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

Elizabeth Park and Elizabeth Bredlau

November 2014

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 below.

TABLE 1
Example Task Design Analysis for Speaking (Hines, 2010)

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:** The test taker can generate language intelligible to native and proficient nonnative English speakers. 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.). 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, Princeton, NJ: ETS. 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-takers' 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: 1) the fixed elements (*e.g.,* nature of the task, order of question elements) common to all questions of the same task type; 2) the elements, such as topic and type of the stimulus text, that can acceptably vary from question to question, also known as variable elements; 3) the rubric; and 4) 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-takers' responses are scored by qualified, trained, certified, and calibrated raters, using a rubric with a range of points between zero and five.

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:

1) most meaningful for the test-taking population and clients of the TOEIC Speaking test, and
2) 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 below presents a sample of a Propose a Solution test question generated from the expanded list of variants.

FIGURE 1
Example of Propose a Solution Alternate Question Format

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.

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 below presents a sample of a Respond to Questions task generated from the expanded list of variants.
FIGURE 2
Sample of Respond to Questions Alternate Question Format

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?

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) (Qu & Cid, 2014). 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. (RM-99-04, TOEFL-MS-15). Princeton, NJ: ETS.

Hines, S. (2010). Evidence-Centered Design: The TOEIC® Speaking and Writing Tests. (TC-10-07). Princeton, NJ: ETS.

Powers, D. E. (2010). Validity: What Does It Mean for the TOEIC® Tests? (TC-10-01). Princeton, NJ: ETS.

Qu, Y. & Cid, J. (2014). Statistical Analyses for the TOEIC® Speaking Item Expansion Pilot Study. Unpublished Manuscript. Princeton, NJ: ETS.

Appendix A: 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 (6575 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 blue font. Adapted from “Evidence-Centered Design: The *TOEIC*® Speaking and Writing Tests” (p. 7.15), by S. Hines, 2010, Princeton, NJ: ETS. Adapted with permission.