ETS TOEFL

Quality Beyond Measure.

TOEFL® Research Report
TOEFL–RR-99
ETS Research Report No. RR–22-10

Evaluating the New *TOEFL ITP*® Speaking Test: Insights From Field Test Takers

Shinhye Lee

December 2022

The *TOEFL®* test is the world's most widely respected English language assessment, used for admissions purposes in more than 130 countries including Australia, Canada, New Zealand, the United Kingdom, and the United States. Since its initial launch in 1964, the TOEFL test has undergone several major revisions motivated by advances in theories of language ability and changes in English teaching practices. The most recent revision, the *TOEFL iBT®* test, contains a number of innovative design features, including integrated tasks that engage multiple skills to simulate language use in academic settings and test materials that reflect the reading, listening, speaking, and writing demands of real-world academic environments. In addition to the TOEFL iBT, the TOEFL Family of Assessments has expanded to provide high-quality English proficiency assessments for a variety of academic uses and contexts. The TOEFL Young Students Series (YSS) features the *TOEFL® Primary™* and *TOEFL Junior®* tests, designed to help teachers and learners of English in school settings. The *TOEFL ITP®* Assessment Series offers colleges, universities, and others an affordable test for placement and progress monitoring within English programs.

Since the 1970s, the TOEFL tests have had a rigorous, productive, and far-ranging research program. ETS has made the establishment of a strong research base a consistent feature of the development and evolution of the TOEFL tests, because only through a rigorous program of research can a testing company demonstrate its forward-looking vision and substantiate claims about what test takers know or can do based on their test scores. In addition to the 20-30 TOEFL-related research projects conducted by ETS Research & Development staff each year, the TOEFL Committee of Examiners (COE), composed of distinguished language-learning and testing experts from the academic community, funds an annual program of research supporting the TOEFL family of assessments, including projects carried out by external researchers from all over the world.

To date, hundreds of studies on the TOEFL tests have been published in refereed academic journals and books. In addition, more than 300 peer-reviewed reports about TOEFL research have been published by ETS. These publications have appeared in several different series historically: TOEFL Monographs, TOEFL Technical Reports, TOEFL iBT Research Reports, and TOEFL Junior Research Reports. It is the purpose of the current TOEFL Research Report Series to serve as the primary venue for all ETS publications on research conducted in relation to all members of the TOEFL Family of Assessments.

Current (2021–2022) members of the TOEFL COE are:

Lorena Llosa – Chair	**New York University**
Beverly Baker	University of Ottawa
Tineke Brunfaut	Lancaster University
Atta Gebril	The American University of Cairo
April Ginther	Purdue University
Claudia Harsch	University of Bremen
Talia Isaacs	University College London
Yasuyo Sawaki	Waseda University
Dina Tsagari	Oslo Metropolitan University
Koen Van Gorp	Michigan University
Wenxia Zhang	Tsinghua University

To obtain more information about the TOEFL programs and services, use one of the following:

E-mail: toefl@ets.org **Web site:** www.ets.org/toefl

ETS is an Equal Opportunity/Affirmative Action Employer.

As part of its educational and social mission and in fulfilling the organization's non-profit Charter and Bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

RESEARCH REPORT

Evaluating the New *TOEFL ITP*® Speaking Test: Insights From Field Test Takers

Shinhye Lee

ETS, Princeton, NJ

In response to the calls for making key stakeholders' perspectives relevant in the test validation process, the study discussed in this report sought test-taker feedback as part of collecting validity evidence and supporting the ongoing field testing efforts of the new *TOEFL ITP*® Speaking section. Specifically, I aimed to investigate the extent to which test takers' perceptions of the newly proposed ITP test tasks are in agreement with the intended characteristics and qualities of the tasks. In addition, I opted to gather insights into whether the speaking tasks are perceived as acceptable by its prospective test takers and also to identify any unwarranted challenges posed for them in completing the tasks. A two-part questionnaire was thus administered during field testing of the new speaking section, and resulting data were analyzed both quantitatively and qualitatively. Findings emerging from the questionnaire data suggest that test-taker perceptions can be used to provide support to corroborate the intended (or hypothesized) properties of the tasks, while pointing to several areas for further monitoring and improvement.

Keywords *TOEFL ITP*®; validation; test-taker perception; speaking assessment

doi:10.1002/ets2.12352

Central to the current conceptualization of validity and validation practices are the interpretations and uses of test scores as well as the consequences of the test use (Kane, 2006; Messick, 1998). This conceptualization assumes a close linkage between the inferences made about the test takers' abilities and the decisions that ensue from those score-based interpretations (Bachman & Purpura, 2008). Accordingly, it is universally recognized that validation is a process of strengthening an argument about the validity of a particular interpretation or test score use, rather than the test itself (Chapelle, 2020).

The examination of validity, and the quest of assembling validity evidence in particular, has typically been approached from the perspectives of test developers and relevant testing programs (Bachman, 2000). Consequently, it is from these stakeholders that specific interpretations and score uses are proposed, constructed, and thus, evaluated (Chapelle et al., 2008). Yet, there has been a growing recognition in educational measurement in general (Moss et al., 2006; Stricker, 2012), and in language assessment more specifically (Cumming et al., 2004; Shohamy, 2001), of the importance of collecting validity evidence through multiple methods and from different groups of stakeholders. Moss (1996), for instance, argued that test score interpretations and uses become "monologic representations" (p. 24) by virtue of neglecting the perspectives of the core end-users (i.e., those who are assessed or who use the assessment information). Moss further pointed out that these users "often have no ready access to reconstruct or challenge these [interpretations] that may have an impact on their lives" (p. 24). The *Standards* (AERA, 2014) also stipulate that validation is the joint responsibility of both the test developers and the test users. Although the test developer is responsible for rationalizing and collecting evidence in support of the proposed test score interpretations for specified uses, it is the test user who ultimately evaluates such evidence in relation to the specific test score use contexts. Above all, validity evidence encompasses not only what supports claims about the qualities of test scores and their intended uses, but what possibly weakens—or refutes—these claims (Crooks et al., 1996; Haladyna & Downing, 2004). Identifying and challenging these "rival hypotheses" (AERA, 2014, p. 23), particularly from the perspectives of multiple score users, is now recognized as one crucial way to expose sources of evidence that may stand to undermine a desired test score inference (Messick, 1998; Xie, 2011).

Corresponding author: S. Lee, E-mail: lee@ets.org

Among key test user groups, the present study aims to draw on the accounts of test takers. In particular, the focus is on gathering test-taker feedback—"the reactions to and opinions about tests" (Brown, 1993, p. 276)—in the context of field testing a new speaking measure included in the *TOEFL ITP®*. Test-taker perspectives were traditionally dismissed or pursued as supplementary information, at best, due to a long-standing view that they represent nonexpert, "lay-people's intuitive judgment" (Sato & Ikeda, 2015, p. 2) as opposed to scientific evidence of validity (Shohamy, 2001). Researchers have also regarded test-taker perceptions as unreliable, due possibly to the influence of personal attributes (e.g., language proficiency, gender, age) on masking accurate evaluations (Eccles et al., 1983; Scouller & Prosser, 1994). In recent years, however, language testing validation research has branched out to incorporate multiple angles and perspectives, including the potential usefulness of incorporating test-taker feedback at various stages of test development (e.g., Ockey et al., 2013) and as part of a program of test validation (e.g., Malone & Montee, 2014).

Research has demonstrated that test-taker input can be used to contribute to an understanding of how tests elicit language knowledge and skills and whether they are actually measuring the intended construct, thereby corroborating the intended claims and uses of the test (Xie, 2011). These studies have addressed test-taker perceptions primarily from the perspectives of test-taking strategies (Barkaoui et al., 2013; Cohen & Upton, 2007), processes involved in performance (Huhta et al., 2006; Sasaki, 2000), test delivery methods (Fox & Cheng, 2015; Stricker & Ward, 2004), and task characteristics (A. Brown, 1993; Elder et al., 2002). Within this line of research, a few notable studies have elicited test-taker feedback on the task types relevant to the current investigation—that is, the speaking task types included in the *TOEFL®* family of assessments.

Huang and Hung (2017), for instance, interviewed Taiwanese English as foreign language (EFL) learners to probe their strategy use while performing integrated speaking tasks. On one hand, the elicitation of certain strategy uses (e.g., synthesizing information from reading and listening input into oral output) lent support for the validity of the score interpretations based on integrated task performance. On the other hand, the researchers claimed that learners' unanimous belief that topical knowledge was an important factor in integrated performance underscored a critical mismatch between the perceived and intended construct measured by integrated tasks. In particular, the majority of learners pinpointed the pivotal role that insufficient topical knowledge plays in arousing unwarranted anxiety in integrated performance, which, in their own views, had an impact on the way they confidently took on the task and verbalized appropriate responses. Some also reasoned that those who are unfamiliar with the given task topic are essentially bogged down with competing cognitive demands, and the resources needed for the subsequent information integration and oral production, which are the key factors for successful task completion, are likely to be diverted. On these grounds, Huang and Hung further raised the issue of whether the aural and textual input provided in the integrated task indeed operates to compensate for "the (dis)advantage topical knowledge might afford the test takers" (p. 174), as scholars and test developers have previously claimed (Read, 1990; Weigle, 2004). In an international, mixed-methods study, Malone and Montee (2014) reported that test takers of the TOEFL iBT® Speaking section made use of the same type of synthesizing and summarizing strategy when performing integrated speaking tasks as test takers in Huang and Hung's (2017) study. Test takers also believed the independent tasks to be highly relevant to the kinds of speaking activities encountered at the university level. Yet the majority of test takers interviewed in the study also prioritized making use of rehearsed responses (e.g., memorized response patterns, response templates) as well as note-taking skills for organizing answers for both task types. For some, the need to structure responses into "short, organized answers" (Malone & Montee, 2014, p. 31) was a function of time pressure so as to ensure their responses were produced during the given amount of time. This belief was, overall, in stark contrast to that of the language instructors participating as another stakeholder group in the study; they believed that formulas or any type of highly structured response patterns are of less priority in preparing students for the TOEFL iBT Speaking section. Malone and Montee thus concluded that what is expected on the speaking section, at least from the viewpoint of test takers, may not entirely be spontaneous speech (as intended), but rather include the use of formulaic, structured modes of oral language, while making use of specific test preparation strategies as needed.

As such, the potential gap between test-taker perceptions and test developers' intention in developing items arguably underscores the importance of corroborating whether test takers perceive the test to be *as relevant* to the claimed purposes and the skills that it purports to measure (Davies et al., 1999). Specifically, this can be considered to have explicit implications for test validity (Sato & Ikeda, 2015). That is, if the test is perceived as irrelevant to the claimed purposes, test takers may put less effort into the test, and the resulting scores may not accurately reflect their ability (Alderson et al., 1995;

Brown & Abeywickrama, 2010). Kane (2006) accordingly stated that "to the extent that students put less effort into their performance on a test than they would on the corresponding tasks in other settings, because the test seems irrelevant, the extrapolation inference would be weakened" (p. 36).

In more practical terms, test-taker feedback constitutes an effective avenue for informing ongoing test development and revision procedures (Stansfield et al., 1990). The pursuit of stakeholder viewpoints and feedback (drawn from test takers in particular) may ultimately be motivated by a desire to ensure that the proposed test tasks are generally acceptable to end-users (Fulcher & Davidson, 2012). Brown (1993) thus asserted that feedback from test takers during test development can be used to develop a fairer and more accessible test for its prospective users. Kenyon and Stansfield (1991) also made the case for gauging test-taker feedback as part of field testing performance-based tasks. The authors outlined a method in which they used test-taker feedback (collected via feedback questionnaires and written comments) to help explain why a particular speaking task may function poorly and thus need to be revised or deleted from an operational version of the test. The information gathered from test takers during this process, as Kenyon and Stansfield (1991) claimed, can add to the test developer's "storehouse of knowledge with the result that he or she will know with some degree of confidence on how future task types are received by the intended test takers" (p. 14). Building upon this argument, Reed (2014) further recommended the use of multiple data collection methods, particularly those described as reflective techniques (e.g., recall protocol, interviews, open-ended questions), in order to elicit test-taker perspectives on various aspects of the trialed test items and tasks. He stressed that this research can have important implications for test design revisions but can serve (more immediately) to prevent major problems from occurring in subsequent, large-scale testing stages.

In summary, test takers' perceptions may be critical for bringing to light a number of factors contributing to and potentially revealing possible threats to the intended score interpretations (Messick, 1998). Such evidence also is of practical relevance and has implications for informing and improving ongoing test development and revision endeavors (Kenyon & MacGregor, 2012). As part of recent field testing efforts of the TOEFL ITP program, the current study thus set out to evaluate how the newly assembled task types (read-aloud, independent, integrated) included in the TOEFL ITP Speaking section are perceived and evaluated by prospective test takers. Based on the aforementioned studies, I opted for gaining insights broadly into two (interrelated) areas in particular: (a) the extent to which the prospective TOEFL ITP test takers' viewpoints correspond to the intended characteristics and purposes of the TOEFL ITP Speaking tasks (as being authentic, eliciting relevant speaking skills, testing at an appropriate and intended difficulty level) and (b) the specific challenges inherent in performing the tasks that point to the need for further areas of improvement or revision. Specifically, following Malone and Montee (2014) and responding to the calls of Reed (2014), I aimed at evaluating the test section from multiple angles, which involved conducting both quantitative and qualitative analyses of test-taker perception data. In terms of the former, I gauged and comparatively analyzed feedback provided to the differing task types on the basis of ratings on a series of Likert scale survey items. These perceptions were further corroborated with qualitative written feedback in order to discover which features of test items and the testing situation in particular are of concern to the test takers. To this end, two specific research questions guided this study:

Research Question 1: What are test takers' perceptions toward the TOEFL ITP Speaking tasks, as evidenced from the Likert scale survey items? Do perceptions differ across task types?

Research Question 2: Which specific concerns do test takers have about completing the TOEFL ITP Speaking tasks, as evidenced from written comments?

The Context

As part of the TOEFL family of assessments, the TOEFL ITP assessment series is designed to measure nonnative English speakers' proficiency in academic English. The TOEFL ITP is a selected-response assessment comprising three test sections: listening comprehension, structure and written expression, and reading comprehension. The test has been made available to colleges, universities, and other providers of English language programs as an on-demand assessment, suited for serving formative (e.g., student placement in English language programs, student progress monitoring) as well as lower-stakes admissions purposes (e.g., admissions to short-term, nondegree programs). See the website http://www.ets.org/toefl_itp/use. Its practicality aspects, such as cost effectiveness, fast score turnaround time, and flexibility in large-scale test administration, have been the driving force behind the growing implementations and wide-ranging applications observed in a variety of institutional contexts (Golubovich et al., 2018).

Alongside the global popularity of the TOEFL ITP uses, recent years have also witnessed the growing needs of score users to expand the test construct and make comprehensive decisions about test takers' proficiency in all relevant skill areas (Golubovich et al., 2018). Specifically, as Collins and Miller (2018) stated, measurement of productive skills—speaking and writing—was recommended as they reflected upon the test takers' ability to communicate and use English. In response to these demands, the TOEFL ITP program has recently explored the addition of a stand-alone speaking test to the current test battery, with the aim of launching its operational version in early 2022.

For the design of speaking tasks, the task types from the TOEFL family of assessments have been initially considered by test developers who are familiar with the TOEFL ITP population. On the one hand, this was deemed as a reasonable starting point, given that the assessments cover the same language-use domain (i.e., language ability for academic purposes), thereby establishing a conceptual justification for using the same types of tasks. On the other hand, the existing TOEFL assessments also afforded a large pool of available items to pull from, which was beneficial given the limited time and expenses to cover new task development. To this end, the TOEFL ITP Speaking section comprised three specific task types selected across the *TOEFL Junior®* and TOEFL iBT Speaking sections: namely, the read-aloud, independent (describe-and-explain), and integrated (listen-and-speak) tasks.

The read-aloud task was selected and modified from the same task type offered in the TOEFL Junior test. The rationale was that the task type allows individuals with relatively low English language skills—who constitute the target TOEFL ITP population in general (Golubovich et al., 2018)—the opportunity to demonstrate foundational skills of speaking (e.g., pronunciation) without coming up with ideas to express (see also Evanini et al., 2015). The independent and integrated speaking task types from the TOEFL iBT (Tasks 1 and 5 in particular) were deemed relevant because the language-use contexts are similar for both assessments—that is, English for academic purposes. On a practical note, the two specific tasks were also made available for TOEFL ITP as a result of the 2019 revision of the TOEFL iBT Speaking section (Papageorgiou et al., 2019). Intertwined with the need to shorten the test section, the tasks were found to not discriminate well among the TOEFL iBT test takers and were thereby deemed necessary to be removed. Given that TOEFL ITP targets a lower proficiency level than TOEFL iBT, test developers decided to use the tasks as the basis for the new speaking test.

Although the TOEFL ITP Speaking tasks were modeled after the existing speaking tasks, some modifications were necessary to make the tasks suitable for its prospective test takers. In Fall 2019, a prototype study was conducted involving a total of 150 test takers with the aim to further inform these revision efforts and rubric development. Most salient areas for improvement were noted for the two TOEFL iBT-based items. For instance, test takers found it particularly challenging to properly understand the double negative used in the prompt of an independent task ("discuss the advantages—or disadvantages—of not doing something"). Although this level of complexity might have discriminated well with the TOEFL iBT test takers, it was determined that it should be avoided with the ITP population. Thus, the prompt was revised to eliminate the double negative for the larger-scale field testing. In terms of the integrated tasks, many test takers generally struggled to include details from the listening stimulus (conversation). Task directions and lead-ins were accordingly revised to simplify the language and provide more context and increased preparation time. For the listening stimulus, only the simplest of problems and solutions in the given conversation were included.

The Study

The present study was coordinated with a larger field test of the new TOEFL ITP Speaking section that was conducted in 2020–2021. The field test involved 1,211 students from 34 institutions that represented three geographical regions—East Asia, Southern Asia, and Latin America—where the ITP has the highest testing volume. The focus of the present study was to gauge test takers' perceptions toward the proposed speaking tasks and further probe the areas of challenges that they may face when performing the tasks.

Methods

I begin this section by outlining the details of the study instruments and the participating test takers. I then report on procedures taken for arranging the final data set and conducting data analysis to answer the guiding research questions.

Table 1 Task Specifications for the *TOEFL ITP®* Speaking Section

Task type	Description	Number of questions	Preparation and response time (in seconds)	Rubrics
Read-aloud (Task 1)	Test takers listen to a campus-related announcement shown on the screen and read it out loud.	1	Preparation time: 60 Response time: 60	Holistic scale of 1–4 Criteria: Fluency, clarity, prosody, and accuracy
Independent (Tasks 2 and 3)	Test takers describe or explain a personal choice involving a variety of familiar, everyday subjects.	2	Preparation time: 45 Response time: 45	Holistic scale of 1–5 Criteria: Delivery, language use, and topic development
Integrated (Task 4)	Test takers listen to a conversation about a campus-related problem and summarize the problem and express a personal opinion about the solutions.	1	Preparation time: 45 Response time: 60	Same as Independent task

TOEFL ITP Speaking Test Forms

Two test forms were created for the field test data collection. Each form included four speaking tasks in the following order: one read-aloud task, two independent tasks, and one integrated task. The speaking test forms were delivered on computer. Table 1 lists the specifications for the speaking test section.

In both test forms, the tasks were ordered according to the expected difficulty. The read-aloud task (Task 1) presented a campus-related announcement (120–170 words) both aurally and visually. It required test takers to read aloud the announcement verbatim. The independent tasks (Tasks 2 and 3) elicited test takers' opinions on familiar and accessible topics. The integrated task (Task 4) required test takers to summarize a problem they heard and provide a solution after listening to a short conversation between two speakers on a campus-related topic.

Questionnaire

A two-part questionnaire was administered within the testing administration platform (i.e., ITS) immediately after test takers had completed the speaking tasks. The first section of the questionnaire related to the speaking tasks, and the other concerned test takers' English speaking skills in general. For the purpose of the current study, I report on the perception data from the first section, as they were meant to evaluate the speaking tasks of interest.

Based upon a review of previous literature (e.g., Kenyon & Stansfield, 1991; Malone & Montee, 2014), the first section of the survey was designed primarily to examine the extent to which test takers perceived the speaking tasks as appropriate and useful in measuring their speaking ability. Accordingly, it was further divided into two parts: Part 1 presented test takers with five Likert scale questions for each of the four speaking tasks, and Part 2 contained an open-ended question, inviting test takers to freely comment on their testing experiences in relation to completing the tasks and any additional concerns they wished to express about the test. All survey items were presented in English in the survey, but for Part 2, test takers were encouraged to express their thoughts in their respective first languages if needed.

Each set of five items in Part 1 was presented along with a screenshot of the corresponding speaking task for test takers' reference. Table 2 details each survey item in Part 1 in terms of its accompanying prompt and response scale as presented in the survey.

The Likert scale questions collectively aimed to gather test takers' viewpoints on the salient issues pertaining to the tasks. For instance, Question 1 concerned the perceived difficulty of the tasks, and Question 2 elicit ratings on whether test takers felt that the tasks afforded them the opportunity to produce speech indicative of their current level of English speaking ability. Along lines similar to those of Question 2, Question 3 asked test takers to self-evaluate their spoken performances on the respective tasks. Questions 4 and 5, respectively, touched upon the domain of authenticity of the tasks (in relation to real-world language-use tasks) and the overall usefulness of the task as a measure of English proficiency.

Table 2 Questionnaire Items and Scale Options

Number	Prompt	Scale option
1	How difficult or easy was this task for you?	1 = Difficult 2 = Somewhat difficult 3 = Neither difficult nor easy 4 = Somewhat easy 5 = Easy
2	Think about your ability to speak in English. How well did your response to this task demonstrate your English speaking ability?	1 = Not well 2 = Not too well 3 = Neutral 4 = Slightly well 5 = Well
3	If others evaluate your English speaking ability based on your response to this task, on a scale of 1 (poor) to 5 (excellent), what score do you expect to receive?	1 = Poor 2 = Below average 3 = Average 4 = Above average 5 = Excellent
4	Think about the English-speaking activities you do in real life. How similar is this task to such real-life speaking activities?	1 = Different 2 = Somewhat different 3 = Neither different nor similar 4 = Somewhat similar 5 = Similar
5	This task was a good way to measure my English ability.	1 = Disagree 2 = Slightly disagree 3 = Neither disagree nor agree 4 = Slightly agree 5 = Agree

Test Takers

The final data set used for the current study came from 616 test takers. Of the 1,211 test takers who initially participated in the field test, 859 test takers agreed to fill out the questionnaire (71% response rate). However, data from 616 test takers were made available for the final analysis after conducting a series of data validation procedures. Specifically, survey responses that were incomplete or tagged with undiscernible IDs were excluded. Responses indicative of low-quality data were also excluded: for example, “speeded” or inattentive responses (i.e., responses associated with unreasonably short response times).

Table 3 provides the demographic information about the sample included in the study. As mentioned, the sample was from the three geographical regions that represent the ITP target testing contexts: East Asia (39.3%), Southern Asia (35.4%), and Latin America (25.3%). Female test takers accounted for over half of the overall sample (55.4%). About two thirds (62%) of the test takers were pursuing degrees at a higher education institution, either in a graduate (39%) or an undergraduate program (23.1%). Another one third (37.9%) were students at secondary-level schools. In terms of language-learning background with regard to English, about two thirds (61.4%) of the sample had between 5 and 10 years of experience in learning English.

Data Analysis

Research Question 1: Quantitative Data

To answer the first research question (*What are test takers’ general perceptions toward the TOEFL ITP Speaking tasks as evidenced from the Likert scale survey items? Do perceptions differ across task types?*), data from the first section of the questionnaire were analyzed. Answers for each Likert scale item were numerically coded on a scale of 1–5, so a 5 represented the most favorable response (e.g., easiest task), and a 1 denoted the least favorable response (e.g., hardest

Table 3 Demographics of the sample included in the study ($N = 616$)

Category	N	%
Total	616	100
Nationality		
East Asia[a]	242	39.3
Southern Asia[b]	218	35.4
Latin America[c]	156	25.3
Gender		
Male	275	44.6
Female	341	55.4
Education status		
Graduate student	240	39
Collegiate level	142	23.1
Secondary level	234	37.9
English learning experience		
1 year or less	59	9.1
1 – 5 years	194	31.5
5 – 10 years	378	61.4

[a] East Asia consisted of examinees from China and Japan. [b] Southern Asia consisted of examinees from Indonesia, Myanmar, and Thailand. [c] Latin America consisted of examinees from Brazil, Colombia, and Mexico.

task). To better summarize and compare the trends associated with the differing types of speaking tasks, I plotted the perception data (frequency of responses) in bar graphs. I further computed a series of repeated measures ANOVA to determine whether test takers' perceptions as indicated in the five survey items differed as a function of task type. In the Results section, I report the effect size as partial η^2, conforming to the range of $\eta^2 = .01$, .06, and .14 as small, medium, and large effect sizes, respectively (Cohen, 1988).

Research Question 2: Qualitative Data

To answer the second research question (*Which specific concerns do test takers have about completing the TOEFL ITP Speaking tasks as evidenced from written comments?*), open-ended responses in the form of written comments in the survey were analyzed to further shed light on the quantitative findings. Of 616 test takers' survey responses, comments from 178 test takers were available for analysis. A subset of responses provided in the respective first languages of the test takers (Spanish: $N = 18$; Indonesian: $N = 14$) were translated using the Amazon AWS Translate app. The Spanish responses were further cross-referenced by a research assistant, a native speaker of Spanish. The same research assistant then entered all verified, translated comments into an Excel spreadsheet for further analysis.

The following coding procedures were adopted to analyze the written comments: (a) initial, open coding phase (Friedman, 2012) for developing a preliminary set of coding schemes, (b) axial and selective phases (Strauss & Corbin, 1998) for refining and collapsing the initial coding schemes and gauging specific patterns in responses, and (c) double and individual coding sessions. As a first step, I imported the prepared Excel data file into NVivo (version 12) in which the written comments were closely examined to establish overarching themes. Specifically, two major categories emerged from the data set: namely, comments pertaining to the speaking test in general and the task types in particular. I then produced subcategories of themes as they emerged, which were later refined into 10 categories upon multiple rereads of the data set. The research assistant, serving as the second coder, then reviewed the generated codes and was trained on coding the written data. To establish intercoder agreement, the research assistant and I jointly coded approximately 20% of the data. Because the written comments were short and straightforward, the agreement between us was high (91%), which further verified the use of the coding schemes. The few cases of disagreement were resolved through a series of discussions; during this process, two coding categories (from the initial 10) were incorporated into existing themes based on low usages and fit with the data. Based on this result, I continued to independently code the rest of the data.

In the following Results section, I report findings aggregated for the frequencies pertaining to a particular coding category as well as any relevant quotes that stand to complement the trends depicted in the Likert responses.

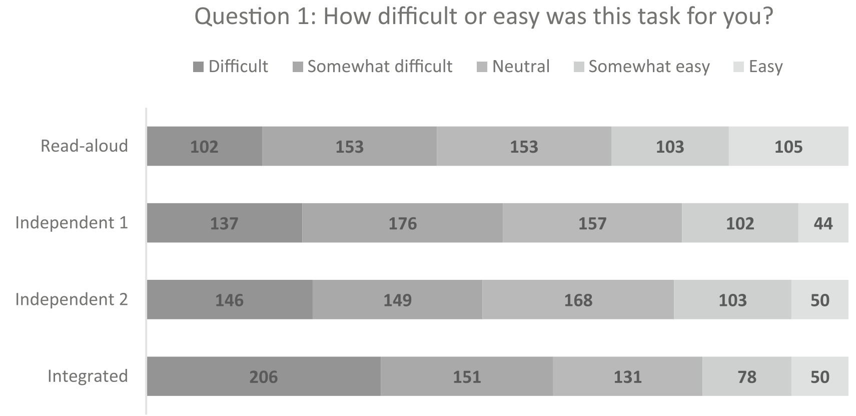

Figure 1 Stacked bar graphs of the frequency of responses to question 1 ($N = 616$). Each row represents results obtained for each task, with individual colored bars respectively representing the five response categories.

Results

Research Question 1

Test takers' responses to the questionnaire are provided for each of the five Likert scale survey questions in the bar graphs. The results for Question 1 are in Figure 1, Question 2 in Figure 2, Question 3 in Figure 3, Question 4 in Figure 4, and Question 5 in Figure 5.

Question 1: How Difficult or Easy Was this Task for you?

As depicted in Figure 1, a noticeable trend in the responses to this question lay in the differing degrees of perceived task difficulty assigned across the task types. Responses to the independent and integrated tasks were resorting to the latter two unfavorable response categories, *difficult* and *somewhat difficult*; in contrast, *somewhat difficult* and *neutral* were the two most common responses for the read-aloud task. This tendency in the responses was most clearly demonstrated for the integrated task as evidenced in the concentration of responses to the category *difficult*; this corresponded to about one third ($N = 206$; 33%) of test takers in the sample.

A one-way repeated measures ANOVA with a Greenhouse–Geisser correction was conducted to determine whether the perceived degree of task difficulty varied as a function of task type. The results showed a significant main effect of task type on test takers' perceptions with a medium effect size, $F(2.77, 1845) = 34.94$, $p < .001$, partial $\eta^2 = .07$. A series of paired samples t-tests further corroborated significant mean differences in ratings among the difficulty assigned to the tasks, particularly between the integrated task and the rest of the task types. The biggest effect size was found for the mean differences between read-aloud and the integrated task ($t[616] = 10.89$, $p < .001$, $d = .42$), whereas those between independent and integrated tasks were smaller (Independent 1 and integrated: $t[615] = 4.37$, $p < .001$, $d = .16$; Independent 2 and integrated: $t[615] = 4.85$, $p < .001$, $d = .16$). The two independent tasks were found to be perceived as similar in difficulty.

In general, test takers attributed relatively higher degrees of difficulty for the independent and integrated tasks and lower for the read-aloud task, a finding that aligns with the intended difficulty of the respective tasks. Furthermore, among all task types, it was the integrated task that the test takers were collectively and markedly evaluating as most difficult, which is in line with previous research findings (Huang & Hung, 2017).

Question 2: How Well Did Your Response to This Task Demonstrate Your English Speaking Ability?

As shown in Figure 3, nearly half of all test takers evaluated their performances as not reflective of their actual speaking abilities, as indicated in the concentration of responses in the bar graphs for the unfavorable response categories *not well* and *not too well*. Notably, it was the integrated task that prompted more unfavorable perceptions, with over half of all test takers in the sample indicating that their performance on this task was not reflective of their actual speaking abilities

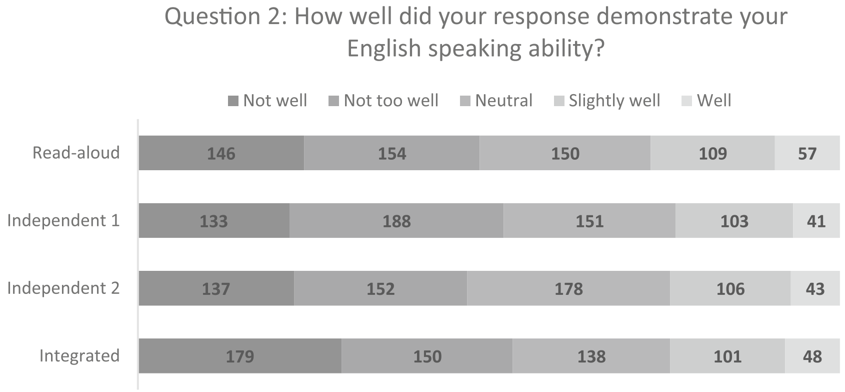

Figure 2 Stacked bar graphs of the frequency of responses to question 2 ($N = 616$). Each row represents results obtained for each task, with individual colored bars respectively representing the five response categories.

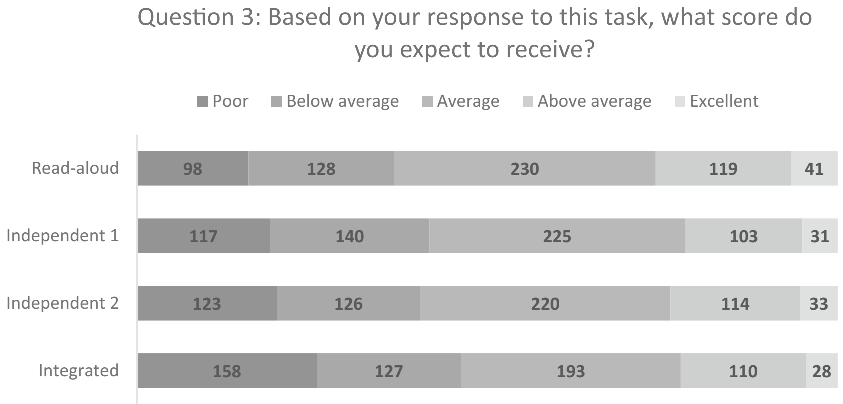

Figure 3 Stacked bar graphs of the frequency of responses to question 3 ($N = 616$). Each row represents results obtained for each task, with individual colored bars respectively representing the five response categories.

($N = 179$; 53.5%). Fewer than 10% of test takers, on the other hand, indicated that they were able to demonstrate their abilities well for all task types.

Results from a one-way ANOVA with Greenhouse–Geisser corrections showed a significant effect for task type but with a marginal effect size: $F(2.80, 1845) = 5.16$, $p = .002$, partial $\eta^2 = .008$. A series of paired samples t-tests further verified the rather consistent response patterns depicted in Figure 1, with a significant mean difference in ratings noted only between the read-aloud and the integrated task, $t(615) = 3.477$, $p < .001$, $d = .11$.

Overall, test takers tended to have similar views regardless of task types; most concurred that their performances on the tasks were less likely to be indicative of their English speaking ability. This viewpoint was relatively more salient for the integrated task, a finding that seems to align with the result depicted in Figure 1. With the integrated task being perceived as challenging in general, test takers may have been inclined to feel that they were not able to demonstrate their ability to the fullest on this particular task type. The read-aloud task, on the other hand, garnered favorable views relative to the other task types, due possibly to the common perception that it was the easier task.

Question 3: On a Scale of 1 (Poor) to 5 (Excellent), What Score Do You Expect to Receive Based Upon Your Performance?

A consistent finding for Question 3 is that the responses were relatively similarly distributed across distinctive response categories. As shown in Figure 3, this was particularly the case among categories of *poor*, *below average*, and *above average*, regardless of task types. A notable exception, however, was the visible peaks displayed in the bar graphs representing *average*; over one third of all test takers in the sample evaluated their performance as *average*. It was only for the integrated task that the trend was somewhat weakened, with a noticeable uptick in the responses for *poor* ($N = 156$; 25%). On the

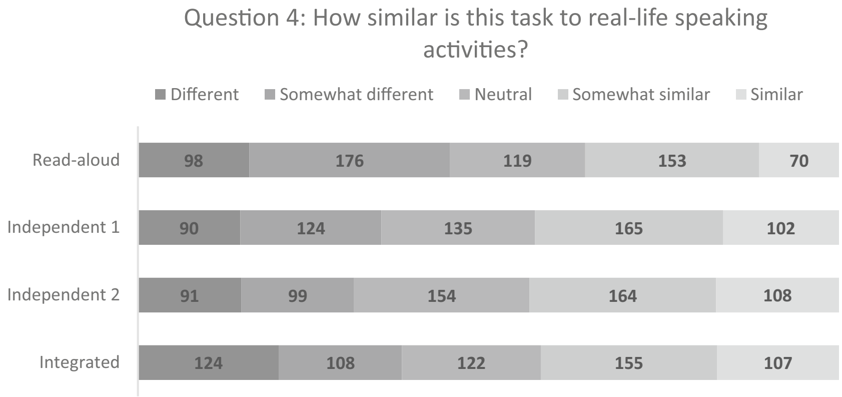

Figure 4 Stacked bar graphs of the frequency of responses to question 4 ($N = 616$). Each row represents results obtained for each task, with individual colored bars respectively representing the five response categories.

other hand, a marginal subset of test takers related their performance across all task types to the most favorable response category, *excellent*. Yet notably, slightly more test takers were inclined to self-evaluate their performance in accordance with this category on the read-aloud task.

A one-way ANOVA of these data indicated a significant effect of task type with a small to medium effect size, $F(2.92, 1845) = 21.76$, $p < .001$, partial $\eta^2 = .03$. Similar to the results for Question 1, post hoc paired samples t-tests revealed that self-assessments were significantly different, particularly between the integrated task and the remaining task types. Above all, the mean difference in self-assessments between the read-aloud and the integrated task exerted the biggest effect size, $t(615) = 7.70$, $p < .001$, $d = .21$. There was no significant difference in the self-evaluations noted between the two independent tasks.

Overall, the findings suggest that the response patterns were mostly varied, encompassing both favorable and unfavorable and, at times, neutral self-assessments across task types. At the same time, test takers were relatively favorable toward their performances on the read-aloud task, and vice versa for the integrated task, a trend in line with findings from Questions 1 and 2.

Question 4: How Similar Is This Task to Real-Life Speaking Activities?

In response to Question 4, test takers tended to similarly evaluate the independent and integrated tasks, as demonstrated in Figure 4. More precisely, test takers indicated that the tasks are similar (or at least somewhat similar) to the kind of English-speaking activities carried out in real life. On the other hand, fewer test takers perceived the read-aloud task in the same manner. In fact, perceptions of the read-aloud task seemed to be rather mixed, although pertaining slightly more to the response category *somewhat different*. This trend in the data represented about one third ($N = 176$; 28.6%) of all test takers in the sample.

A one-way ANOVA confirmed a significant effect of task type on the extent to which test takers perceived the tasks to be authentic and relevant, $F(2.83, 1845) = 18.57$, $p < .001$, partial $\eta^2 = .03$. The read-aloud task, as indicated from the paired samples t-tests, was associated with a lower degree of authenticity relative to the other task types. The biggest effect size observed was for the mean difference in ratings between the read-aloud and Independent Task 2, $t(615) = -6.78$, $p < .001$, $d = .23$. Moreover, although with a marginal effect size, the same independent task was rated significantly higher in authenticity than the integrated task, $t(615) = 3.52$, $p < .001$, $d = .10$.

In sum, the results showed that test takers viewed the independent and the integrated tasks as relevant and authentic in relation to the English-speaking activities they carry out in real life. The read-aloud task, on the other hand, was viewed as less relevant to the test takers' actual language-use contexts.

Question 5: This Task Was a Good Way to Measure My English Ability

For Question 5, the response patterns were similar across all four tasks; that is, a majority of test takers consistently held positive viewpoints as to whether the tasks were good measures of their English ability. Figure 5 further demonstrates

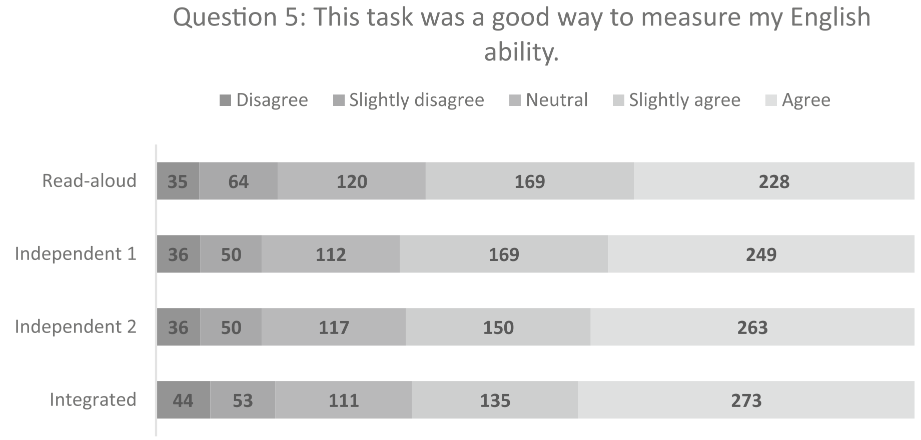

Figure 5 Stacked bar graphs of the frequency of responses to question 5 ($N = 616$). Each row represents results obtained for each task, with individual colored bars respectively representing the five response categories.

that such a favorable reaction is relatively more evident for the independent and the integrated tasks, as indicated by the concentration of responses for the favorable response category *agree* and the least for *disagree*. This gap in responses (i.e., between the most favorable and least favorable options), on the other hand, seemed the least evident for the read-aloud task.

With small to marginal effect size, a one-way ANOVA revealed a significant effect for task types, $F(2.68, 1845) = 3.24$, $p = .026$, partial $\eta^2 = .005$. The paired samples' t-tests further revealed that, similar to the previous findings, significant mean differences in ratings mainly existed between the read-aloud and the other task types. In particular, the read-aloud task was consistently evaluated lower relative to the Independent Task 1 ($t[615] = -2.40$, $p = .017$, $d = .06$), the Independent Task 2 ($t[615] = -2.50$, $p = .011$, $d = .08$), and the integrated task ($t[615] = -1.91$, $p = .050$, $d = .06$), although the latter was not statistically significant. All effect sizes, however, were generally marginal.

Thus, the results appear to show that all task types consistently garnered favorable reactions from test takers in terms of their usefulness as a measure of English speaking ability. Among the task types, however, test takers were more positive about the independent and the integrated tasks relative to the read-aloud task, although the independent and integrated tasks were perceived as more difficult.

Research Question 2

For the open-ended question in Part 2 of the questionnaire, 178 test takers (29% of the sample) shared their overall opinions about the speaking test. Of all the available written comments, 141 test takers (79%) commented on the test as a whole, and a smaller subset of 37 test takers (21%) provided comments on a specific task type. Accordingly, I categorized the written comments into two major categories: "General" and "Task-specific." Note that the latter responses did not specifically differentiate between the two independent tasks (Tasks 2 and 3); hence, a uniform category of "Independent task" was used to represent any relevant comments made about the particular task type. As Table 4 summarizes, the general responses were categorized into eight major themes that provided a glimpse into test takers' overall testing experiences as well as the features of the test that they had perceived as potentially challenging. In what follows, I present the direct quotes and comments that test takers made with regard to the major themes.

General Comments: Positive Test-Taking Experiences

A positive outcome was that about one third (29.1%) of test takers reported having enjoyed the experience of completing the speaking test in general. Some comments corroborated the Likert responses discussed above, particularly in relation to favorably evaluating the authenticity (e.g., "*it was ok for me … similar to situations we have in daily life*") as well as the usefulness of the test as measuring English speaking ability (e.g., "*This test is good to measure our speaking skills*"; "*It is good. Directly measure the ability to speak on various topics*").

Table 4 Question 6: Additional Comments About The Speaking Test

		Task-specific ($N = 37$)		
Theme	General ($N = 141$)	Read-aloud ($N = 11$)	Independent ($N = 17$)	Integrated ($N = 9$)
Short amount of time	34% (48)	18.2% (2)	52.9% (9)	22.2% (2)
Positive experience	29.1% (41)	–	–	–
Technical difficulty	10% (14)	–	5.9% (1)	11.1% (1)
Self-reflection on performance	7.8% (11)	–	–	–
Test anxiety; nervousness	7.8% (11)	9.1% (1)	–	–
Test difficulty	5.7% (8)	9.1% (1)	5.9% (1)	–
Task contents	4.3% (6)	–	35.3% (6)	44.4% (4)
Task instructions	1.4% (2)	63.6% (7)	–	22.2% (2)

Note. The dash (–) indicates no specific responses coded for the theme.

Another 11% of responses pertained to test takers evaluating the test in relation to their performances. Relevant test takers shared that the test motivated them to enhance their speaking skills in general as well as focus on specific aspects of their speaking ability that they need to further improve. Three representative comments included the following:

Example 1. *Test Taker 193*.

“I was surprised there is a speaking test, but surely it help me to improve my ability to speak, not just ability to read and listen.”

Example 2. *Test Taker 145*.

“This is my first time to do the ITP test, I’m a little nervous. But, I know I have lots of to improve, like speaking and listening. I also should improve my speed of speaking.”

Example 3. *Test Taker 137*.

“The speaking test was very good for measuring the speaking ability, and I have a lot to learn in order to accomplish this skills.”

General Comments: Factors Contributing to Test Difficulty

Corroborating the previous findings on the overall difficulty of the tasks (see Figure 1), the majority of written comments pinpointed a variety of challenging factors that test takers encountered while completing the test. Of all general responses, more than one third of the comments (34%) were primarily about the time limits imposed on the tasks—for planning and responding—being too short (e.g., “*Give a little bit more time for the response*,” “*We need some long time to talk*”). Notably, these comments seemed to reflect the belief that the amount of time provided in the test had limited test takers’ ability to complete the tasks, thereby inadequately representing their capabilities. Moreover, some comments pointed toward the specific aspects of the spoken performance that seemed to have been the most impacted. The following comments exemplify the key concerns expressed.

Example 4. *Test Taker 1*.

“It’d be okay to leave a bit less time to prepare and give us more to reply. Many of us do have what we actually consider good fluency and coherence when speaking but I didn’t feel I got the chance to actually prove it.”

Example 5. *Test Taker 311*.

“Give the speaking tests a few more seconds to prepare and perform. Sometimes, it was cut short way too abruptly when I was about to finish my speech, and way too soon for me to prepare a proper sentences. I was not able to show what I can do.”

Example 6. *Test Taker 10.*

“You should consider to extend the response time to give the candidate chance to explore more regarding their own speaking ability. Also, always give extra time in any aspect of the test for candidates with blind or visual impairment.”

A subset of comments (10%) revealed that test takers experienced technical issues (e.g., audio cutting off during the test) arising from the testing platform designed specifically for field testing. A few other test takers (7.8%) touched upon the emotional aspects of taking a speaking test—interestingly, about half of these responses came from novice test takers of either the TOEFL ITP or a computerized speaking test in general. In particular, these sensations of anxiety, as reported by the test takers, were exacerbated by the presence of a timer, as well as the amount of time imposed. The following comments illustrate these observations:

Example 7. *Test Taker 45.*

“I’m so nervous because I can see the count downtime. It makes me can’t focus. Not sure my answer is good.”

Example 8. *Test Taker 130.*

“I was very nervous because it was my first online application test, and also because seeing time, the ideas didn’t come to my mind easily.”

Task-Specific Comments: Read-Aloud Task

In terms of the Task-specific category, the trend was that a specific task type was related to a particular theme. As Table 4 shows, the majority of comments (63.6%) for the read-aloud task, for example, addressed the need for more clarity in task instructions. That is, test takers were most confused about what they were being asked to do; specifically, they were torn between reading the excerpt as-is or providing its summary. The differing instructions given between the written (“read the text”) and aural (“respond to the task”) input for the task also seemed to have enhanced confusion. Some representative comments included the following:

Example 9. *Test Taker 4.*

“the instructions for the first question are not clear, because in the voice it is said to respond, so it’s a bit confusing whether to respond or just read.”

Example 10. *Test Taker 114.*

“I wasn’t sure about the task on the first question, I wasn’t sure if I was asked to response or to read the text. I think you should reformulate the instructions so they can be understandable.”

One particular comment identified a concern about accessibility aspects of the task, given that it requires test takers to directly read the text presented on screen. The comment pinpointed how such delivery would disadvantage test takers with special needs.

Example 11. *Test Taker 28.*

“This kind of test is not fully accessible for visually impaired persons. The first task was an impossible task for blind people.”

Task-Specific Comments: Independent Tasks

For the independent task type, many comments (52.9%) focused on the time limit imposed for preparation. These comments generally suggested that the given 45-second planning time was too short for formulating a response, let alone brainstorming and synthesizing ideas. Example 12 exemplifies these comments:

Example 12. *Test Taker 343.*

“so little time to prepare the answer. even if I should speak in bahasa, I need more time to prepare the answer of advantage of social media, or the reason of my decision. I can’t delivered in 45 seconds.”

The topic of the task was the second most frequent theme noted for the independent tasks. Specifically, over a third of responses (35.3%) reflected a belief that the topics were either not culturally universal or biased toward certain groups of individuals equipped with relevant background knowledge. The following two examples demonstrate these comments.

Example 13. *Test Taker 616*.

“Specific question such as, school subject or childhood friends, are difficult. Because not everyone has one, and making up stories about it, and in nonnative language make it harder.”

Example 14. *Test Taker 112*.

“Maybe a more general topic on the second question and make it less personal so we can all talk about it.”

Task-Specific Comments: Integrated Task

The integrated task was the least commented on, a finding that seemed to be at odds with the results from the Likert questions that are consistently noted for their high task difficulty. For this task, relevant comments were made on the integration of the listening skills (44.4%) in particular; test takers perceived it as irrelevant for strictly measuring one’s speaking ability. Example 15 is a particular case in point:

Example 15. *Test Taker 564*.

“This speaking test also examined the listening test. When I didn’t concentrate on particular thing because of some distraction, I felt lost.”

The multiple requirements of the integrated task—that is, to summarize a given problem, list possible solutions, and reason which solution is better—seemed to have posed a problem for the test takers. Specifically, Example 16 describes the added pressure of meeting all such requirements within the short amount of time given for responding.

Example 16. *Test Taker 176*.

“We don’t have to make up stories like questions 2 and 3, but the duration of answering a lot of things in 1 minute, I think is not enough, it’s 3 questions in one question.”

In summary, the qualitative survey data relevant to Research Question 2 showed a mixture of both positive and negative reactions to the test as a whole as well as to specific task types. In particular, it was apparent that test takers were able to pinpoint specific factors contributing to the challenges that they had faced during testing, and by extension, how these elements had posed problems for adequately demonstrating their speaking abilities. These factors are further elaborated upon in the following section.

Discussion

In response to the calls for making key stakeholders’ perspectives relevant in the test validation process (Moss et al., 2006), the study discussed in this report sought test-taker feedback as part of collecting validity evidence and supporting the ongoing field testing efforts of the new TOEFL ITP Speaking section. Specifically, I aimed at investigating the extent to which test takers’ perceptions of the newly proposed ITP test tasks are in agreement with the intended characteristics and qualities of the tasks. In addition, I opted for gathering insights into whether the speaking tasks are perceived as acceptable by its prospective test takers and identifying any unwarranted challenges posed for them in completing the tasks. A two-part questionnaire was thus administered during field testing, from which resulting data were analyzed both quantitatively and qualitatively.

Four major findings emerging from the questionnaire data suggest that test takers’ perceptions can be used to provide support to corroborate the intended (or hypothesized) properties of the tasks. First of all, responses to Question 1 revealed that test takers evaluated the read-aloud task to be generally easy but found the independent and integrated tasks to be difficult, a finding that generally confirms the intended difficulty of the tasks (Evanini et al., 2015). Such explicit distinctions of differing levels of perceived difficulty may stem from the relatively less-demanding nature of what is being required of the read-aloud task (e.g., read out loud a given reading passage) as opposed to the independent or integrated task types (e.g., spontaneously speaking and expressing opinions), as discussed by Elder et al. (2002). Naturally, read-aloud task requirements may have also afforded the novice, and potentially, less-proficient test takers than those recruited in the current study, to feel more at ease, or at least sufficiently confident that they can successfully take on the task — hence, leading to the relatively favorable self-evaluations expressed in survey Question 3. This finding, therefore, may speak to the kinds of tasks that are potentially suitable, and in need of further consideration, for accommodating the target ITP population — that is, tasks associated with visual references (e.g., text- or picture-based description tasks) or other kinds of controlled task types (e.g., listen-and-repeat, otherwise known as elicited imitation task).

Second, as indicated from responses to Question 4, test takers also associated the independent and integrated task types — tasks that directly tap into communicative skills — with relatively high levels of authenticity. This finding lends support to the intended claims of these task types that they simulate the real-life, academic language-use situations (Butler et al., 2000; Cumming et al., 2004). The read-aloud task, on the other hand, garnered mixed reactions, which seems to align with the limited usage of the very task type in the real-world setting. Although reading aloud is a common activity in language classroom contexts (McKay, 2006), it is seldom used in broader language-use contexts, particularly for communicative purposes (Brown & Abeywickrama, 2010). Primarily advanced in their academic studies, test takers in this study also could have been more inclined to perceive the specific task type to be relatively irrelevant.

Third, responses from Question 5 demonstrated that test takers concurrently held favorable opinions toward all task types, indicating their usefulness in evaluating English speaking ability. In particular, this trend in responses was most salient for the independent and integrated task types, which further underscores a high level of acceptability of those very task types, as evidenced from previous studies (Huang & Hung, 2017; Malone & Montee, 2014). Finally, the positive written comments also confirmed that the test tasks offer learning opportunities and encourage test takers to reflect upon what they need to further improve, thereby inducing positive washback for learning (Barkaoui et al., 2013).

Yet another consistent finding is that test takers' perceptions were unfavorable when asked to evaluate the extent to which performance on the speaking tasks reflected their true ability. With specific regard to responses from Question 2 on the survey, test takers largely perceived that their performances were not the best representations of their actual speaking ability. Further, test takers tended to consider their performance to be average or below average, as evidenced in responses from Question 3. Collectively, this pattern of perceptions was most markedly evident for the independent and integrated tasks, both of which were associated with higher degrees of perceived difficulty (see Figure 1).

With some limitations, qualitative comments can be leveraged to help explain the unfavorable reactions, while pointing to potential areas for monitoring and further research. In relation to the independent and integrated tasks, test takers in this study raised a wide range of issues with relevant task-performance conditions as well as other content-related concerns. Specifically, the role that the amount of time allotted to planning and responding plays in task performance, which happened to constitute one of the most salient issues for test takers, should be further understood in consideration of the precise characteristics of the target TOEFL ITP population. Although extant research generally attests to the validity of the operational length of time offered in TOEFL speaking tasks (e.g., Inoue & Lam, 2021), little is known as to the extent to which specific groups of test takers, presumably those with lower language proficiency, are indeed affected by differing timed conditions (see also Ellis (2009) for a relevant discussion). With respect to planning times, O'Grady (2019) provided preliminary insights into how low-ability speakers take advantage of the extended planning time to compensate for limitations in their language knowledge to complete test tasks. For instance, O'Grady reported that test takers' performance on a narrative speaking task with 5 minutes of planning time was found to be significantly higher relative to test performance resulting from a 30-second planning condition. On the other hand, the same effect of extended planning was not found for advanced-level test takers, who presumably have access to a wider range of language knowledge to perform the task regardless of the extra time given for preparation. Given these findings, the even shorter planning time offered in the TOEFL ITP (20–30 seconds) could have led the test takers in the current study to become more susceptible to forming a critical or negative stance toward the testing situation, and presumably, of their own performance (Iwashita et al., 2001).

On this ground, O'Grady (2019) further urged that speaking tests developed specifically for accommodating less-proficient learners "involve a period of planning to meet Swain's (1985, p. 42) requirement that tests should "bias for the best" performance" (p. 522). This claim aligns with Wigglesworth and Elder's (2010) argument for the provision of preparation time to promote test fairness; that is, preparation time is needed to reduce test anxiety and allow test takers to give their best performance. Follow-up research that involves fine-grained timed conditions and controls for test-taker proficiency levels is thus needed to further verify the concerns raised in the study; yet, as Inoue and Lam (2021) remind us, such research needs to strike a balance between what is feasible within the operational parameters and adherence to the specific demands of test takers, which poses a specific challenge for test developers and researchers alike.

It is worth noting that test-taker feedback on the read-aloud task revealed how task instructions could be misunderstood, to the extent that test takers believed it affected their performance. In their comments, some test takers noted that there were differences in the aural (narrations) and visual (textual) instructions provided for the task. Notably, similar sentiments were not reported during an earlier phase of development (i.e., a prototype study), which further underscores the importance of conducting a larger-scale field test, one that preferably involves an adequately motivated and larger sample of the target testing population so a variety of opinions and voices can be heard (Reed, 2014). Most importantly, this finding points to the need to change the directions and to investigate the extent to which responses to the very task on the field test were impacted by the directions. Follow-up research could therefore benefit from trialing a modified version of the task lead-in on the extent to which prospective test takers engage with and perform on the task as intended.

These results, although informative, should be interpreted with some caution. Test takers' perceptions might differ from their actual behavior (performance); thus, some issues discussed above may need to be further investigated before making definitive conclusions. In addition, some of the issues that test takers noted (or their negative perceptions of the tasks in general) could be due to the lack of exposure to and familiarity with the testing environment in general. Indeed, over half of all candidates ($N = 350$; 57% of the sample) included in the study were first-time test takers of the TOEFL ITP; as exemplified in the written comments, some also reported that they had minimal experience in taking a computer-mediated speaking test. Such a sense of novelty, as Fox and Cheng (2015) described, may lead test takers to be more sensitive and, in a sense, critical toward the testing environment and the accompanying performance conditions. Yet in light of launching a new test section, as in the case of the TOEFL ITP Speaking test, the affective state of test takers, and their unfamiliarity with the testing situation in particular, should still be carefully considered and monitored (Jennings et al., 1999; Reed, 2014).

Conclusion

The present study was motivated by a desire to obtain test-taker perspectives to identify any substantial gaps in task designs or testing environment that might critically undermine the deployment of the new TOEFL ITP Speaking section. By means of a questionnaire, it gathered preliminary evidence in regard to supporting the intended claims of the proposed test tasks and revealed further areas for improvement.

Several limitations of the study deserve mention. Data used in the study were from a subset of test takers within the larger field tested sample, who may not have been representative of the full population of institutional test takers. In addition, statistical analyses on the survey data were mainly descriptive and performed on the full analysis set; the results we observed might have been different at an examinee subgroup level. The reason for not conducting such group-level comparisons, however, was that using a convenience sample in the field test, and the composition of a possible subgroup (e.g., in terms of language proficiency and academic background) might result in substantial differences within the group. Nevertheless, a more in-depth, qualitatively driven investigation, one that possibly controls for examinee background, might provide a richer understanding of how the employed task types function across differing demographics of the test-taker population. Similarly, richer test-taker feedback—both in quantity and quality—could have been generated if the open-ended question were mandatory to complete. Although this requirement was not imposed (in the interest of limited time) during the field testing, a follow-up study with a more sophisticated data elicitation instrument could be conducted to tap into a fuller picture of test takers' perceptions. Most importantly, test takers' perceptions were not considered in direct relation to their actual task performances. In fact, the written feedback in regard to specific task design features (e.g., time allotment) hypothesized effects on task performance. Employing such a study design could be more immediately useful in justifying the usefulness of test-taker accounts in validation research (Xie, 2011). Along similar lines, the reliability of self-report data should be taken into consideration when interpreting the study results. As mentioned above, further empirical research is needed to verify the claims put forth by the test takers in the current study.

These limitations notwithstanding, it is hoped that the current study findings could constitute an initial step toward accumulating validity evidence in support of the new speaking section of the TOEFL ITP. The findings provide insights into making candidates' accounts essential in the test validation process and, ultimately, leveraging their experiences for forming an evaluative judgment to claim that the proposed speaking tasks are functioning as expected and thus can be used for their intended purpose.

References

Alderson, C. J., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press.

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

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing*, *17*(1), 1–42. https://doi.org/10.1177/026553220001700101.

Bachman, L. F., & Purpura, J. E. (2008). Language assessments: Gate-keepers or door-openers? In B. Spolsky & F. Hult (Eds.), *Handbook of educational linguistics* (pp. 456–468). Blackwell. https://doi.org/10.1002/9780470694138.ch32

Barkaoui, K., Brooks, L., Swain, M., & Lapkin, S. (2013). Test-takers' strategic behaviors in independent and integrated speaking tasks. *Applied Linguistics*, *34*(3), 304–324. https://doi.org/10.1093/applin/ams046

Brown, A. (1993). The role of test-taker feedback in the test development process: Test-takers' reactions to a tape-mediated test of proficiency in spoken Japanese. *Language Testing 10*(3), 277–303. https://doi.org/10.1177/026553229301000305

Brown, H. D., & Abeywickrama, P. (2010). *Language assessment: Principles and classroom practices* (2nd ed.). Pearson Education.

Butler, F. A., Eignor, D., Jones, S., McNamara, T., & Suomi, B. K. (2000). *TOEFL 2000 speaking framework: A working paper (TOEFL Monograph No. 20)*. ETS.

Chapelle, C. A. (2020). *Argument-based validation in testing and assessment*. Sage. https://doi.org/10.4135/9781071878811

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). *Building a validity argument for the test of English as a foreign language*. Routledge.

Cohen, A., & Upton, T. (2007). "I want to go back to the text": Response strategies on the reading subset of the new TOEFL®. *Language Testing*, *24*(2), 209–250. https://doi.org/10.1177/0265532207076364

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

Collins, J. B., & Miller, N. H. (2018). The TOEFL (ITP): A survey of teacher perceptions. *Shiken: JALT Testing and Evaluation SIG Newsletter*, *22*(2), 1–13.

Crooks, T. J., Kane, M. T., & Cohen, A. S. (1996). Threats to the valid use of assessments. *Assessment in Education: Principles, Policy and Practice*, *3*(3), 265–286. https://doi.org/10.1080/0969594960030302

Cumming, A., Grant, L., Mulcahy-Ernt, P., & Powers, D. E. (2004). A teacher verification study of speaking and writing prototype tasks for a new TOEFL. *Language Testing*, *21*(2), 107–145. https://doi.org/10.1191/0265532204lt278oa

Davies, A., Brown, A, Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing*. Cambridge University Press.

Eccles, J. S., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J., & Midgley, C. (1983). Expectancies, values, and academic behaviors. In J. I. Spence (Ed.), *Achievement and achievement motives* (pp. 75–146). W. H. Freeman.

Elder, C., Iwashita, N., & McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: What does the test-taker have to offer? *Language Testing*, *19*(4), 347–368. https://doi.org/10.1191/0265532202lt235oa

Ellis, R. (2009). The differential effects of three types of task planning on the fluency, complexity, and accuracy in L2 oral production. *Applied Linguistics*, *30*(4), 474–509. https://doi.org/10.1093/applin/amp042

Evanini, K., Heilman, M., Wang, X., & Blanchard, D. (2015). *Automated scoring for the TOEFL Junior® Comprehensive Writing and Speaking test*. (Research Report No. RR-15-09). ETS. https://doi.org/10.1002/ets2.12052

Fox, J., & Cheng, L. (2015). Walk a mile in my shoes: Stakeholder accounts of testing experience with a computer-administered test. *TESL Canada Journal*, *32*(9), 65–85. https://doi.org/10.18806/tesl.v32i0.1218

Friedman, D. A. (2012). How to collect and analyze qualitative data. In A. Mackey & S. M. Gass (Eds.), *Research methods in second language acquisition: A practical guide* (pp. 180–200). Wiley Blackwell.

Fulcher, G., & Davidson, F. (2012). The *Routledge handbook of language testing*. Routledge. https://doi.org/10.4324/9780203181287

Golubovich, J., Tolentino, F., & Papageorgiou, S. (2018). *Examining the applications and opinions of the TOEFL ITP® assessment series test scores in three countries* (Research Report No. RR-18-44). ETS. https://doi.org/10.1002/ets2.12231

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *23*(1), 17–27. https://doi.org/10.1111/j.1745-3992.2004.tb00149.x

Huang, H-T. D., & Hung, S-T. A. (2017). EFL test-takers' feedback on integrated speaking assessment. *TESOL Quarterly*, *51*(1), 166 – 179. https://doi.org/10.1002/tesq.330

Huhta, A., Kalaja, P., & Pitkänen-Huhta, A. (2006). Discursive construction of a high-stakes test: The many faces of a test-taker. *Language Testing*, *23*(3), 326 – 350. https://doi.org/10.1191/0265532206lt331oa

Inoue, C., & Lam, D. M. K. (2021). *The effects of extended planning time on candidates' performance, processes, and strategy use in the lecture listening-into-speaking tasks of the TOEFL iBT® test* (TOEFL Research Report No. RR-93). ETS. https://doi.org/10.1002/ets2.12322

Iwashita, N., McNamara, T., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to task design. *Language Learning*, *51*(3), 401 – 436. https://doi.org/10.1111/0023-8333.00160

Jennings, M., Fox, J., Graves, B., & Shohamy, E. (1999). The test-takers' choice: An investigation of the effect of topic on language test performance. *Language Testing*, *16*(4), 426 – 456. https://doi.org/10.1177/026553229901600402

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (pp. 17 – 64). Greenwood Publishing.

Kenyon, D. M., & MacGregor, D. (2012). Pre-operational testing. In G. Fulcher & F. Davidson (Eds.) *The Routledge handbook of language testing* (pp. 295 – 306). Routledge.

Kenyon, D. M., & Stansfield, C. (1991, April 4 – 6). *A method for improving tasks on performance assessments through field testing* [Paper presentation]. *Annual meeting of the National Council on measurement in education*, Chicago, IL, United States.

Malone, M. E., & Montee, M. (2014). *Stakeholders' beliefs about the TOEFL iBT® test as a measure of academic language ability* (TOEFL iBT Report No. 22, ETS Research Report No. RR-14-42). ETS. https://doi.org/10.1002/ets2.12039

McKay, P. (2006). *Assessing young language learners*. Cambridge University Press. https://doi.org/10.1017/CBO9780511733093

Messick, S. (1998). Validity: A matter of consequences. *Social Indicators Research*, *45*(1 – 3), 35 – 44. https://doi.org/10.1023/A:1006964925094

Moss, P. A. (1996). Enlarging the dialogue in educational measurement: Voices from interpretive research traditions. *Educational Researcher*, *25*(1), 20 – 29. https://doi.org/10.3102/0013189X025001020

Moss, P. A., Girard, B. J., & Haniford, L. C. (2006). Chapter 4: Validity in educational assessment. *Review of Research in Education*, *30*, 109 – 162. https://doi.org/10.3102/0091732X030001109

Ockey, G. J., Koyama, D., & Setoguchi, E. (2013). Stakeholder input and test design: A case study on changing the interlocutor familiarity facet of the group oral discussion test. *Language Assessment Quarterly*, *10*(3), 292 – 308. https://doi.org/10.1080/15434303.2013.769547

O'Grady, S. (2019). The impact of pre-task planning on speaking test performance for English-medium university admission. *Language Testing*, *36*(4), 505 – 526. https://doi.org/10.1177/0265532219826604

Papageorgiou, S., Wu, S., Hsieh, C.-N., Tannenbaum, R. J., & Cheng, M. (2019). *Mapping the TOEFL iBT® test scores to China's standards of English language ability: Implications for score interpretation and use*. (Research Report No. TOEFL-RR-89) ETS. https://doi.org/10.1002/ets2.12281

Read, J. (1990). Providing relevant content in an EAP writing test. *English for Specific Purposes*, *9*(2), 109 – 121. https://doi.org/10.1016/0889-4906(90)90002-T

Reed, D. J. (2014). Field testing of test items and tasks. In A. J. Kunnan (ed.), *The companion to language assessment* (pp. 1 – 17). Wiley Blackwell.

Sasaki, M. (2000). Effects of cultural schemata on students' test-taking processes for cloze tests: A multiple data source approach. *Language Testing*, *17*(1), 85 – 114. https://doi.org/10.1177/026553220001700104

Sato, T., & Ikeda, N. (2015). Test-taker perception of what test items measure: A potential impact of face validity on student learning. *Language testing in Asia*, *5*(10), 1 – 16. https://doi.org/10.1186/s40468-015-0019-z

Scouller, K., & Prosser, M. (1994). Students' experiences in studying for multiple choice question examinations. *Studies in Higher Education*, *19*(3), 267 – 279. https://doi.org/10.1080/03075079412331381870

Shohamy, E. (2001). Democratic assessment as an alternative. *Language Testing*, *18*(4), 373 – 391. https://doi.org/10.1177/026553220101800404

Stansfield, C. W., Kenyon, D. M., Paiva, R., Doyle, F., Ulsh, I., & Cowles, M. A. (1990). The development and validation of the Portuguese speaking test. *Hispania 73*(3), 641 – 651. https://doi.org/10.2307/343942

Strauss, A., & Corbin, J. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (2nd ed.). Sage.

Stricker, L. J. (2012). Testing: It's not just psychometrics. (Research Memorandum No. RM-12-07). ETS.

Stricker, L. J., & Ward, W. C. (2004). Stereotype threat, inquiring about test takers' ethnicity and gender, and standardized test performance. *Journal of Applied Social Psychology*, *34*(4), 665 – 693. https://doi.org/10.1111/j.1559-1816.2004.tb02564.x

Swain, M. (1985). Large scale communicative testing: A case study. In Y. Lee, C. Fok, R. Lord & G. Low (Eds.), *New directions in language testing* (pp. 35 – 46). Pergamon Press.

Weigle, S. (2004). Integrating reading and writing in a competency test for non-native speakers of English. *Assessing Writing*, *9*(1), 28 – 47. https://doi.org/10.1016/j.asw.2004.01.002

Wigglesworth, G., & Elder, C. (2010). An investigation of the effectiveness and validity of planning time in speaking test tasks. *Language Assessment Quarterly*, *7*(1), 1–24. https://doi.org/10.1080/15434300903031779

Xie, Q. (2011). Is test taker perception of assessment related to construct validity? *International Journal of Testing*, *11*(4), 324–348. https://doi.org/10.1080/15305058.2011.589018

Suggested citation:

Lee, S. (2022). *Evaluating the new* TOEFL ITP® *Speaking test: Insights from field test takers* (TOEFL Research Report No. RR-99). ETS. https://doi.org/10.1002/ets2.12352

Action Editor: Jonathan Schmidgall

Reviewers: Spiros Papageorgiou and Lin Gu

Find other ETS-published reports by searching the ETS ReSEARCHER database at http://search.ets.org/researcher/