language skills to work collaboratively; network productively; profit from opportunities available in the company's international markets; and be ready to express their knowledge, expertise, and ideas on professional matters clearly and accurately. Furthermore, responses from Taiwan revealed that TOEIC scores are helpful in identifying the staff possessing the needed English skills to attend international conferences and bring firsthand international information into the company to share with colleagues.

Using Scores to Inform Decisions Related to Promotions and Employee Training

Anthony (2003) reported that international businesses use a range of TOEIC scores to inform promotion decisions. For example, IBM Japan and Toyota Automobile use a TOEIC score of 600 points as part of the criteria for promotion to department head. Matsushita Electric uses TOEIC scores of 650 points for promotion into overseas work, and SMK uses a TOEIC score of 730 points for awarding bonuses of 10,000 yen per month. Moreover, consistent with its hire-from-within policy, Procter & Gamble uses the TOEIC Listening and Reading tests to assess whether internal employees have the requisite English skills to be eligible for promotion or whether a professional development plan needs to be implemented to help promote employees to more advanced positions (ETS, 2014b; Stahl et al., 2007).

TOEIC scores are also used to inform training decisions, such as establishing a baseline for the type and level of English training employees need. For instance, NEC employees who receive scores lower than 470 points are assigned to basic English courses focusing on building a solid set of fundamental skills in listening, grammar, and vocabulary. Employees who receive scores between 470 and 725 points are asked to strengthen their basic skills and improve their communication skills in writing and conversation. Employees who score above 730 points focus on acquiring skills that include making presentations, with the ultimate goal of increasing their English-language ability to a level suitable for conducting business smoothly in English (ETS, 2007b). Moreover, the Japanese childcare manufacturer Pigeon Corporation assigns employees whose scores are less than 500 points to an elementary training class and assigns employees who score between 500 and 699 points to an intermediate training class (ETS, 2014a). Furthermore, Bristol-Myers Squibb assigns employees with less than 700 points to an in-house English development training program to enable employees to have seamless communication with colleagues globally (ETS, 2007f).

Companies also use the TOEIC test to monitor progress in English learning. For instance, the Latin America food company Empresas Carozzi uses the TOEIC tests to monitor employee progress in acquiring knowledge of English and asks employees to retake the test at regular intervals (ETS, 2007e). Moreover, the Banyan Tree Samui hotel chain uses the TOEIC tests to help identify areas in need of improvement for employees relative to their job titles (ETS, 2011b).

Our analysis from the responses to the test-use examples also provided insight into the perceived relationship between TOEIC scores and on-the-job performance. Responses from managers from Brazil suggested that TOEIC scores help them gauge employees' readiness to take on more challenging work, including assignments to international posts. For instance, the respondents from Taiwan noted that possessing strong English skills can help companies forgo hiring interpreters to conduct meetings to talk about diverse global issues and can speed up the business decision-making process. They also remarked that the TOEIC score is helpful in identifying employees who will have an easier time adapting to the company's corporate environment, thus potentially reducing employee turnover; that is, the TOEIC tests strengthened the managers' confidence that employees would be less prone to making mistakes, would present information accurately both in internal and external communications, or would be better equipped to understand messages conveyed in meetings. For instance, respondents from Spain (CNMV) suggested that understanding of English facilitates achieving cross-border transactions, such as contract negotiations, mergers and acquisitions, and foreign investments.

Discussion

Our analysis of test-use examples provided insight into how companies use TOEIC scores to inform HR decisions related to hiring, promotion, training, and on-the-job performance. We note the limitations of our sample, as the respondents represent only a subset of all TOEIC users. As such, the sampled respondents may have had a positive bias toward the TOEIC tests. As a result, we may have failed to obtain a fully accurate picture of the diverse uses of the test, particularly negative ones. Nonetheless, our results represent the voices of an important segment of TOEIC users and as such have led to useful insights into test score uses.

We suggest that future studies examine the linguistic skills that are required by the end users of TOEIC tests to investigate which skills both employees and employers require in terms of linguistic and functional communicative proficiency. Such studies should be conducted cooperatively between researchers, teachers, employers, and test developers to collaboratively develop materials and tests that reflect authentic workplace contexts and to clearly lay out the limitations of such measures and the derived inferences. Additional collaboration may involve identifying and building future courses of action for test developers to provide additional support and services to test users (e.g., assessment literacy, the development of assessments measuring additional components of workplace English, or algorithms to help analyze the various variables relevant to informing HR decisions) for more meaningful and relevant score-based decisions and interpretations.

In closing, we reiterate that this study served as an initial step in analyzing consequences of using TOEIC scores on personnel decision making and English-related workplace tasks. It is meant to start a discussion on how users use test scores to inform HR decisions. There were several unanswered questions, which future studies could help address. For instance, such studies should examine the weight attributed to scores in hiring candidates (e.g., are the scores the primary source of evidence, or are they considered in light of additional sources, such as interviews, and if so, which models are used to weight the various sources of evidence used to inform HR decisions?). We suggest conducting studies using multiple methods, such as focus groups, surveys, and/or interviews with HR managers, to investigate test score use in greater depth. We also suggest conducting quantitative studies, such as utility analyses (Boudreau, 1988) with TOEIC scores as a predictor in the model, which would allow us to quantify and describe the impact (usefulness) of TOEIC scores on HR personnel selection processes.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Anthony, L. (2003, September 4–6). *Integrating the TOEIC® into the university English curriculum.* Paper presented at the 42nd annual convention of the Japan Association of College English Teachers, Sendai, Japan.

Boudreau, J. W. (1988). *Utility analysis for decisions in human resource management* (Working Paper No. 88-21). Ithaca, NY: Cornell University School of Industrial and Labor Relations, Center for Advanced Human Resource Studies.

Educational Testing Service. (2007a). *Copper-mining company certifies the English fluency of 60 percent of its supervisors and managers*. Retrieved from https://www.ets.org/Media/Tests/TOEIC/pdf/Minera_Los_Pelambres_Story.pdf

Educational Testing Service. (2007b). *Providing English proficiency solutions to a global solutions company*. Retrieved from https://www.ets.org/Media/Resources_For/English_Language_Learning/NEC_case.pdf

Educational Testing Service. (2007c). *Recruiting made more efficient with TOEIC®: The case of Thai International*. Retrieved from https://www.ets.org/Media/Tests/Test_of_English_for_International_Communication/ThaiInternational.pdf

Educational Testing Service. (2007d). *Recruiting talented professionals: France and Malaysia.* Retrieved from https://www.ets.org/Media/Tests/Test_of_English_for_International_Communication/FranceandMalaysia.pdf

Educational Testing Service. (2007e). *The TOEIC® test is "food for thought" in expansion of company's international marketplace*. Retrieved from https://www.ets.org/Media/Tests/TOEIC/pdf/54160c_toeic_carozzi.pdf

Educational Testing Service. (2007f). *Using English to be a part of the global community*. Retrieved from https://www.ets.org/Media/Resources_For/English_Language_Learning/BMS_case.pdf

Educational Testing Service. (2008). *TOEIC® test scores help recruit qualified volunteers*. Retrieved from https://www.ets.org/Media/Tests/TOEIC/pdf/9788_TOEIC_Regatta_Testimonial.pdf

Educational Testing Service. (2010). *TOEIC® test helps Shanghai Expo recruiters find qualified volunteers*. Retrieved from https://www.ets.org/s/toeic/pdf/toeic_world_expo_testimonial.pdf

Educational Testing Service. (2011a). *TOEIC® Listening and Reading test helps Air France connect with the world*. Retrieved from https://www.ets.org/s/toeic/pdf/toeic_air_france_testimonial.pdf

Educational Testing Service. (2011b). *TOEIC® Listening and Reading test helps Banyan Tree Samui communicate with its guests*. Retrieved from https://www.ets.org/s/toeic/pdf/banyan_tree_samui_success_story.pdf

Educational Testing Service. (2016). *Examinee handbook for TOEIC® Speaking and Writing tests.* Retrieved from https://www.ets.org/s/toeic/pdf/speaking-writing-examinee-handbook.pdf

Educational Testing Service. (2013). *User guide for the TOEIC® Listening and Reading test*. Retrieved from https://www.ets.org/s/toeic/pdf/toeic-listening-reading-test-user-guide.pdf

Educational Testing Service. (2014a). *Leading Japanese company uses TOEIC® test to promote globalization*. Retrieved from https://www.ets.org/s/toeic/pdf/pigeon_corp_japan.pdf

Educational Testing Service. (2014b). *Procter & Gamble invests in English communication*. Retrieved from http://www.toeic.com.hk/english/News/20140826.html

Educational Testing Service. (2014c). *Why English matters* [PowerPoint slides]. Princeton, NJ: Author.

Educational Testing Service. (2015). *Examinee handbook for TOEIC® Listening and Reading test.* Retrieved from https://www.ets.org/Media/Tests/TOEIC/pdf/TOEIC_LR_examinee_handbook.pdf

Harzing, A., Köster, K., & Magner, U. (2011). Babel in business: The language barrier and its solutions in the HQ-subsidiary relationship. *Journal of World Business*, *46*, 279–287. https://doi.org/10.1016/j.jwb.2010.07.005

Institute for International Business Communication. (2011). *TOEIC® newsletter: TOEIC scores for new recruits in FY2011*. Tokyo, Japan: Author.

Kim, H. H. (2013). Needs analysis for English for specific purpose course development for engineering students in Korea. *International Journal of Multimedia and Ubiquitous Engineering*, *8*, 279–288. https://doi.org/10.14257/ijmue.2013.8.6.28

Marra, M. (2012). English in the workplace. In B. Paltridge & S. Starfield (Eds.), *The handbook of English for specific purposes* (pp. 175–192). Malden, MA: John Wiley. https://doi.org/10.1002/9781118339855.ch9

Moritoshi, P. (2001). *The Test of English for International Communication (TOEIC®): Necessity, proficiency levels, test score utilization, and accuracy.* Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.199.475&rep=rep1&type=pdf

Newton, P. E. (2010). The multiple purposes of assessment. In *International encyclopedia of education* (3rd ed., pp. 392–396). Oxford, UK: Elsevier.

Ojanperä, M. (2014). *Effects of using English in business communication in Japanese-based multinational corporations* (Unpublished master's thesis). University of Oulu, Oulu, Finland. Retrieved from http://jultika.oulu.fi/files/nbnfioulu-201402131106.pdf

Park, K. (2013). English language and skill mismatch: The case of South Korea. *African and Asian Studies*, *12*, 391–414. https://doi.org/10.1163/15692108-12341274

Peltokorpi, V. (2010). Intercultural communication in foreign subsidiaries: The influence of expatriates' language and cultural competencies. *Scandinavian Journal of Management*, *26*, 176–188. https://doi.org/10.1016/j.scaman.2010.02.003

Peltokorpi, V., & Vaara, E. (2014). Knowledge transfer in multinational corporations: Productive and counterproductive effects of language-sensitive recruitment. *Journal of International Business Studies*, *5*, 600–622. https://doi.org/10.1057/jibs.2014.1

Piekkari, R. (2006). Language effects in multinational corporations: A review from an international human resource management perspective. In G. K. Stahl & I. Björkman (Eds.), *Handbook of research in international human resource management* (pp. 536–550). Northampton, MA: Edward Elgar. https://doi.org/10.4337/9781845428235.00038

Stahl, G. K., Björkman, I., Farndale, E., Morris, S. S., Paauwe, J., Stiles, P., . . . Wright, P. M. (2007). *Global talent management: How leading multinationals build and sustain their talent pipeline.* Fontainebleau, France: INSEAD.

Appendix

List of Examples Sent to Human Resource Managers

We would like your help in expanding the list of examples associated with low and high levels of English proficiency in the workplace. To illustrate the types of statements we are seeking, we have provided examples. These examples come from our review of published research about English in the workplace. We have organized this information along the potential impact of English on key employment stages such as hiring and selection, training and development, on-the-job performance, promotion, and career mobility as well as international assignments.

Please send us three to five examples based on your direct observations of employees or conversations with human resource managers or others in relation to how English-language proficiency impacts employment decisions.

Impact of English Skills on:

Hiring and Selection

- Hiring applicants with strong English skills reduces the amount of resources spent on language and job training delivered in English.
- Hiring employees with strong English-language skills makes it easier to promote from within if higher level positions require a stronger command of the English language.
- Having a lingua franca strengthens corporate identity; we thus focus our hiring efforts on employees with high levels of English proficiency.
- Language-sensitive recruitment helps narrow down a large pool of applicants to a more manageable number.

Training and Professional Development

- Low proficiency in English may prevent employees from participating in corporate training or professional development programs.
- Low proficiency in English may interfere with how much employees are able to benefit from corporate training or professional development programs delivered in English.
- High English proficiency helps ensure that employees are able to contribute their knowledge and expertise.
- High proficiency in English helps employees accurately and clearly capture the message intended for discussion.
- High proficiency in English helps employees convey their thoughts and ideas in relation to professional matters clearly and accurately.
- High proficiency in English makes it easier to implement new learning or new policies.
- A higher degree of English fluency opens the channels of communication across employees.

On-the-Job Performance

- Proficiency in English leads to creating strong team-building opportunities.
- Proficiency in English has helped develop English-language skills for interaction with managers, to prepare them for visits, and technical inspection from outside the company.
- Low proficiency in English contributes to broken promises and human oversight leading to disappointing service outcomes.
- English-language misunderstandings may stand in the way of major cross-border transactions (e.g., contract with a supplier, merger and acquisition, foreign direct investment), which may in turn cause significant economic losses.

- Low proficiency in English might lead employees to understand only the basic message in a meeting, not the contextual nuances, which can sometimes be critical.
- Low proficiency in English might lead to written reports and e-mails taking longer to write and having more errors.
- Employees' lack of confidence in English leads to customers not having confidence in them, which might lead to doing business with the competitor.
- Low proficiency in English skills might lead individuals to appear "flat," "nonverbal," or "lacking in insight" given their limited vocabulary in English.

Promotions, Career Mobility, and International Assignments

- Limited English-language skills may reduce the chances of being promoted.
- Employees fluent in English will have an easier time developing international networks.
- Employees who are highly proficient in English are more likely to be chosen for international assignments.

TOEIC® Research Studies

126084-126084 • SED618E.5 • Printed in U.S.A.

808527

Copyright © 2018 Educational Testing Service. All rights reserved. ETS, the ETS logo, MEASURING THE POWER OF LEARNING, GRE, GRADUATE RECORD EXAMINATIONS, TEXTEVALUATOR, TOEIC, TOEFL, and TOEFL IBT are registered trademarks of Educational Testing Service (ETS) in the United States and other countries. TOEIC BRIDGE is a trademark of ETS. SAT is a registered trademark of the College Board. All other trademarks are property of their respective owners.

Measuring the Power of Learning.®

www.ets.org