Research Report

Interpretation and Use of a Workplace English Language Proficiency Test Score Report: Perspectives of TOEIC® Test Takers and Score Users in Taiwan

ETS RR–23-10

Ching-Ni Hsieh

December 2023

ETS

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Interpretation and Use of a Workplace English Language Proficiency Test Score Report: Perspectives of TOEIC® Test Takers and Score Users in Taiwan

Ching-Ni Hsieh

ETS, Princeton, NJ

Research in validity suggests that stakeholders' interpretation and use of test results should be an aspect of validity. Claims about the meaningfulness of test score interpretations and consequences of test use should be backed by evidence that stakeholders understand the definition of the construct assessed and the score report information. The current study explored stakeholders' uses and interpretations of the score report of a workplace English language proficiency test, the TOEIC® Listening and Reading (TOEIC L&R) test. Online surveys were administered to TOEIC L&R test takers and institutional and corporate score users in Taiwan to collect data about their uses and interpretations of the test score report. Eleven survey respondents participated in follow-up interviews to further elaborate on their uses of the different score reporting information within the stakeholders' respective contexts. Results indicated that the participants used the TOEIC L&R test scores largely as intended by the test developer although some elements of the score report appeared to be less useful and could be confusing for stakeholders. Findings from this study highlight the importance of providing score reporting information with clarity and ease to enhance appropriate use and interpretation.

Keywords workplace English; score reporting; stakeholder perception; TOEIC®; Taiwan; English language proficiency; validity; score use; score interpretation

doi:10.1002/ets2.12373

Validity as a concept has evolved from the validation of test scores themselves toward the evaluation of score interpretations and uses in recent years (O'Leary et al., 2017). Researchers have argued that how test users interpret test results (MacIver et al., 2014) and the interpretability of score reports (Van der Kleij et al., 2014) should be considered as aspects of validity. MacIver et al. (2014) proposed the concept of *user validity* and define it as "the overall accuracy and effectiveness of interpretation resulting from the test output"; the concept focuses on "the validity of the interpretations in use and the decisions that form part of these interpretations" (p. 155). This perspective of validity provides a useful framework for examining actual test uses and interpretations of score report information — the focus of the current study.

Within the context of language assessment, researchers have suggested that claims about the meaningfulness of test score interpretation and consequences of test use should be backed by evidence that stakeholders understand the definition of the construct assessed and the score report information (Bachman & Palmer, 2010). To provide evidence to support valid use and interpretation of score reports, Tannenbaum (2019) proposed that evidence is needed to demonstrate that (a) users interpret features of score reports as intended, (b) subgroups of stakeholders understand the same reported information as intended, and (c) stakeholders can act on the reported information consistent with expectations, among others. Other researchers in score reporting research have called for ongoing evaluation of the consequences and use of score reports that involve various stakeholder groups and recommended considering separate stakeholder groups when investigating the use and interpretation of score report information (Van den Heuvel et al., 2014).

Quality score reports are critical to valid use of test results as they are the communication tools that connect test developers with the stakeholders. To date, much has been discussed and researched about the process of developing score reports (Zapata-Rivera et al., 2012; Zenisky & Hambleton, 2012) and the guidelines for designing quality score reports (Goodman & Hambleton, 2004; Hambleton & Zenisky, 2013). Despite the documented best practices and guidelines for developing quality score reports, research continues to find that test takers and score users have difficulties understanding

Corresponding author: C.-N. Hsieh, E-mail: chsieh@ets.org

score reporting information as intended across testing purposes and contexts (e.g., admissions, credentialing, hiring, formative assessment; see Kannan et al., 2018; Kim et al., 2019; Van der Kleij et al., 2014; Zapata-Rivera et al., 2021). Through evaluating comprehension and aspects of score reporting information, actual uses of scores, and the effectiveness of ancillary materials, several empirical studies collectively show that the uses and interpretations of test scores differ among stakeholder groups and are influenced by the backgrounds of the users (e.g., educational level, language proficiency, country of origin, institution type, profession; see Golubovich et al., 2018; Kannan et al., 2018; Kim et al., 2019; Van der Kleij et al., 2014; Zapata-Rivera et al., 2021).

The use and interpretation of test performance feedback provided in score reports is a topic that has been under-researched in the existing literature on score reporting (Kim et al., 2019). Test performance feedback provides stakeholders with more elaborate information about the strengths and weaknesses of test takers' knowledge, skills, or abilities, in addition to the single numeric scores. One common type of feedback is performance level descriptors (PLDs). PLDs are statements that describe specific knowledge and skills test takers typically demonstrate at a given performance level (Cizek et al., 2005) and can facilitate teachers' diagnostic judgments about students' skill mastery status (Jang, 2013). PLDs are widely reported in standards-based English language proficiency assessments in the United States (Baron et al., 2020) and in large-scale language assessments such as the TOEFL iBT® test (Gomez et al., 2007), the TOEFL ITP® tests (Powers et al., 2017), the TOEIC® tests (Liao, 2010), and the Eiken Test in Practical English Proficiency in Japan (Sawaki & Koizumi, 2017). Though commonly reported, one drawback of the PLDs is that the descriptors are identical for all test takers at a particular performance level. Thus, they have been criticized for not providing useful, personalized performance profiles and only offering limited information for learning and instruction (Sawaki & Koizumi, 2017). Other typical types of test performance feedback include test section scores (e.g., listening, reading, grammar, vocabulary scores; see Papageorgiou & Choi, 2018), subscores for groups of items within a test section using score augmentation method (Haberman & Sinharay, 2010), and cognitive diagnostic feedback based on subskill mastery (Kim, 2015; Kunnan & Jang, 2009). Given that test performance feedback presented in score reports are typically intended to help test takers identify skill improvement areas and provide educators with more detailed information about students' performance and inform remedial instruction, appropriate interpretation of the feedback by stakeholders is thus a necessary precondition for adequate use and planning of appropriate actions based on test results. Consequently, the extent to which stakeholders interpret and use test performance feedback as intended warrants further research investigation.

Despite the potential pedagogical value of test performance feedback, research in score reporting has shown that users do not necessarily know how to use or interpret feedback information as intended (Clark et al., 2022; Kim et al., 2019; Min et al., 2022). The literature suggests that multiple reasons could contribute to problems with understanding performance feedback, ranging from the assessment literacy of the users (Kim et al., 2019) to the design of the score report (Zapata-Rivera et al., 2021) and the complexity of the assessment and/or reporting information (Van den Heuvel et al., 2014). Carless and Boud (2018) further identified several barriers to students' ineffective use of feedback. One major challenge was that students did not recognize the value of feedback. The researchers argued that it was important to provide training to help students understand the value of feedback and develop capabilities to make sense of the feedback information they receive and act on it. Taken together, the existing body of research suggests that stakeholders need to have the assessment literacy or the knowledge and skills to use assessment results appropriately.

As the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 2014) indicates, it is critical to recognize that communicating score reporting information in a clear, accessible, and appropriate manner is an important responsibility of the test developers to support stakeholders' interpretation and use of test results. In order to provide quality score reports that meet the needs and backgrounds of stakeholders, researchers have advised keeping the audience of the score report in mind (Zapata-Rivera & Katz, 2014), advocated for soliciting stakeholder input at multiple points in score report development processes to identify potential interpretation and use issues (Van den Heuvel et al., 2014), and emphasized the need for conducting ongoing evaluation of test impact and use (Hambleton & Zenisky, 2013). The current study focused on the evaluation of stakeholders' uses and interpretations of an English language proficiency test score report to collecte validity evidence and support claims made of the test results. The following section details the study context and the test.

Study Context

The current study investigated stakeholders' uses and interpretations of a score report of the TOEIC Listening and Reading test (TOEIC L&R test) and focused on collecting validity evidence to support claims about the meaningfulness of the test scores, the appropriateness of score-based decisions, and the consequences of test use. The TOEIC L&R test is a standardized language proficiency test that measures everyday English listening and reading skills needed to work in an international environment. Many companies, academic institutions, and language programs worldwide use the TOEIC L&R test to measure English language proficiency of students and business professionals. The full-length version of the TOEIC L&R test contains a listening and a reading section. Currently, the test includes 100 multiple-choice listening questions and 100 multiple-choice reading questions. The test questions simulate real-life situations that are relevant to the global workplace. The listening section takes approximately 45 minutes, and the reading section takes about 75 minutes.

Scores on the TOEIC L&R test are determined by the number of correct answers. The raw listening and reading section scores are then converted to scaled scores (for a detailed description of the scoring process and its statistical specifications, see Cid et al., 2018). The scaled section scores range from 5 to 495, for a total scaled score of 10 to 990 with 5-point increments. The TOEIC L&R test allows test takers to (a) measure their English proficiency, (b) qualify for a new position and/or promotion in a company, (c) enhance professional credentials, (d) monitor progress in English, (e) set learning goals, and/or (f) involve their employers or academic institutions in advancing their English abilities. The test section and total scaled scores can be used to evaluate an individual's ability to communicate in English; apply for jobs; hire, train, and promote employees; and better prepare students for employment opportunities. See the TOEIC Examinee Handbook (ETS, 2022a) for details on the intended uses of the test.

The TOEIC L&R score report provides the following information: (a) test-taker background information; (b) test date, and the date after which use of the scores is not recommended; (c) listening, reading, and total scaled scores; (d) listening and reading score descriptors; (e) listening and reading abilities measured; (f) percent correct of abilities measured; and (g) a footnote on how to read the abilities-measured information (see Appendix A for an example TOEIC L&R score report). In this study, the score descriptors and the percent correct of abilities were referred as test performance feedback provided in the TOEIC L&R score report.

The TOEIC listening and reading test sections were developed using evidence-centered design (ECD; Schedl, 2010), an approach to test design that requires specifying claims about test-takers' abilities. Following ECD, all test items in the listening and reading test sections are designed to elicit evidence of test takers' abilities with respect to one or more of the claims about listening or reading abilities. For the TOEIC listening test section, the test tasks are designed to elicit test-taker's abilities regarding the following claims: The test taker can understand (a) gist in short spoken text, (b) gist in extended spoken text, (c) details in short spoken text, (d) details in extended spoken text, and (e) purpose of implied meaning/pragmatic understanding. For the TOEIC reading test section, the test tasks are designed to elicit test taker's abilities regarding the following claims: The test taker can (a) make inferences based on information that is explicitly stated in texts, (b) understand specific (factual) information in tables and passages, (c) connect information across multiple sentences in single texts and across two texts, (d) understand vocabulary, and (e) understand grammar. Thus, the "abilities measured" are directly tied to the construct definition (i.e., they reflect how the construct is defined through the claims test developers think are important to make about listening or reading comprehension). They are operationalized in the TOEIC L&R test through test items that are designed to elicit evidence with respect to specific claim(s). The percent correct of abilities-measured information presented in the TOEIC L&R score report indicate the percentage of questions associated with each claim about the ability answered correctly. This information provides more insight into test-taker performance with respect to how it has been defined and operationalized in the TOEIC L&R test.

The footnote in the score report provides a brief description of the abilities measured and states that test takers should not compare the percentages across those obtained from different test forms or administrations. This is because the abilities-measured data are not equated across test forms, unlike the TOEIC L&R section and total scaled scores. Consequently, a repeat test taker may expect more variations in performances across test forms and occasions for the abilities-measured data compared to the scaled scores. Thus, stakeholders need to exercise caution if they desire to use the abilities-measured data to monitor one's progress in English because the changes in the abilities-measured information can result from changes in test taker's *actual* abilities and minor differences between test forms. The footnote also implies that test takers may compare their percent correct of abilities-measured information to that of test takers who have taken

the same TOEIC L&R test form; although, due to the relatively low number of questions in some abilities, users should be prudent when comparing the percentages across test takers who have taken even the same form. With these cautions in mind, stakeholders of the TOEIC L&R test can use the abilities-measured information to identify test takers' strengths and weaknesses for the purpose of informing future study. The information, however, is not intended to be used for making any high-stakes decisions, such as admissions, hiring, or promotion.

It should be noted that the current study was part of a larger study that examined the use and interpretation of the TOEIC L&R test scores and the abilities-measured information. The study was also a follow-up study of Hsieh (2023). In that study, only perspectives of test takers in Japan were included, which limited the generalizability of the study findings to other TOEIC L&R test stakeholder groups and test-taking population. Additionally, in Japan, the different elements of the TOEIC L&R score report, including the listening and reading score descriptors, the abilities-measured information, and the footnote message, are translated from the original English wording and presented in Japanese, whereas in many other countries where the test is administered the original English texts are used to present all the score reporting information. It is unclear whether this difference in score reporting language has an impact on stakeholders' understanding and uses of the score report; further investigation is warranted.

The audience of score reports typically consists of both test takers and score users such as educators, decision makers, or policymakers. When evaluating the use and interpretation of score reports, research has proposed that it is necessary to separate out stakeholder groups, consider the needs of disparate groups separately, and explicitly seek input from different groups (Van den Heuvel et al., 2014). Within the context of the TOEIC L&R test, test takers as well as corporate and institutional score users are key stakeholders whose opinions and perceptions play an important role in the meaningfulness of score interpretations (e.g., understanding the meaning of scores and construct assessed) and appropriate test uses (e.g., using scores as intended, understanding score reports). This study solicited the perspectives of test takers and score users in Taiwan, exploring the uses and interpretations of the TOEIC L&R score report information in educational institutions where students are developing English skills for everyday and workplace purposes and in domestic and multinational corporations for various purposes. The study results could contribute to test validation research and have practical implications for the design and development of score reports. The research questions (RQs) guiding the study included the following: .

1. How do TOEIC L&R test takers and institutional and corporate score users in Taiwan use the test scores?
2. How do TOEIC L&R test takers and institutional and corporate score users in Taiwan use the abilities-measured information presented in the score report?
3. How well do TOEIC L&R test takers and institutional and corporate score users in Taiwan understand the abilities-measured information presented in the score report?
4. How do the TOEIC L&R test takers and institutional and corporate score users in Taiwan perceive the usefulness of the score report?

Methods

Participants

The research participants included three TOEIC L&R test stakeholder groups in Taiwan: test takers, score users in educational institutions, and score users in domestic and multinational corporations. The participants responded to an online survey and 11 survey respondents also participated in follow-up interviews (see Table 1). The survey respondents and the interviewees voluntarily participated in the study and did not receive any monetary compensation for their participation.

Survey Respondents

Test takers. A total of 1,444 TOEIC L&R test takers who took the test in Taiwan between January 2021 and January 2022 participated in the online survey. The participants' ages ranged from 18 to 66 ($M = 25.7$, $SD = 7.9$). Two-thirds (62.8%) of the participants were females and the remaining (36.6%) were males; eight participants preferred not to report their gender. The participants varied in the number of times they had taken the TOEIC L&R test. Around half of the participants (51.9%) had taken the test one or two times at the time of data collection; 71.2% of the respondents optionally reported their most recent TOEIC L&R test total scores ($M = 710.6$, $SD = 133.6$). This reported mean score

Table 1 Background Information of Interview Participants by Stakeholder Group

ID	Stakeholder group	Age	Gender	Profession	Test-taking experience
1	Test taker	19	Female	College student	3 times
2	Test taker	22	Male	College student	>4 times
3	Test taker	19	Female	College student	1 time
4	Test taker	30	Female	Airline company employee	>4 times
5	Test taker	26	Male	Postgraduate student	>4 times
6	Test taker	19	Female	College student	3 times
7	Institutional score user	55	Female	College English instructor	Yes
8	Institutional score user	37	Female	College language center administrator	Yes
9	Institutional score user	38	Male	Senior high school English teacher	Yes
10	Corporate score user	38	Female	Manufacturing company HR manager	Yes
11	Corporate score user	38	Female	Foreign bank HR manager	Yes

Note. The score users were asked whether they had taken the TOEIC L&R test, but not asked the number of times they had taken the test.

was 145 scaled score point higher than the mean score of test takers from Taiwan who took the test in year 2021 ($M = 565$, $SD = 204$; see ETS, 2022b).

Institutional score users. Fifty-nine institutional score users participated in the online survey. The majority was female ($n = 51$, 86.4%). More than half of the participants (54.2%) were faculty members, administrators, or staff at higher-education institutions and the remaining (45.8%) taught or worked at regular or vocational high schools at the time of data collection. Thirty-five of the participants (59.3%) had taken the TOEIC L&R test and the remaining had never taken the test. Fourteen optionally reported their most recent TOEIC L&R test total scores ($M = 883.2$, $SD = 107.2$).

Corporate score users. Fifty-four corporate score users participated in the online survey (32 females, 22 males). They were working in various industries at the time of data collection. Many were employed either in the manufacturing (46.3%) or information technology (27.8%) sector. Around two-thirds of the corporate score users (63%) were human resources (HR) personnel; the remaining played different roles such as managers, executives, sales, customer services in their respective corporations. Most of the corporate score users ($N = 47$, 87%) had taken the TOEIC L&R test. Sixteen optionally reported their most recent TOEIC L&R test scores ($M = 639.7$, $SD = 205.4$).

Interview Participants

Test takers. Out of the 1,444 survey respondents, 217 expressed an interest in participating in a follow-up interview. Twenty-two of those (approximately 10%) were randomly selected to have representation of gender, age, profession, and experience taking the TOEIC L&R test and invited to participate. Due to attrition and challenges between scheduling and conducting the interviews, six agreed and participated. They were four females and two males. Four were undergraduate students; one, graduate student; and one, an airline company employee.

Institutional score users. Out of the 59 survey respondents, six expressed an interest in a follow-up interview. All six users were contacted, and three agreed to participate (two females, one male). They included one English instructor at a private four-year university, one language center administrator at a top-tier public university, and one English teacher at a private senior high school.

Corporate score users. Out of the 54 survey respondents, four expressed an interest in a follow-up interview. All four were contacted, and two agreed to participate. Both were female and senior HR managers. One worked in a manufacturing company headquartered in the United States; the other worked in a foreign bank's Taipei branch office.

Instruments

T his section describes the research instruments used in the study, including the online surveys and the interview questions.

Online Surveys

Three online surveys were created for data collection: (a) test-taker survey, (b) institutional score user survey, and (c) corporate score user survey. Each of the surveys had three parts. Part 1 contained questions about the participants' background information, such as age, gender, TOEIC L&R test-taking experience, and the uses of TOEIC L&R test section scores, scaled scores, and the abilities-measured information. An image of the abilities-measured information was embedded in this part of the survey to help the participants answer the question (see Appendix B, Question 6). The background questions in the three surveys varied when appropriate to reflect the different professions or status of the three stakeholder groups (e.g., student vs. corporate HR manager). Part 2 contained five comprehension questions (CQs) that assessed stakeholders' understanding of the abilities-measured information presented in the TOEIC L&R score report. The five questions were identical in all three surveys. Part 3 consisted of one set of Likert-scale questions that were used to gather stakeholders' perceptions of the different TOEIC L&R score report elements. These questions were identical in all three surveys. At the end of each survey, the respondents were asked to indicate their interest in participating in a follow-up interview. Those who expressed interests provided their contact information.

The survey questions were reviewed by ETS's editorial reviewers, assessment developers, psychometricians, and research scientists who had expertise in score reporting research. After the review process, the English version of the survey was translated into Chinese by the researcher; a site coordinator in Taiwan reviewed the Chinese versions of the surveys to ensure the accuracy of the translation. The Chinese versions of the surveys were piloted with two TOEIC L&R test takers in Taiwan. They were asked to respond to the questions and provide feedback on problematic or confusing questions, statements, and survey design. The pilot resulted in a few rounds of revisions on the wordings and answer options with the intention of maximizing the interpretability for the study participants. The survey was finalized after all reviewers reported that all the questions were clear. The final Chinese versions of the surveys were uploaded to an online survey platform, Alchemer, for data collection. Links to the three online surveys were sent to the site coordinator for dissemination (see Data Collection section).

Interview Questions

Three sets of semi-structured interview questions were used to guide the interview process. The interview questions asked the participants to (a) describe their experience taking or using the TOEIC L&R tests and their educational and/or employment backgrounds, (b) elaborate on how they use the TOEIC L&R test score reporting information for various purposes, (c) comment on how they use the abilities-measured information and the types of decisions they make based on the information, and (d) suggestions for improvements of the TOEIC L&R score report. Additional probing questions were used whenever relevant and appropriate.

Data Collection

To address the research questions, I collaborated with ETS's local partner in Taiwan to collect data. A local site coordinator from the local partner helped to (a) review the research instruments, (b) recruit survey participants, and (c) distribute the online surveys to the recruited participants.

The recruitment of test takers took place in Spring 2022. The site coordinator sent out an email recruitment letter to the TOEIC L&R test takers who took the test between January 2021 and January 2022 in Taiwan and invited them to participate in the study. This testing window was chosen to ensure that the recruited participants' most recent TOEIC L&R test scores were still valid at the time of data collection. Those who agreed to participate were provided the link to the online survey.

The recruitment of the score users was conducted by the local partner's sales team that had direct contacts with institutional and corporate score users in Taiwan. The team reached out to score users in different types of educational institutions, including public and private colleges and regular and vocational high schools, and corporations in various industries (e.g., financial services, manufacturing, aviation, banking, information technology) and invited the score users and/or decision-makers in these institutions and organizations to participate in the study. Those who agreed to participate were provided the links to the online surveys.

Upon completion of the survey data collection, I conducted the one-on-one follow-up interviews to further explore how the stakeholders used and interpreted the TOEIC L&R test results in their specific use contexts. The interviews

followed a semi-structured format that included the questions from the interview protocol and additional topics that emerged from the discussion. During each interview session, I screen-shared a sample TOEIC L&R test score report to allow the participants to view and reflect on their experiences using the test scores and the abilities-measured information. The interview was conducted in Mandarin Chinese, the participants' and the researcher's first language, via Zoom video conferencing. The interview sessions lasted between 30 and 40 minutes and were recorded for further analyses.

Data Analysis

Survey responses were downloaded and compiled from the Alchemer system in early summer 2022. Descriptive statistics and frequency counts were used to analyze the survey responses, using IBM SPSS Statistics (Version 27). To analyze the interview data, I first translated the interviews from Chinese to English. I then read and re-read the transcripts in their entirety several times to create an initial coding scheme as it related to the research questions. The coding scheme was used to guide a systematic analysis of the interview data. A second coder, who was a researcher at ETS and had a background in language education and assessment, discussed the initial coding scheme with me until the final coding categories were finalized. We then independently coded three interview transcripts using the final coding scheme to check coder agreement. Discrepant cases were discussed until 100% agreement was reached. After that, I coded the remaining interview data.

Results

Findings are summarized for each of the research questions that addressed how TOEIC L&R stakeholders in Taiwan used and interpreted the test results, their understanding of the abilities-measured information, and their perceptions of the usefulness of the score report.

RQ1. How do TOEIC L&R test takers and institutional and corporate score users in Taiwan use the test scores?

Tables 2 to 4 present the numbers and percentages of the survey respondents with respect to the ways they used the TOEIC L&R test scores. The participants could select multiple uses and thus the total percentages in the tables exceeded 100%.

For the test takers (see Table 2), the most frequent uses were "to evaluate English proficiency" (71.2%), "to meet school graduation requirement" (42.1%), "to improve English proficiency" (37.2%), and "to apply for jobs" (34.3%). Some test takers took the test "to get an internationally recognized English proficiency test certificate" (18.6%), "to meet school admissions requirement" (11.7%), "to pass job-related exams" (10.5%), and "to get promoted at work" (7.9%). The results indicated that test takers largely used the TOEIC L&R test scores as intended. Twenty-six test takers indicated "other" uses, which included to waive required English language courses in college, to apply for scholarship, to become a Chinese-English bilingual teacher, and to advance in career.

For the institutional score users (see Table 3), the most frequent uses were "to monitor students' progress in English-language classes or programs" (59.3%), "as one of the graduation requirements for undergraduate and/or degree programs" (54.2%), "to admit students to undergraduate degree programs, as contributing documentation for English proficiency"

Table 2 Test Takers' Uses of the TOEIC L&R Test Scores

Use of the TOEIC L& R section and total scores	*n*	%
To evaluate my English proficiency	1,028	71.2
To meet school graduation requirement	608	42.1
To improve my English proficiency	537	37.2
To apply for jobs	496	34.3
To get an internationally recognized English proficiency test certificate	269	18.6
To meet school admissions requirement	169	11.7
To pass job-related exams	151	10.5
To get promoted at work	114	7.9
Other	26	1.8

Note. $N = 1{,}444$.

Table 3 Institutional Score Users' Use of the TOEIC L&R Test Scores

Use of the TOEIC L&R section and total scores	*n*	%
To monitor students' progress in English-language classes or programs	35	59.3
As one of the graduation requirements for undergraduate and/or degree programs	32	54.2
To admit students to undergraduate degree programs, as contributing documentation for English proficiency	27	45.8
For scholarship programs, as contributing documentation for English proficiency	23	39.0
To place students into English-language classes or programs	19	32.2
To guide the selection and design of instructional materials or curriculum	10	16.9
To admit students to graduate degree programs, as contributing documentation for English proficiency	4	6.8
Other	0	0.0

Note. $N = 59$.

Table 4 Corporate Score Users' Use of the TOEIC L&R Test Scores

Use of TOEIC L&R section and total scores	*n*	%
To screen job applicants	38	67.9
To promote employees	32	57.1
To select employees for overseas work assignments	25	44.6
To assign employees to different roles	24	42.9
To assess employees' English proficiency	21	37.5
To place employees into different English language training programs	12	21.1
Other	3	5.4

Note. $N = 54$.

(45.8%), "for scholarship programs, as contributing documentation for English proficiency" (39.0%), and "to place students into English-language classes or programs" (32.2%). A few users reported that they used the test results to guide instruction (16.9%) or for graduate school admissions (6.8%).

For the corporate score users (see Table 4), the most frequent uses were "to screen job applicants" (67.9%), "to promote employees" (57.1%), "to select employees for overseas work assignments" (44.6%) and "to assign employees to different roles" (42.9%). Some users used the test results "to assess employees' English proficiency" (37.5%) and "to place employees into different English language training programs" (21.4%). Three users indicated "other" uses; one commented that her company used the test results to determine merit increases, and two mentioned that they didn't know how to use the test results.

Results of the interviews with the six test takers corroborated findings from the survey responses. When asked how they used the TOEIC L&R test scores, all the test takers said that their primary uses were to evaluate their English language proficiency, meet college graduation requirements, or apply for jobs in the future. Three test takers (IDs 1, 2, and 3) mentioned that they also took the test to apply for colleges and two (IDs 5 and 6) used the scores to waive required English language courses in college.

The institutional score users used the TOEIC L&R test results for various purposes and the primary use was to determine if students met the college graduation requirements. One user who was a language center administrator commented that her college used the test for multiple purposes. She said:

> In my college, we use TOEIC L&R test scores for placement purposes and college admissions. We also have a graduation requirement and the TOEIC L&R test is one of the tests the school accepts for meeting the requirement. The cut score for graduation is 550 on the TOEIC L&R test total score. (ID 8)

Another user who taught English to senior high school students (ID 9) reported that the TOEIC L&R test was not a required test for his students. He did not use the test results to guide instructions because not all of his students took the test, and he did not believe that the test content was aligned to his school's English curriculum. He stated that he encouraged his students to take the TOEIC L&R test because the test results could provide additional evidence of their English language proficiency when they applied for college.

The two corporate score users (IDs 10 and 11) mentioned that they used the TOEIC L&R test scores to evaluate and monitor progress of the English language proficiency of their employees who participated in short-term language training programs. One user (ID 10), an HR manager at a fitness equipment manufacturing company, mentioned that her company also used the TOEIC L&R total scores to make hiring decisions. The other user (ID 11), an HR manager at a foreign bank, said that her bank did not use the TOEIC L&R test scores to make any high-stakes decisions, such as hiring, promotion, or overseas job assignments. Instead, the bank used the scores to evaluate its employees' English language proficiency for language training purposes.

RQ2. How do TOEIC L&R test takers and institutional and corporate score users in Taiwan use the abilities-measured information presented in the score report?

Table 5 shows how the different stakeholder groups used the abilities-measured information. The participants could select multiple uses and thus the total percentages exceeded 100%. For test takers and corporate score users, the most common use of the information was to identify the strengths and weaknesses of test takers' language skills, although the percentages differed substantially across groups (67.5% and 37.0% respectively). Another frequent use of the abilities-measured information was to monitor skill improvement over time, and this was especially the case for institutional score users (57.6%). Small percentages of the stakeholders (between 11.1% and 16.7%) indicated that they did not pay much attention to the abilities-measured information or did not know how to use this information (between 11.6% and 27.8%). Eight test takers and two corporate score users provided additional comments. These comments generally showed that the stakeholders were either not aware of the abilities-measured information prior to responding to the surveys or did not understand the information.

During the interview sessions, the participants were presented with a sample TOEIC L&R test score report and asked how they used or interpreted the abilities-measured information. Four test takers (IDs 1, 4, 5, and 6) reported that they used the information mainly to identify their strengths and weaknesses and determine what skills they needed to improve, corroborating their survey responses. One test taker, who was a college student and had taken the test three times said:

> I would look at the percent correct information at the bottom of the score report to find my strengths and weaknesses. Based on these results, I would look for practice materials to work on my weak areas. (ID 6)

Two test takers (IDs 2 and 3) said that they initially did not pay much attention to the abilities-measured information, and that they did not always find the feedback helpful. Interestingly, both test takers changed their perceptions of the feedback after they paid some attention to the information and recognized the value of the feedback.

> I didn't pay much attention to the percent correct information in the first two times when I took the TOEIC L&R test, because at that time, I didn't think that this information was important or helpful. But when I took the test for the third time, I looked at the feedback more closely and I found it very helpful. (ID 2)

> Actually, I never paid attention to the percent correct information on the score report before I participated in this study. I noticed the information for the first time when I responded to the online survey. It was also at that time that I started to think about how to interpret the abilities-measured information and I also realized that I could actually use the information as feedback to improve my English skills. (ID 3)

Table 5 Use of the Abilities-Measured Information by Stakeholder Group

	Test takers		Institutional score users		Corporate score users	
Use of abilities-measured information	*n*	%	*n*	%	*n*	%
To identify strengths and weaknesses	975	67.5	30	50.8	20	37.0
To monitor progress over time	648	44.9	34	57.6	18	33.3
I don't pay much attention to this information.	161	11.1	7	11.9	9	16.7
I don't know how to use this information.	168	11.6	8	13.6	15	27.8
Other	8	0.6	0	0.0	2	3.7

The above comments demonstrated that the test takers might not fully understand or appreciate the feedback because of a lack of understanding of its meaning or its intended uses. The finding also reflects one of the barriers identified in Carless and Boud (2018) regarding test takers' ineffective use of feedback. The change of perceptions, nonetheless, pointed to a possibility that test takers' assessment literacy could be enhanced through training or guidance on how to use or interpret test performance feedback information.

All three institutional score users interviewed reported that they did not use the abilities-measured information for making any high-stakes decisions, such as for college admissions or graduation requirements. They also mentioned that they typically did not ask for or have access to their students' full TOEIC L&R test score reports and thus they did not use the abilities-measured data. In addition, the score users commented that they did not use the feedback to inform their instructions because the test was not required for all students and actually very few students in two users' schools (IDs 7 and 9) took the test. One said:

> Honestly, I never paid attention to the abilities-measured information until I participated in this study. Actually, because of this research study, I was reminded that maybe I should look at the information in the future. (ID 7)

When asked how she would use this piece of information, she responded:

> Well, like for reading, it shows the percent correct for vocabulary, grammar, and other reading skills. I would use the information to identify my students' strengths and weaknesses and to select instructional materials and design class activities. (ID 7)

Contrastively, a user who was a high-school English teacher commented:

> We only receive students' TOEIC L&R section and total scores in an EXCEL sheet from the test distributor in Taiwan. We never receive the abilities-measured information, so I don't really know how my students perform in these abilities. Also, because the TOEIC L&R test is not a required test for my students, it's just a standardized test that we encourage our students to take so that they can use the results for college applications, or they can use the scores to waive English language courses once they enter college. We will not use the abilities-measured information to inform our classroom instruction because it's not aligned to our curriculum and not all the students take the test. (ID 9)

The two corporate score users interviewed reported that they never used or considered the abilities-measured information for making any high-stakes decisions, such as hiring, promotions, or overseas job assignments. Whenever their companies used the TOEIC L&R test results for making important decisions about individuals, both said that only the total scaled scores were considered. Additionally, they mentioned that they hardly, if ever, paid attention to the abilities-measured information until they responded to the study surveys. The two users also commented that they did not think that other domestic or multinational corporations in Taiwan that used the TOEIC L&R test results would consider the abilities-measured information for making high-stakes decisions.

Taken together, the interviews provided further evidence to supplement the survey results, confirming that the abilities-measured information was generally used as intended by the stakeholders and that no inappropriate, high-stakes decisions were made based on the data.

RQ3. How well do TOEIC L&R test takers and institutional and corporate score users in Taiwan understand the abilities-measured information?

Table 6 shows the percentages of the participants in each stakeholder group answering the comprehension questions (CQ1 to CQ5) correctly. It should be noted that the five questions, corresponding to Survey Questions 7 to 11 (see Appendix B), had different numbers of answer options (between two and four). These comprehension questions were used to evaluate the stakeholders' understanding of the abilities-measured information, following the methodology used in previous score reporting research studies (e.g., Kannan et al., 2021). CQ1 and CQ3 assessed whether the participants understood the percent correct of abilities-measured information as shown in the bar graphs. The results showed that the three stakeholder groups had a good understanding of the bar graphs and that the great majority of the participants

Table 6 Percentages of Participants Getting the Comprehension Questions (CQ) Correct by Stakeholder Group

Question (brief description of question)	Test takers ($N = 1,444$)	Institutional score users ($N = 59$)	Corporate score users ($N = 54$)
CQ1 (inference test takers can make based on one ability measured)	64.8%	59.3%	55.4%
CQ2 (inference test takers can make based on the abilities-measured data)	43.6%	52.5%	26.8%
CQ3 (the ability measured with the highest percent correct)	90.4%	89.8%	85.7%
CQ4 (comparison of results across test forms or administrations)	42.7%	42.4%	42.9%
CQ5 (comparison of test results with those of a different test taker)	30.1%	33.9%	30.4%
All	54.3%	55.6%	48.2%

answered CQ3 correctly (group mean ranging from 88.9% to 90.4%). CQ2 evaluated whether the participants understood that the different pieces of percent correct of abilities-measured information should not be compared because the number of items measuring each ability differed slightly. The percentages of the participants answering this question correctly varied across the stakeholder groups, ranging from 26.8% for the corporate score users to 52.5% for the institutional score users, revealing that the corporate score users had a weaker understanding of the information.

Given that the footnote in the score report explicitly states that the abilities-measured information cannot be compared across test forms and administrations, CQ4 and CQ5 were designed to evaluate whether the participants understood this message. The results showed that, regardless of the stakeholder groups, less than half of the participants answered the questions correctly (between 30.1% and 42.9%), demonstrating a limited understanding of the footnote message.

Some variations regarding how well the three groups comprehended the abilities-measured information were observed. The average percentages taken across the five comprehension questions indicated that the test takers and institutional score users did roughly equally well (54.3% for test takers; 55.6% for institutional score users), whereas the corporate score users performed slightly less well with an average of 48.2%. Compared to the findings of Hsieh (2023) where TOEIC L&R test takers in Japan ($N = 834$) answered, on average, 60.9% of the same five comprehension questions correctly, the relatively lower percentages seen in the current study raised the concern that the TOEIC L&R test takers and score users in Taiwan may have more challenges in understanding the abilities-measured information and the explanatory texts in the footnote message. As mentioned earlier, the TOEIC L&R score report in Hsieh (2023) was presented in Japanese to the test takers and research participants in Japan and the one used in the current study was presented in the English language. It was speculated that this difference in the language used in the score reports might have impacted stakeholders' comprehensions of the abilities-measured information.

The interview data provided further insights into how the test takers and score users interpreted the abilities-measured information. Most of the 11 participants interviewed did not know how the abilities-measured data was computed. When asked what the percent correct of abilities-measured data meant, one responded:

> I think the percent correct of abilities-measured information was calculated based on the comparison of my performances in these language skills against other test takers who took the test on the same day. (ID 2)

Another test taker who had taken the TOEIC L&R test 10 times at the time of the interview said that:

> I didn't understand what the abilities-measured information meant. I think this information is showing the percentile ranking, like how you perform in relation to others. (ID 4)

When asked whether she had read the footnote that explained how to interpret the abilities-measured information, she reported:

> I wasn't aware that there was a footnote in the score report until I participated in this study. I think most test takers wouldn't notice it and would interpret the abilities-measured data the way I did. (ID 4)

The five score users provided limited input regarding how they interpreted or used the abilities-measured information because all of them reported that they had never seen the information or ever paid much attention to it before participating

Table 7 Helpfulness of the TOEIC L&R Score Report Elements

	Test takers		Institutional score users		Corporate score users	
	M	*SD*	*M*	*SD*	*M*	*SD*
Listening score	3.14	0.62	3.32	1.23	3.59	0.71
Reading score	3.15	0.62	3.31	1.23	3.57	0.74
Total score	3.20	0.63	3.59	0.95	3.56	0.86
Score descriptors	3.05	0.67	3.19	1.28	3.22	0.94
Percent correct of abilities measured	3.10	0.65	2.98	1.51	3.15	0.99
Footnote	3.04	0.67	3.05	1.40	3.07	1.09

in the study. When asked whether they would use the information in the future, one institutional score user who was a college English instructor said:

> Actually now that I'm aware of the abilities-measured information, I think I will pay some attention to it in the future. (ID 7)

RQ4. How do the TOEIC L&R test takers and institutional and corporate score users in Taiwan perceive the usefulness of the score report?

Table 7 presents the means and standard deviations of the participants' ratings on the usefulness of the different score report elements on a scale of 1 to 4. The three stakeholder groups generally found that the different score report elements helpful. Unsurprisingly, given that the TOEIC L&R section and total scaled scores were the information most score users used for making important, high-stakes decisions as the survey results and interview data suggested, they were also rated as the most useful elements of the score report.

The three stakeholder groups' perceptions of the listening and reading score descriptors were generally positive, although the test takers gave the least favorable mean rating ($M = 3.05$) among the three groups. The score descriptors were intended to provide test takers with a snapshot of what they could do based on their TOEIC L&R test scores. As mentioned earlier, these descriptors are provided as one piece of performance feedback in broad score levels. It is possible that the information was not perceived as very helpful because the descriptors were not created based on each individual test taker's test performances. The interview data provided some evidence to substantiate this possibility. One test taker who had taken the TOEIC L&R test more than four times commented:

> The score descriptors are a bit too vague for me. For example, in this sample score report, it only describes test takers who score around 200 on the listening section and what these test takers can do. The descriptors do not provide any actionable feedback or tell you what you can do next. After I took the test a few times, I felt that I could just skip this feedback section altogether because it was not very helpful. (ID 5)

Another test taker commented that the score descriptors were too difficult to understand and said:

> Many of my friends who took the TOEIC L&R test didn't understand the score descriptors. A lot of the vocabulary that you use in the descriptors are unfamiliar to us, like my friends who scored low on the TOEIC L&R test, they couldn't really understand what the descriptors meant. (ID 6)

When asked how the score report could be improved, she suggested:

> It would be good if you can simplify the score descriptors and use simpler vocabulary. You may also want to reduce the number of texts here. Right now it's too text-heavy and very overwhelming. You can change the texts and replace them with some kind of graphical representations, like pie charts or bar graphs. I think visual representations will be more intuitive and easier to understand for most test takers. (ID 6)

The abilities-measured information was perceived as useful overall, although the institutional score users rated the information as the least useful among all the reporting elements, with a mean rating of 2.98. The institutional score users

were arguably the group of stakeholders who could benefit most from this piece of information because it provided them with information about students' strengths and weaknesses that could inform instructional practices. The fact that they rated the percent correct of abilities-measured information as least useful signals that this feedback may not meet all the institutional score users' needs for feedback. This could be partly because some institutional score users, particularly classroom teachers, did not necessarily have access to this piece of information as commented above or that the test content was not aligned to the course content. Since the feedback was not tied to the students' classroom performances or learning content, the abilities-measured information was perceived as not directly relevant to the local teaching and learning contexts and unhelpful by some teachers.

The footnote was rated as the least useful score report element by the test takers and the corporate scores users. The footnote was printed at the bottom of the score report with small font and was not very visible. It seems reasonable to assume that many test takers did not notice its existence. As one test taker (ID4) noted in the interview, when asked what she thought about the footnote, she reported that she did not know that there was a footnote until she participated in this study. For the corporate score users, since most of them did not use the abilities-measured information to make important decisions, they might not find the footnote useful or needed. It is also possible that the footnote message was confusing or difficult to understand for some score users, as shown in the results of the comprehension questions. This lack of understanding was likely to lead to a negative perception of the usefulness of the footnote itself for the score users.

Discussion and Conclusion

Proper understanding and use of score report information is central to the concept of validity (Tannenbaum, 2019). This study surveyed three stakeholder groups of the TOEIC L&R test in Taiwan and investigated how test takers and score users interpreted and used the test results. Results of the survey responses and interview data converged and showed that the stakeholders used the TOEIC L&R test section and totaled scaled scores to evaluate a test taker's language proficiency and make various decisions about individuals as intended by the test developer. When considering how they would use the abilities-measured feedback information, stakeholders indicated that they primarily used the information to identify test takers' strengths and weaknesses, but some used it to monitor skill improvement. No unintended use of the information for making high-stakes decisions about individuals was reported. Stakeholders' comprehension of the abilities-measured information demonstrated that they had a generally good understanding the abilities assessed by the TOEIC L&R test. These findings provide evidence substantiating the claims about the meaningfulness of the TOEIC L&R test scores and the appropriateness of score-based decisions.

Regarding score report interpretation, the study findings further provide theoretical implications by corroborating previous research, which suggests that test takers and score users often experience difficulties understanding certain score report information (Kannan et al., 2018; Kim et al., 2019; Whittaker et al., 2011; Zapata-Rivera & Katz, 2014). The results revealed that the abilities-measured information, despite its theoretical foundation, did not effectively convey information in a way understandable to some stakeholders and raised the need for score report refinements. In addition, many stakeholders indicated in the survey that they used the abilities-measured information to monitor test takers' progress in English, showing that these survey respondents were not fully aware of or knowledgeable about the need for caution when using the information to make comparisons across test forms and administrations. Part of the reasons why stakeholders in this study did not have a full understanding of the abilities-measured information could be that the explanatory texts were confusing and/or the footnote was overlooked by users due to the design of the score report (e.g., small font size, the position of the footnote).

Another issue that is worthy of attention is that the corporate score users had the lowest level of feedback comprehension, on average, among the three stakeholder groups. Since the TOEIC L&R test's corporate score users in Taiwan typically only use the total scaled scores to make decisions about individuals, it is reasonable to assume that they do not have a practical need to use the abilities-measured information or a real need to understand its meaning. Nevertheless, as the findings revealed, score users' awareness of the existence of the abilities-measured information was raised as a result of participating in this study and their attitudes toward using such information changed. This finding implies that score users' engagement with test performance feedback can be promoted when they are instructed on how to use and interpret the feedback information as intended (Carless & Boud, 2018; Min et al., 2022). The result also points to the importance of enhancing stakeholders' assessment literacy and awareness of different types of information provided in test score reports.

The score descriptors and abilities-measured information were perceived as not very useful by some test takers and institutional score users. This finding revealed a need for refining the descriptors with more digestible and easily accessible information. Given that the TOEIC L&R test is a standardized language proficiency test, it is typical that the test score report does not include the kinds and levels of feedback that would be tailored to personal performance profiles or sufficient to support classroom instruction. Furthermore, the test content is not linked to any local curriculum or educational contexts and therefore the test performance feedback is not expected to forge a strong link between assessment and classroom instruction, which was pointed out by one institutional score user. On the other hand, one college English instructor showed an interest in using the performance feedback to select instructional materials and design classroom activities, suggesting that a bridge can be made between standardized language assessment and classroom instruction through external feedback presented in standardized, large-scale workplace English test score report.

Results of this study contribute to the growing body of score reporting literature that focuses on the underlying validity arguments around test score interpretation and use (Zapata-Rivera et al., 2021; Zapata-Rivera & Katz, 2014). By investigating how the three stakeholder groups use and interpret the TOEIC L&R test scores and performance feedback information, the study intends to inform future design and development of score reports with more meaningful and accessible score reporting information to facilitate stakeholder comprehension and use of assessment results. The study findings demonstrated that test takers and score users had varying degrees of comprehension of the abilities-measured information and perceived differently the usefulness of the different score report elements. These findings highlight the importance of keeping in mind the diversity of stakeholder needs and backgrounds when communicating test results and feedback information to the intended audience. Although developing audience-specific score reports as suggested by previous studies (e.g., Kannan et al., 2018; Zapata-Rivera & Katz, 2014) may be challenging for large-scale standardized language assessments that have heterogenous stakeholder population such as the TOEIC L&R test, one alternative to addressing the needs of diverse stakeholder groups is to provide interpretive or supporting materials that can be useful for various users. Providing resources in English and in the test taker's native language, or having bilingual versions of score reports, for example, may facilitate accurate interpretations of test results.

The study findings have practical implications to inform further refinements of the TOEIC L&R score report. The following recommendations are made based on the results:

1. Improve the wording of the footnote and convey the message at a level of understanding appropriate for the test takers and score users receiving or using the score report. Use language that will be easily understood by the different stakeholder groups and to the extent possible, use a layperson's description of the technical terms.
2. Reduce the length of the listening and reading score descriptors and focus on the key message to avoid overwhelming individuals reading the score report.
3. Provide additional print and digital resources to support appropriate interpretation and use of the score report and distribute the resources as widely as possible.
4. Conduct studies to evaluate user perceptions of different score report designs with clear visual representations of the feedback information.
5. Engage local partners, test distributors, and stakeholders in discussions about the research findings and how they might be used to enhance score report comprehension and assessment literacy. This can be done in several formats, such as webinars, face-to-face meetings, workshops in which the research report or findings serve as the springboard for brainstorming possible actions.

Some limitations of the study need to be pointed out. First, the research participants were TOEIC L&R test takers and score users in Taiwan. Thus, the results cannot be generalized to other stakeholders outside this context. In addition, the number of interview participants was very small due to challenges in recruitment, which also limited the generalizability of the findings. Second, as the study findings revealed, some score users only had access to partial information of the score report; this lack of prior exposure to the full report, especially the abilities-measured information, might have influenced how the participants responded to the survey questions, even though images of the information were presented in the survey to minimize the potential impact. Future research could address this limitation by surveying score users who have had prior experience with using the full score report to further verify how stakeholders use and interpret the feedback information. Third, the English version of the TOEIC L&R score report is given to test takers and score users in Taiwan. As the study finding implied, an individual's English language proficiency may play a role in how well one understands the score reporting information. Because the study relied on self-reported TOEIC L&R test scores and only a small percentage

of the participants reported their most recent test scores, the study did not consider English language proficiency as a variable that may impact the participants' comprehension of the test performance feedback. Future research should investigate this possibility by using official TOEIC L&R test scores that research participants receive to investigate the role of language proficiency on one's ability to understand the score report information. Findings of such investigations can then be used to inform whether a more audience-centered score report, such as having bilingual versions of the score report, is needed to support different target audiences (Zapata-Rivera & Katz, 2014). Future studies should also examine test takers' and score users' assessment literacy, or lack thereof, that allows them to adequately use and interpret score reporting information. Such investigation can provide insights into the specific assessment literacy skills needed to more effectively interpret test results and performance feedback presented in score reports and guide the development of resources to enhance stakeholders' assessment literacy.

References

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

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

Baron, P., Linquanti, R., & Huang, M. (2020). Validating threshold scores for English proficiency assessment uses. In M. K. Wolf (Ed.), *Assessing English language proficiency in U.S. K-12 schools* (pp. 161–184). Routledge. 10.4324/9780429491689-9

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, *43*(8), 1315–1325. https://doi.org/10.1080/02602938.2018.1463354

Cid, J., Wei, Y., Kim, S., & Hauck, C. (2018). Statistical analyses for the updated TOEIC® Listening and Reading test. In D. Powers & J. Schmidgall (Eds.), *The research compendium for the TOEIC tests* (Vol. *III*, pp. 4.1–4.16). ETS.

Cizek, G. J., Bunch, M. B., & Koons, H. (2005). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice, 23*(4), 31–50. https://doi.org/10.1111/j.1745-3992.2004.tb00166.x

Clark, A. K., Nash, B., & Karvonen, M. (2022). Teacher assessment literacy: Implications for diagnostic assessment systems. *Applied Measurement in Education*, *35*(1), 17–32. https://doi.org/10.1080/08957347.2022.2034823

ETS. (2022a). *TOEIC® Listening and Reading test: Examinee handbook*. ETS. https://www.ets.org/content/dam/ets-org/pdfs/toeic/toeic-listening-reading-test-examinee-handbook.pdf

ETS. (2022b). *TOEIC® 2021 report on test takers worldwide—TOEIC Listening and Reading test*. ETS. https://www.ets.org/content/dam/ets-org/pdfs/toeic/toeic-listening-reading-report-test-takers-worldwide.pdf

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. TOEFL RR-84). ETS. https://doi.org/10.1002/ets2.12231

Gomez, P. G., Noah, A., Schedl, M. A., Wright, C., & Yolkut, A. (2007). Proficiency descriptors based on a scale-anchoring study of the new TOEFL iBT® reading test. *Language Testing, 24*(3), 417–444. https://doi.org/10.1177/0265532207077209

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education, 17*(2), 145–220. https://doi.org/10.1207/s15324818ame1702_3

Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika, 75*(2), 209–227. https://doi.org/10.1007/s11336-010-9158-4

Hambleton, R. K., & Zenisky, A. (2013). Reporting test scores in more meaningful ways: Some new findings, research methods, and guidelines for score report design. In K. F. Geisinger (Ed.), *American Psychology Association handbook of testing and assessment in psychology, Vol. 3. Testing and assessment in school psychology and education* (pp. 479–494). American Psychological Association. https://doi.org/10.1037/14049-023

Hsieh, C.-N. (2023). *Evaluating the use and interpretation of the TOEIC® linstening and reading test score report: Perspectives of test takers in Japan* (Research Report No. RR-23-02). ETS. https://doi.org/10.1002/ets2.12364

Jang, E. E. (2013). Diagnostic assessment in language classrooms. In A. J. Kunnan (Ed.), *Companion to language assessment* (pp. 113–126). Wiley-Blackwell.

Kannan, P., Zapata-Rivera, D., & Bryant, A. D. (2021). Evaluating parent comprehension of measurement error information presented in score reports. *Practical Assessment, Research, and Evaluation, 26*(12), 1–21. https://doi.org/10.7275/rgwg-t355

Kannan, P., Zapata-Rivera, D., & Leibowitz, E. A. (2018). Interpretation of score reports by diverse subgroups of parents. *Educational Assessment, 23*(3), 173–194. https://doi.org/10.1080/10627197.2018.1477584

Kim, A. A. (2015). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing, 32*(2), 227–258. https://doi.org/10.1177/0265532214558457

Kim, A. A., Chapman, M., Kondo, A., & Wilmes, C. (2019). Examining the assessment literacy required for interpreting score reports: A focus on educators of K-12 English learners. *Language Testing, 37*(1), 54–75. https://doi.org/10.1177/0265532219859881

Kunnan, A. J., & Jang, E. E. (2009). Diagnostic feedback in language assessment. In M. H. Long & C. J. Doughty (Eds.), *The handbook of language teaching* (pp. 610–627). Blackwell. https://doi.org/10.1002/9781444315783.ch32

Liao, C.-W. (2010). *TOEIC® Listening and Reading test scale anchoring study* (TOEIC® Compendium TC10-05). ETS. https://www.ets.org/Media/Research/pdf/TC-10-05.pdf

MacIver, R., Anderson, N., Costa, A.-C., & Evers, A. (2014). Validity of interpretation: A user validity perspective beyond the test score. *International Journal of Selection and Assessment, 22*(2), 149–164. https://doi.org/10.1111/ijsa.12065

Min, S., Zhang, J., Li, Y., & He, L. (2022). Bridging local needs and national standards: Use of standards-based individualized feedback of an in-house EFL listening test in China. *Language Testing, 39*(3), 425–452. https://doi.org/10.1177/02655322211070990

O'Leary, T. M., Hattie, J. A. C., & Griffin, P. (2017). Actual interpretations and use of scores as aspects of validity. *Educational Measurement: Issues and Practice, 36*(2), 16–23. https://doi.org/10.1111/emip.12141

Papageorgiou, S., & Choi, I. (2018). Adding value to second-language listening and reading subscores: Using a score augmentation approcah. *International Journal of Testing, 18*(3), 207–230. https://doi.org/10.1080/15305058.2017.1407766

Powers, D. E., Schedl, M. A., & Papageorgiou, S. (2017). Facilitating the interpretation of English language proficiency scores: Combining scale anchoring and test score mapping methodologies. *Language Testing, 34*(2), 175–195. https://doi.org/10.1177/0265532215623582

Sawaki, Y., & Koizumi, R. (2017). Providing test performance feedback that bridges assessment and instruction: The case of two standardized English language tests in Japan. *Language Assessment Quarterly, 14*(3), 234–256. https://doi.org/10.1080/15434303.2017.1348504

Schedl, M. A. (2010). *Background and goals of the TOEIC® Listening and Reading test redesign project*. (TOEIC Compendium TC10-02). ETS. https://www.ets.org/Media/Research/pdf/TC-10-02.pdf

Tannenbaum, R. J. (2019). Validity aspects of score reporting. In D. Zapata-Rivera (Ed.), *Score reporting research and applications* (pp. 9–18). Routledge. https://www.routledge.com/Score-Reporting-Research-and-Applications/Zapata-Rivera/p/book/9780815353409

Van den Heuvel, J. R., Zenisky, A., & Davis-Becker, S. (2014, April 3–7). *Applying lessons learned in educational score reporting to credentialing* [Paper presentation]. American Educational Research Association, Philadelphia, Pennsylvania, United States. https://www.alpinetesting.com/wp-content/uploads/2017/09/Applying-Lessons-Learned-in-Educational-Score-Reporting-to-Credentialing.pdf

Van der Kleij, F. M., Eggen, T. J. H. M., & Engelen, R. J. H. (2014). Towards valid score reports in the Computer Program LOVS: A redesign study. *Studies in Educational Evaluation, 43*, 24–39. https://doi.org/10.1016/j.stueduc.2014.04.004

Whittaker, T. A., Williams, N. J., & Dodd, B. G. (2011). Do examinees understand score reports for alternate methods of scoring computer-based tests? *Educational Assessment, 16*(2), 69–89. https://doi.org/10.1080/10627197.2011.582442

Zapata-Rivera, D., Andrews-Todd, J., & Oliveri, M. E. (2021). Communicating assessment information in the context of a workplace formative task. *Journal of Writing Analytics, 5*, 324–341. https://wac.colostate.edu/docs/jwa/vol5/zapata-rivera.pdf

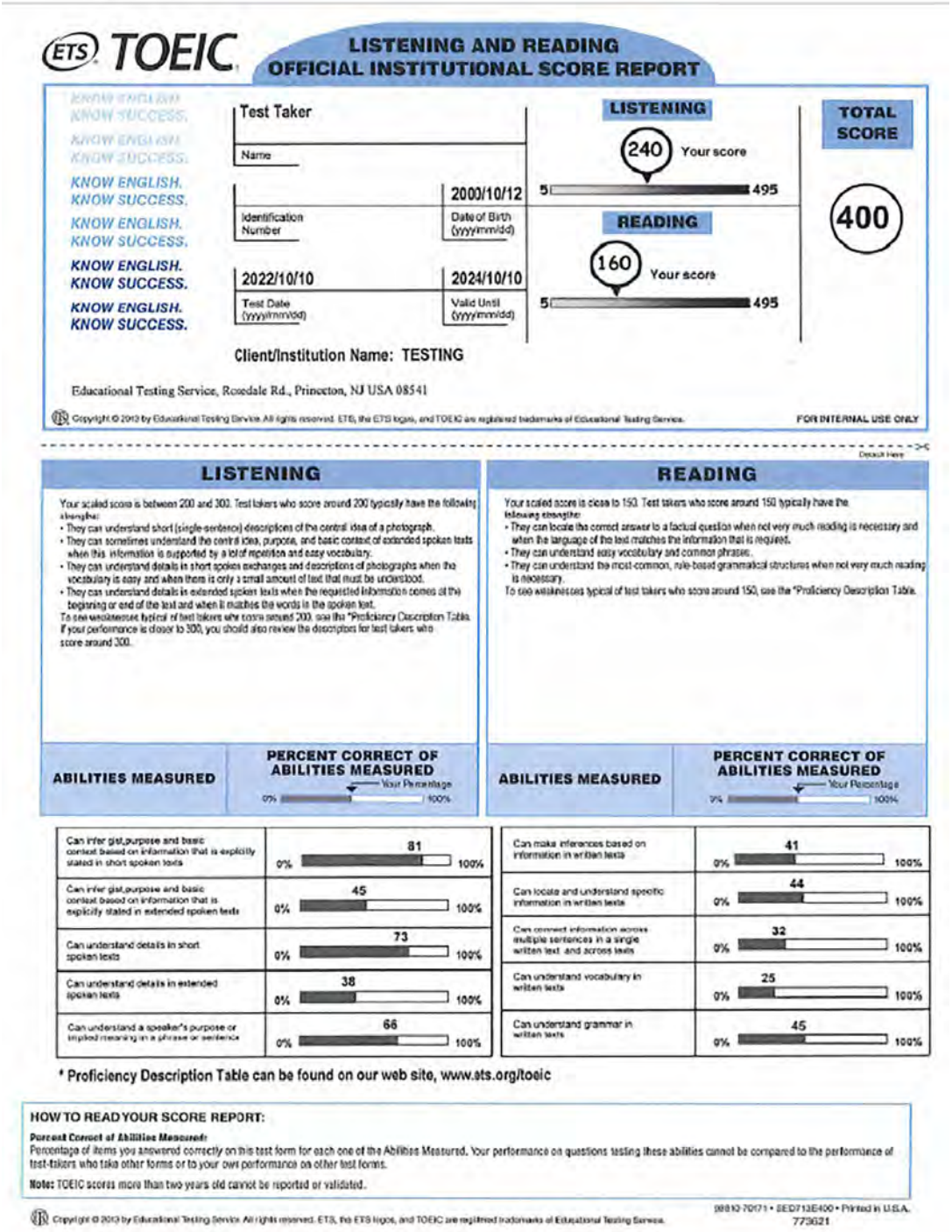
Zapata-Rivera, D., & Katz, I. R. (2014). Keeping your audience in mind: Applying audience analysis to the design of interactive score reports. *Assessment in Education: Principles, Policy & Practice, 21*(4), 442–463. https://doi.org/10.1080/0969594X.2014.936357

Zapata-Rivera, D., VanWinkle, W., & Zwick, R. (2012). *Applying score design principles in the design of score reports for CBAL teachers* (ETS Research Memorandum No. RM-12-20). ETS.

Zenisky, A., & Hambleton, R. K. (2012). Developing test score reports that work: The process and best practices for effective communication. *Educational Assessment: Issues and practice, 31*(2), 21–26. https://doi.org/10.1111/j.1745-3992.2012.00231.x

Appendix A

Sample TOEIC Listening and Reading Test Score Report



ETS TOEIC

LISTENING AND READING
OFFICIAL INSTITUTIONAL SCORE REPORT

KNOW ENGLISH. KNOW SUCCESS.

Test Taker

Name

Identification Number

2000/10/12
Date of Birth (yyyy/mm/dd)

2022/10/10
Test Date (yyyy/mm/dd)

2024/10/10
Valid Until (yyyy/mm/dd)

LISTENING
240 Your score
5 495

READING
160 Your score
5 495

TOTAL SCORE
400

Client/Institution Name: TESTING

Educational Testing Service, Rosedale Rd., Princeton, NJ USA 08541

Copyright © 2013 by Educational Testing Service. All rights reserved. ETS, the ETS logo, and TOEIC are registered trademarks of Educational Testing Service. FOR INTERNAL USE ONLY

Detach Here

LISTENING

Your scaled score is between 200 and 300. Test takers who score around 200 typically have the following strengths:
- They can understand short (single-sentence) descriptions of the central idea of a photograph.
- They can sometimes understand the central idea, purpose, and basic context of extended spoken texts when this information is supported by a lot of repetition and easy vocabulary.
- They can understand details in short spoken exchanges and descriptions of photographs when the vocabulary is easy and when there is only a small amount of text that must be understood.
- They can understand details in extended spoken texts when the requested information comes at the beginning or end of the text and when it matches the words in the spoken text.

To see weaknesses typical of test takers who score around 200, see the *Proficiency Description Table. If your performance is closer to 300, you should also review the descriptors for test takers who score around 300.

ABILITIES MEASURED	PERCENT CORRECT OF ABILITIES MEASURED (Your Percentage, 0% – 100%)
Can infer gist,purpose and basic context based on information that is explicitly stated in short spoken texts	81
Can infer gist,purpose and basic context based on information that is explicitly stated in extended spoken texts	45
Can understand details in short spoken texts	73
Can understand details in extended spoken texts	38
Can understand a speaker's purpose or implied meaning in a phrase or sentence	66

READING

Your scaled score is close to 150. Test takers who score around 150 typically have the following strengths:
- They can locate the correct answer to a factual question when not very much reading is necessary and when the language of the text matches the information that is required.
- They can understand easy vocabulary and common phrases.
- They can understand the most common, rule-based grammatical structures when not very much reading is necessary.

To see weaknesses typical of test takers who score around 150, see the *Proficiency Description Table.

ABILITIES MEASURED	PERCENT CORRECT OF ABILITIES MEASURED (Your Percentage, 0% – 100%)
Can make inferences based on information in written texts	41
Can locate and understand specific information in written texts	44
Can connect information across multiple sentences in a single written text and across texts	32
Can understand vocabulary in written texts	25
Can understand grammar in written texts	45

* Proficiency Description Table can be found on our web site, www.ets.org/toeic

HOW TO READ YOUR SCORE REPORT:

Percent Correct of Abilities Measured:
Percentage of items you answered correctly on this test form for each one of the Abilities Measured. Your performance on questions testing these abilities cannot be compared to the performance of test-takers who take other forms or to your own performance on other test forms.

Note: TOEIC scores more than two years old cannot be reported or validated.

Copyright © 2013 by Educational Testing Service. All rights reserved. ETS, the ETS logo, and TOEIC are registered trademarks of Educational Testing Service.

[8810-7017] • SED713E400 • Printed in U.S.A.
773621

Appendix B

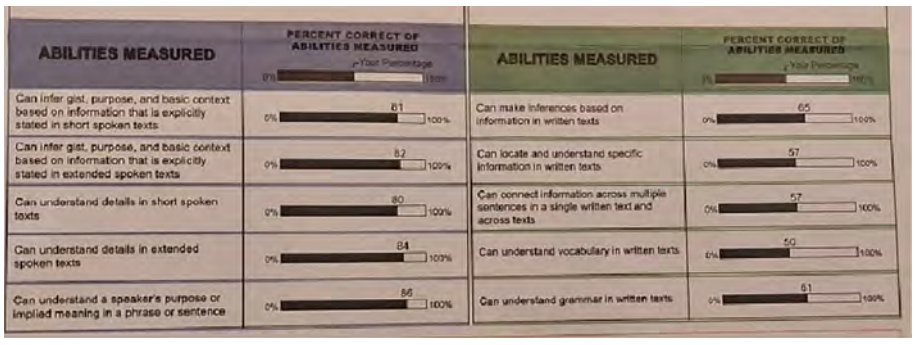
Sample Online Survey

TOEIC Listening and Reading Score Report Study

Test-Taker Online Survey

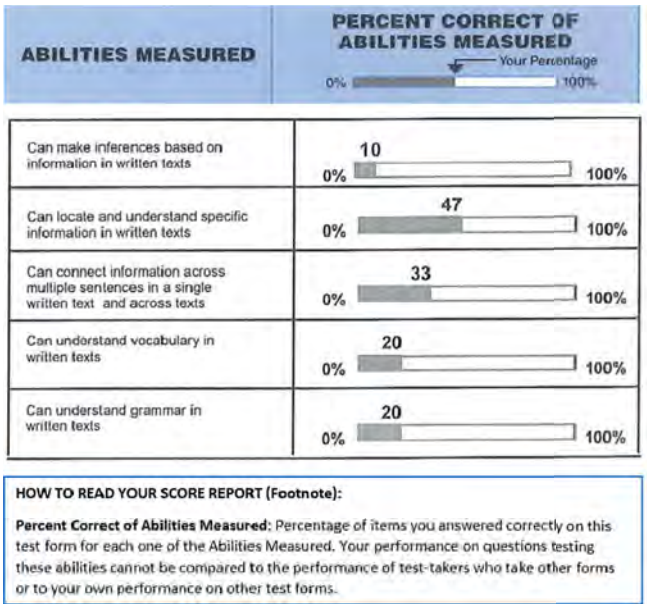
Part 1. Background information

1. Gender ___ male ____ female ____ other _____ prefer not to answer _____
2. Age _____
3. Which of the following best describes your current status?
   - A. I am employed full-time (including self-employed).
   - B. I am employed part-time.
   - C. I study part-time.
   - D. I am a full-time student.
   - E. Other
4. How many times have you taken the TOEIC Listening and Reading test?
   - A. 1
   - B. 2
   - C. 3
   - D. 4 or more
5. What did you use your TOEIC Listening and Reading section and total scores for? (Check all that apply)
   - A. To apply for jobs
   - B. To get promoted at work
   - C. To pass job-related exams
   - D. To evaluate my English proficiency
   - E. To improve my English proficiency
   - F. To meet a school language requirement
   - G. To get an internationally recognized English proficiency test certificate
   - H. Other (Please specify)
6. In your TOEIC Listening and Reading score report, there is a section about the abilities measured (See sample image below). How do you use the abilities-measured information?
   - A. I use it to compare the strengths and weaknesses of my reading and listening skills.
   - B. I use it to monitor the improvement of my reading and listening skills over time.
   - C. I don't pay much attention to the abilities-measured information.
   - D. I don't know how to use this information.



Part 2. Comprehension questions

In this section, you will see a snapshot of TOEIC Listening and Reading score report of a hypothetical test taker, Jing. The snapshot shows (1) the reading abilities measured, (2) percent correct of the reading abilities measured, and (3) the footnote that explains how to read the percent correct of abilities measured. Use the information provided in the snapshot to answer questions 7 to 11.



7. Based on the information presented in the snapshot above, what can you infer about Jing's ability to connect information across multiple sentences in a single written text and across texts?

   A. Jing correctly answered 33 questions measuring this ability on this test.
   B. Jing's performance on this test indicates that she has mastered 33% of this ability.
   C. Jing correctly answered 33% of the questions measuring this ability on this administration of the test.
   D. I don't know.

8. Based on the information presented in the bar graphs, which of the following inferences can be made?

   A. Jing's ability to make inferences based on written texts is comparable to her ability to understand vocabulary from written texts.
   B. Jing's ability to understand vocabulary in written texts is comparable to her ability to understand grammar in written texts.
   C. Since the numbers of questions in each ability measured are different, performances cannot be compared across the different abilities measured.

9. Based on the information presented in the bar graphs, which of the following represents the ability where Jing got the highest percentage of questions correct on this test?

   A. Locating and understanding specific information in written texts.
   B. Connecting information across sentences in written texts.
   C. Making inferences based on information in written texts.

10. Can Jing compare the percent correct of abilities-measured information she received on this test to the results of a TOEIC Listening and Reading test she took on a different day?

   A. Yes.
   B. No.

11. Can Jing compare the percent correct of abilities-measured information she received on this test to information her friend received on a TOEIC Listening and Reading test her friend took?
    A. Yes.
    B. No.

Part 3. Perception of TOEIC Listening and Reading score report

12. To what extent did you find the following information useful to have in the TOEIC Listening and Reading score report?

	Not at all useful	Not very useful	Somewhat useful	Very useful
Listening score				
Reading score				
Total score				
Listening and reading proficiency descriptions				
Listening and reading abilities measured				
Bar graphs of percent correct of abilities measured				
How to read your score report footnote				

Appendix C

Interview Questions

TOEIC Listening and Reading Test-Taker Interview Questions

1. Please describe your TOEIC Listening and Reading test-taking experience. (e.g., When did you take the test? How did you prepare for the test?)
2. What did you use your TOEIC Listening and Reading test scores for? Please provide some examples.
3. What is the most important piece of information on the TOEIC Listening and Reading test score report for you? Why is it important?
4. Did you find the information provided in your TOEIC Listening and Reading test score report easy to understand? Why or why not?
5. How do you interpret the percent-correct scores on the abilities measured in the TOEIC Listening and Reading test score report? (Show a sample TOEIC Listening and Reading test score report to the test taker.)
6. Did you have any difficulty understanding the TOEIC Listening and Reading test score report?
    a. If yes, please describe the difficulties you had.
    b. If yes, what did you try to do to understand the score report?
7. If you could change one thing about the TOEIC Listening and Reading test score report, what would it be and why?

Institutional score user interview questions

1. Please describe your teaching context and how your school uses the TOEIC Listening and Reading test scores.
2. Do you help your students prepare for the TOEIC Listening and Reading test? If yes, please describe how you help them prepare for the test.
3. Do you receive your students' TOEIC Listening and Reading test scores?
    a. If yes, what do you use your students' test scores for?
    b. If no, would you like to receive your students' test scores? Why or why not?

4. Do you have access to your students' TOEIC Listening and Reading test score reports? If yes, please answer the following questions regarding the abilities-measured information presented in the TOEIC Listening and Reading test score report.
    a. How useful is this piece of information to you?
    b. Do you use this piece of information to make any high-stakes decisions?
    c. Do you have any difficulties in using this piece of information?
    d. Do you think that the information on the abilities measured accurately reflects your students' English listening and reading abilities? Why or why not?
5. Here is a sample TOEIC Listening and Reading test score report. Please have a look. If you could change one thing about the TOEIC Listening and Reading test score report, what would it be and why?
6. Is there any additional information you would like to receive about your students' TOEIC Listening and Reading test performances? Please explain.

Corporate score user interview questions

1. Please describe your organization and how your organization uses the TOEIC Listening and Reading test scores.
2. Do you require job applicants or your current employees to provide their TOEIC Listening and Reading test scores?
    a. If yes, what do you use the test scores for?
    b. If no, would you like to receive their test scores? Why or why not?
3. Do you require job applicants or your current employees to provide their TOEIC Listening and Reading test score report for various purposes? If yes, answer the following questions regarding the abilities-measured information shown in the TOEIC Listening and Reading test score report.
    a. How do you use this piece of information?
    b. Do you use this piece of information to make any high-stakes decisions?
    c. How useful is this piece of information to you or your organization?
    d. Do you have any difficulties in using this piece of information?
4. Here is a sample TOEIC Listening and Reading test score report. Please have a look. If you could change one thing about the TOEIC Listening and Reading test score report, what would it be and why?

Suggested citation:

Hsieh, C.-N. (2023). *Interpretation and use of a workplace English language proficiency test score report: Perspectives of TOEIC® test takers and score users in Taiwan* (Research Report No. RR-23-10). ETS. https://doi.org/10.1002/ets2/12373

Action Editor: Jonathan Schmidgall

Reviewers: Lin Gu and Saerhim Oh