Research Report

Investigating Fairness Claims for a General-Purposes Assessment of English Proficiency for the International Workplace: Do Full-Time Employees Have an Unfair Advantage Over Full-Time Students?

ETS RR–24-06

Jonathan Schmidgall
Yan Huo
Jaime Cid
Youhua Wei

December 2024

ETS Research Memorandum 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

ETS RESEARCH REPORT

Investigating Fairness Claims for a General-Purposes Assessment of English Proficiency for the International Workplace: Do Full-Time Employees Have an Unfair Advantage Over Full-Time Students?

Jonathan Schmidgall, Yan Huo, Jaime Cid, & Youhua Wei

ETS Research Institute, Princeton, New Jersey, United States

The principle of fairness in testing traditionally involves an assertion about the absence of bias, or that measurement should be impartial (i.e., not provide an unfair advantage or disadvantage), across groups of test takers. In more general-purposes language testing, a test taker's background knowledge is not typically considered relevant to the measurement of language proficiency; consequently, if there are systematic differences in background knowledge between groups of test takers this background knowledge should not provide an unfair advantage or disadvantage. As a general-purposes assessment of English for everyday life and the international workplace, the TOEIC® Listening and Reading test is designed to assess the listening and reading comprehension skills of second language (L2) users of English. In this study, we investigated whether a group of test takers with more workplace experience (full-time employees) have an unfair advantage over test takers with less workplace experience (full-time students). We conducted DIF analysis using nine forms of the test (1,800 items) and flagged 18 items (1.0%) for statistical differential functioning. An expert panel reviewed the items and concluded that none of the items could be clearly identified as biased in favor of employed (or student) test takers. Follow-up analyses using score equity assessment found that test scores do not unfairly advantage fulltime employed (versus student) test takers. Finally, we performed a content review using two expert panels that led to examples of how workplace-oriented content is incorporated into test items without disadvantaging full-time students (versus full-time employees). The results of these analyses provide support for claims about the impartiality (or fairness) of TOEIC Listening and Reading test scores for postsecondary test takers and add to current research on the role of background knowledge and fairness for more general-purposes language assessments.

Keywords English for specific purposes; validity argument; TOEIC® test; assessment use argument; background knowledge; reading comprehension; listening comprehension; full-time workers; full-time employees; fairness claims; international workplace

doi:10.1002/ets2.12380

Fairness is a core principle in assessment and is deeply interwoven with the concepts of equity, validity, and reliability (Bachman & Palmer, 2010; Kunnan, 2018). It typically involves the expectation that all test takers are given an equal opportunity to demonstrate their knowledge and skills without any bias or disadvantage arising from aspects of their identify or background, such as racial, socioeconomic, or disability status (American Educational Research Association [AERA] et al., 2014). When a test developer fails to adequately address fairness concerns in the test development and quality control process, the meaning of test scores can be compromised, leading to biased interpretations about test takers' knowledge and abilities. Thus, fairness in testing is not just an ethical obligation; it is also a crucial aspect of test quality and the trustworthiness of test scores.

Although perceptions and definitions of fairness evolve over time (see Sireci & Randall, 2021), fairness in assessment is often defined as the degree to which the qualities of a test, test administration, and test scores are consistent and valid across different groups of test takers (ETS, 2015; Xi, 2010). This definition does not imply that there will be no differences in how difficult the test is for different groups of test takers who may legitimately differ in terms of their proficiency. Rather, it emphasizes that test takers should not be advantaged or disadvantaged based on aspects of their background or identity that are not directly relevant to the knowledge and skills being measured by an assessment.

Corresponding author: Jonathan Schmidgall, Email: jschmidgall@ets.org

The authors would like to thank IIBC Research and Development staff for participating in the content review and for their input on this study.

Closely related to the principle of fairness is the concept of absence of bias or impartiality in testing (AERA et al., 2014; Bachman & Palmer, 2010; Kunnan, 2007, 2018; Stoynoff, 2013). Bias is said to occur when aspects of a test's content, administration, or scoring systematically and improperly advantage one group of test takers over another (Camilli, 2006). Thus, bias—or an unfair advantage—can occur when test content somehow advantages a group of test takers over others based on background characteristics that are irrelevant to the skills being measured. In language assessment, investigations of potential bias often focus on aspects of test takers' identity or background (e.g., gender, native or first language [L1], age, academic major; Kunnan, 2018, p. 173).

Because fairness is a core principle of assessment and integral to test quality, test developers should clearly communicate how they address concerns about fairness within a broader argument for test use (Kunnan, 2018; Xi, 2010). In this study, we focus on a particular claim in the validity argument for a general-purposes test of English reading and listening comprehension: the impartiality or fairness of score interpretations across student versus employee test takers. We begin by unpacking the issue of test takers' background knowledge as a potential fairness issue in language assessment and the potential role of background knowledge in the assessment of second language (L2) reading and listening comprehension. This discussion provides a rationale for a claim about the impartiality of TOEIC® Listening and Reading (TOEIC L&R) test score interpretations for student and employed test takers, which we propose to investigate using a mixed methods research design.

Fairness in General-Purposes Versus Specific-Purposes Language Testing

One consideration for fairness in language testing derives from how the construct is defined. A construct definition typically includes a statement of the knowledge, skills, and abilities (KSAs) to be measured and the context in which KSAs will be measured. This context is often characterized as the target language use (TLU) domain (Bachman & Palmer, 2010), which may be viewed on a continuum from general (e.g., English for the workplace) to specific (e.g., English for aviation; Douglas, 2000). Researchers in the field of language for specific purposes focus on conceptualizing and defining TLU domains along this continuum within the two primary subbranches of language for academic purposes and language for occupational purposes (Knoch & Macqueen, 2016).

Within the context of language for occupational purposes, more specific TLU domains have a narrower focus, such as the English required to work in the aviation industry as a pilot or air traffic controller. More general domains have a broader focus that is not defined in reference to a specific industry or occupation. Consequently, TLU domain definitions for assessments that are more specific (e.g., aviation English) will correspond very closely to a specific real-world context (e.g., English used in communication between pilots and air traffic controllers). Domains that are more general (e.g., English for the workplace) are abstractions whose features are expected to broadly generalize across relevant real-world contexts (e.g., English used for email correspondence, meetings, presentations, travel) and should be applicable across a wide variety of industries and occupations.

One implication of the specificity of the TLU domain is the degree to which a test taker's background or content knowledge is expected to be relevant to the construct (Douglas, 2000), which has implications for fairness and bias. Typically, background knowledge is not represented in constructs that are bound to more general domains but is considered relevant for more specific domains. More general domains often include language tasks whose contextual features are broadly defined and thus may be expected to be relevant across subdomains and specific communicative situations. For example, topics related to the more general domain of workplace English should be familiar to language users across a variety of industries and career trajectories. To the extent that topics (or other contextual features) favor one subgroup within the domain (e.g., marketing managers) over others, they represent a source of potential bias and a threat to fairness or the impartiality of score interpretations (Bachman & Palmer, 2010). However, when the domain is more specific, background knowledge becomes an integral part of the construct and would not pose the same threat to fairness.

Background Knowledge as a Source of Bias in L2 Reading and Listening Assessment

Researchers typically differentiate background knowledge from language knowledge, skills, and proficiency (e.g., Douglas, 2000; Bachman & Palmer, 1996, 2010). Background knowledge refers to the prior knowledge of an individual test taker, such as knowledge of (or familiarity with) test content, such as the topic of a reading or listening comprehension

passage. Background knowledge may involve familiarity with any aspect of the context of test task input, including the setting, purposes, tone, norms, and genre (Douglas, 2000).

Background knowledge is part of many models of language production and comprehension. In the situation model of comprehension (Kintsch, 1998), which has influenced models of foreign language reading comprehension (e.g., Grabe, 2009; Grabe & Yamashita, 2022), readers construct a mental model of a text by integrating the text's most important propositions with their background knowledge. Thus, it is inevitable that language users will bring background knowledge to comprehension activities.

In her dissertation, Banerjee (2019) provided a comprehensive review of terms that have been used to characterize the nature of a test taker's background (or prior, content, topical) knowledge. The terms *prior knowledge* and *background knowledge* typically refer to more domain-general knowledge, or the "sum of what an individual knows" (Alexander et al., 1991, p. 333, as cited in Banerjee, 2019). The terms *content knowledge* and *topical knowledge* refer to more domain-specific knowledge. Other researchers have followed the practice of separating background knowledge into domain-general and domain-specific knowledge (Cai, 2013; Cai & Kunnan, 2018). In L2 reading and listening assessment, the impact of background knowledge on performance has been shown to be affected by the specificity of the test and test takers' level of language proficiency. Research has found support for a "two threshold" effect wherein test takers at lower levels of language proficiency have difficulty utilizing their background knowledge effectively and test takers at higher levels of language proficiency are able to use their linguistic knowledge to compensate for less background knowledge (Cai, 2013; Cai & Kunnan, 2018; Chung & Berry, 2000; Clapham, 1996; Ridgway, 1997).

In studies that have included L2 comprehension tests (or texts) of varying specificity, background knowledge has been found to better predict performance on more specific tests (Chung & Berry, 2000). This finding has led researchers to argue that "as texts become more specific, background knowledge becomes more important" (Chung & Berry, 2000, p. 208). Jensen and Hansen (1995) concluded that the effect of background knowledge was more substantial for technical versus nontechnical lectures in a listening comprehension test. In another study, when economics majors completed three specific-purposes reading comprehension tests that varied in the degree of specificity, the correlation between their specific-purposes test performance and general-purposes reading comprehension decreased with more specific tests (Tarlani-Aliabadi et al., 2022).

Other studies have found that background knowledge may impact performance on more general-purposes L2 comprehension tests, but the practical implications may be relatively trivial. In the context of more general-purposes assessment (i.e., TOEFL®), Hale (1988) concluded that test takers performed better on reading comprehension passages related to their general field of study (i.e., humanities and social sciences vs. biological and physical sciences), but the effect size was very small, translating to about 3 scaled score points on a scale that ranged from 237 to 677. Hill and Liu (2012) found that approximately 3% (two of 70) of TOEFL Reading comprehension items exhibited evidence of statistical bias (i.e., C level DIF) based on background knowledge but no evidence of passage-level bias. Jensen and Hansen (1995) found that overall, prior knowledge had a significant main effect for some (more technical) listening test passages but the effect size was trivial (i.e., ranged from $pr^2 = .03$ to .09). Karami & Alavi (2012) found that approximately 9% of items in a general academic English test of grammar, vocabulary, and reading comprehension exhibited statistical bias (i.e., large differential item functioning [DIF]) based on background knowledge but no bias in overall scores. In an analysis of strategy use with TOEFL iBT® Reading passages, Lee (2011) found that there was no difference in strategy use when test takers completed more familiar (versus less familiar) general academic reading passages; in other words, increased background knowledge did not impact strategy use. However, test takers reported more comfort and confidence with more familiar reading passages.

In summary, research investigating the potential effect of background knowledge in L2 reading and listening comprehension has found that background knowledge is expected to have a greater impact on performance for tests with higher specificity (i.e., more specific purposes), and for test takers above a minimum threshold of language proficiency.

The Context for This Study: Fairness in the TOEIC Listening and Reading Test Validity Argument

The TOEIC L&R test was designed to evaluate the English listening and reading proficiency of individuals whose first language is not English in everyday and workplace contexts (ETS, 2022a). The TLU domain referenced by the TOEIC tests is an example of a more general workplace domain (Knoch & Macqueen, 2016, p. 293). This TLU domain is not specific to any particular industry or occupation, and test content samples a variety of everyday and workplace contexts

and topics that are expected to be familiar to a broad range of young adults and adults who are L2 English users. The TOEIC L&R test includes settings and situations related to everyday activities (such as travel, entertainment, health, and dining out) and general workplace contexts (e.g., corporate development, finance, manufacturing, offices, personnel, and purchasing; see ETS, 2022a, pp. 3–4). Topical content that is situated in these contexts is developed in a manner that minimizes the use of industry-specific or otherwise specialized vocabulary and discourse genres.

Thus, as a more general-purposes language test, the abilities measured by the TOEIC L&R test are not intended to include specialized background knowledge of workplace contexts. In other words, such specialized knowledge—typically gained through experience—should not advantage or disadvantage test takers. This has implications for several claims about the fairness or impartiality of score interpretations that are stated in the validity argument for the test.

Bachman and Palmer's (2010) assessment use argument (AUA) is a comprehensive framework for constructing a validity argument that has been used for TOEIC program tests (e.g., Schmidgall, 2017; Schmidgall et al., 2021). An AUA consists of a set of interconnected statements (claims) about qualities of test scores, test score interpretations, intended uses, and consequences of test uses. One important quality of test score interpretations is that they are impartial to all groups of test takers. This claim about the impartiality of scores is elaborated in additional statements, including the following:

(1) TOEIC L&R test questions do not include response formats or content that favors or disfavors some test takers.
(2) Interpretations about test takers' listening and reading comprehension skills based on TOEIC L&R test scores are equally meaningful across different groups of test takers.

In a validity argument, claims are expected to be supported by evidence from test design, quality control documentation and procedures, and research (Bachman & Palmer, 2010). For example, Statement 1 above is supported by quality control procedures wherein test questions go through a multistep content and fairness review prior to administration that is designed to minimize the possibility that items could contain content that would unfairly advantage one group of test takers over another (e.g., employed vs. students, older vs. younger, males vs. females). All test questions receive a fairness review (see Zieky, 2013) from assessment development staff who are trained in ETS's standards for quality and fairness (ETS, 2015). In addition, a statistical technique called differential item functioning (DIF) analysis is conducted for gender as a routine part of the preliminary item analysis carried out after each test administration to identify (and if need be, remove) any items that are gender biased. Test takers are also provided information about how to contact ETS directly if there are concerns about test content. Statement 2 above has received empirical support as well. For example, Yoo and Manna (2017) examined the factorial invariance of a correlated two-factor model for TOEIC L&R across five groups (i.e., test-taker background characteristics): gender, age, employment status, time spent studying English, and time having lived in a country where English is the main language. Their analysis found that strict measurement invariance and structural invariance was upheld across all subgroups; thus, the underlying structure of the construct was the same across subgroups. In a subsequent study, Yoo et al. (2019) used score equity assessment (Dorans, 2004) to examine the extent to which subgroup membership (gender, age, educational background, language exposure, previous experience with the assessment) influenced the outcomes of the statistical and psychometric methods used in producing test scores; results provide evidence in support of the comparability and meaning of test scores across the subgroups studied.

Stakeholders who may question whether TOEIC L&R test score interpretations are impartial for students versus employees—due to potential differences in background knowledge of workplace contexts and topics—may desire to see more evidence to support Statements 1 and 2. Previous research on the role of background knowledge in L2 listening and reading comprehension suggests that background knowledge can play a role in performance. As a group, full-time employees tend to score higher on the TOEIC L&R test than full-time students (e.g., ETS, 2023, p. 8). These mean score differences would not be a problem if there are genuine differences in proficiency between the groups. However, if full-time students and full-time employees of equal ability did not have equal probability of getting the same score on the test (presumably, due to background knowledge), that would undermine claims about fairness and may require reconsidering claims about the extent to which the test is measuring background knowledge.

To further investigate these claims about fairness in the TOEIC L&R validity argument, we posed the following research questions:

1. Do TOEIC L&R test items show evidence of bias based on occupational status (i.e., full-time employed vs. full-time students)?
2. Is there any impact of employment-status-related bias on TOEIC L&R test scores?

Method

Materials

The TOEIC L&R test is administered as a paper- or computer-based test and consists of two separately timed sections, listening comprehension and reading comprehension, with 100 items in each section. The listening section takes 45 minutes to complete and is paced by an audio recording. The reading section is self-paced and takes up to 75 minutes to complete. An overview of the test's content, administrative procedures, scoring and score interpretations, measurement quality, and intended uses can be found in the TOEIC L&R test *Examinee Handbook* (ETS, 2022a) and *Score User Guide* (ETS, 2022b).

The analyses performed for this study used content and test taker data from nine previously administered TOEIC L&R test forms. The first five forms, administered in the same year, were designated as A1 to A5 for this study. The remaining four forms, administered in a different year, were designated as B1 to B4 for this study. In forms A1 to A5, no items were excluded from operational scoring. In forms B1 to B4, four items were excluded from operational scoring because these items did not function as intended as detected by quality control procedures. In the operational administration, forms A1 through A3 were administered in the same morning session; forms A4 and A5 were administered in an afternoon session. Likewise, forms B1 and B2 were both administered in a morning session; forms B3 and B4 were administered in the afternoon. The reliability—the extent to which the scores remain consistent over repeated administrations of the same or alternated forms—of the listening and reading sections in the test forms was estimated based on the internal consistency method KR-20 reliability (Kuder & Richardson, 1937). The empirical reliability estimates for the listening and reading sections in all forms were sufficiently high and ranged from 0.91 to 0.94.

The TOEIC background questionnaire (BQ) was used to identify test takers' demographic characteristics. The BQ is a survey that gathers information about TOEIC test takers' educational background, work experience, English language use and study, and the TOEIC test-taking experience. The BQ was administered to test takers during the administration of the TOEIC L&R test, prior to beginning the Listening test section.

Participants

The number of test takers varied between 10,000 and 20,000 on each of the nine TOEIC L&R test forms. The sample sizes for the listening and reading sections are equal because all test takers completed both listening and reading sections. The demographic characteristics of test takers (gender, age, occupational status) are presented in Table 1 for forms A1 through A5 and forms B1 to B4.

The total sample of test takers for each form was partitioned into subsamples based on test takers' occupational status. In the TOEIC BQ, Question 3 asks, "Which of the following best describes your current status?" (ETS, 2022a, p. 22). In this question, Option 1 is "I am employed full-time (including self-employed)," Option 2 is "I am employed part-time and/or study part-time," Option 3 is "I am not employed," and Option 4 is "I am a full-time student." Test takers who self-reported either Option 1 or Option 4 were included in the analysis and subsequently identified as full-time employee (reference group) or full-time student (focal group); test takers who self-reported Option 2 or Option 3 were excluded from the DIF analysis by occupational status. For a separate DIF analysis by age, which included all test takers regardless

Table 1 Compositions of Test Takers in Forms

		Gender		Age		Occupational status[a]	
Session time	Form	Female	Male	Under 22	At least 22	Full-time student	Full-time employee
Morning	A1	36%	64%	27%	73%	39%	53%
	A2	43%	57%	29%	71%	43%	47%
	A3	38%	62%	27%	73%	40%	51%
Afternoon	A4	40%	60%	34%	66%	51%	38%
	A5	42%	58%	34%	66%	51%	38%
Morning	B1	40%	60%	27%	73%	38%	53%
	B2	40%	60%	25%	75%	35%	55%
Afternoon	B3	43%	57%	35%	65%	54%	35%
	B4	43%	57%	35%	65%	52%	37%

[a] Percentages for occupational status include only full-time students and employees and do not sum to 100%.

of employment status, individuals under age 22 and test takers who were 22 or older were classified into two age groups. The under 22 age group was considered as the focal group, and the 22 or older group was the reference group.

As shown in Table 1, test takers within the same session had more similar compositions in terms of gender, age, and occupational status than test takers who took the test in separate sessions within the same administration. In general, the morning sessions (i.e., A1 to A3; B1 and B2) had more full-time employees and test takers who were at least 22 years old, compared with the afternoon sessions (A4 and A5; B3 and B4). In addition, both the morning and afternoon sessions had larger percentages of male test takers than female test takers and had more test takers who were at least 22 years old. These patterns are consistent across these two administrations (A and B) and they are common in TOEIC test administrations.

Procedures and Data Analysis

To investigate whether the TOEIC L&R test unfairly advantages employed test takers, this study included analyses conducted at the item level (i.e., for item-level performance) and at the test score level (i.e., for test section performance). At the item level, DIF analysis was used to evaluate a differential performance of test items across subgroups with the same English proficiency, as well as content reviews by a DIF panel of content subject matter experts. At the test score level, we evaluated score comparability across subgroups of full-time employed and full-time student test takers. This psychometric approach (described in the following sections that describe Part 2) compares test scale scores derived for each subgroup with those from the total group.

To supplement the analyses at the item and test score levels, we coordinated independent content reviews of a TOEIC L&R test form to gain additional insight into how workplace-oriented content (e.g., settings, topics, vocabulary) is incorporated into test items. This content analysis was conceived of and conducted after Parts 1 and 2 of the study were already completed. The content analysis was intended to complement the statistical analyses performed in Parts 1 and 2 with illustrations of "typical" items that have workplace content but do not show evidence of bias.

Part 1: Item-Level DIF Analysis

DIF analysis evaluates the extent to which the psychometric characteristics of each individual item varies between the focal and reference groups and thus provides statistical evidence for whether items favor the focal group or not. The group of interest in DIF analysis is referred to as the focal group, whereas the group to be compared is referred to as the reference group. Using the Mantel-Haenszel delta difference (MH D-DIF) statistics and its statistical significance, the DIF procedure as used at ETS classifies test items into three categories of A, B, or C, in which the severity of DIF increases from A to C (Zwick, 2012). According to the classification described by Zieky (1993, p. 342), items classified as A have negligible or nonsignificant DIF; items classified as B have slight to moderate DIF, and items classified as C are moderate to large DIF. Classification levels, particularly levels B and C, are further distinguished by the direction of the signs. Items classified or flagged as B+ and C+ indicate that the items favor the focal group. Items flagged as B- and C- indicate that the items favor the reference group. In the current study, all the DIF analyses were conducted in statistical analysis software (SAS) by adapting the SAS macro programs for MH D-DIF analysis from Hao (2013), with full-time students as the focal group and full-time employees as the reference group. The DIF analyses were completed by the second, third, and fourth authors of this study.

Items that were classified as C level[1] (either C- or C+) items by the analysis were reviewed by a DIF panel (Zieky, 2016). The DIF panel consisted of a group of five assessment development experts at ETS. The panel examined each flagged item individually to judge whether an item exhibited evidence of bias for the focal or reference group. If an item is judged to exhibit clear evidence of bias, the item should be removed from scoring to minimize bias in test scores and promote the principle of fairness. The panel was observed by the first author of this study.

Part 2: Test-Level Analysis

Evidence of bias at the item level is concerning, but bias in overall test scores would raise more severe concerns about overall fairness and appropriate use of the test for specific subgroups. DIF analysis is the statistical approach typically used to identify individual items that may be biased, but DIF analysis does not quantify bias at the level of the test score. Score equity assessment (SEA) is a method used to evaluate the equity of test scores and to address construct validity

from a fairness perspective (Dorans, 2004). The central idea of SEA is to assess if the test score measures the same construct across subgroups as compared with the total group through the subpopulation invariance requirement of equating (Dorans, 2004). In other words, if subgroup membership has no impact on test (scaled) scores, test scores derived from the subgroup should be comparable with test scores derived from the total group regardless of test takers' background. More specifically, equating functions derived from different subgroups should produce only negligible differences if the test measures the same construct across all the subgroups.

The general procedure for the equating analysis is based on item response theory (IRT) methodology (Lord & Novick, 1968). The specific IRT model used in this study is the two-parameter logistic (2PL) IRT model, which specifies item discrimination and item difficulty parameters in addition to ability. The IRT equating procedure included three steps. The initial step was to estimate the IRT item parameters for the subgroups of full-time students and full-time employees, through the multiple-group IRT calibration, using the R mirt package (Chalmers, 2012). The second step was to transform the IRT item parameter estimates to put them on the same scale of the reference forms, which serves as the benchmark to equate the test scores. The transformation technique is the Stocking-Lord method (Stocking & Lord, 1983). After the transformation, IRT true-score equating (Kolen & Brennan, 2014) was conducted to equate the new forms separately for each subgroup. The same item calibration, scale transformation, and equating procedures that were conducted on both subgroups, respectively, were also applied to the total groups for the forms analyzed. Thus, in this study to evaluate the equating comparisons between the total group and subgroups, the equating relationship was obtained using three different groups of test takers: the full-time student subgroup, the full-time employee subgroup, and the total group.

The discrepancies based on the two subgroups of test takers were compared to the conversion based on the total group in each form. The discrepancies were evaluated by difference that matters (DTM), as proposed by Dorans et al. (1994). The DTM was defined as a half of the score unit. Since the score unit of TOEIC L&R section scores is 5 points, the DTM was set at 2.5 points for the unrounded scale scores. Any difference that is less than this criterion was deemed to be negligible. The DTM of 2.5 points is well below the standard error of measurement of 25 points for each TOEIC L&R test section (see ETS, 2022b, p. 19). All score-level analyses were completed by the second, third, and fourth authors of this study.

Part 3: Content Review

A content review was conducted to identity examples of test items that incorporate workplace-oriented content (e.g., workplace settings and workplace-oriented language) and evaluate the extent to which item content could be expected to provide an unfair advantage for employed or student test takers. Since this review would result in the disclosure of item content in this publication, we identified another previously administered TOEIC L&R test form that was similar in its characteristics to the test forms used in Parts 1 and 2. Prior to the content analysis, we performed DIF analysis using the same procedures described in Part 1. Any items flagged for C level DIF were reviewed by a DIF panel.

The content review was conducted by two groups of experts in two phases. In the first phase, two TOEIC L&R assessment development content leads reviewed the test form (stimulus material and items) to identify exemplar TOEIC L&R test items that contain workplace-oriented content. For each item judged to be an exemplar, the content leads commented on how workplace content is incorporated (e.g., the nature of the setting, the topic[s], specific vocabulary used or implied) and whether this content could be expected to unfairly advantage employed (vs. student) test takers. In the second phase of the content analysis, all the items identified in the first phase were examined and discussed by an external panel. Since this subset of items from the TOEIC L&R test form had already been identified as situated in workplace contexts, the external panel focused on the issue of whether each item had content that could be expected to unfairly advantage employed (or student) test takers. This panel consisted of three employees of the Institute for International Business Communication (IIBC), the TOEIC program's partner in Japan. The participating IIBC employees were familiar with TOEIC L&R test content, proficient in English, and familiar with the target groups of test takers (full-time students, full-time employees).

Results

Part 1: Item-Level DIF Analysis

Tables 2 and 3 present the DIF flag results, along with the scale score means and standard deviations of subgroups, for both the listening and reading sections in the nine forms analyzed for the DIF analysis. Each section has 100 items for analysis.

Table 2 Occupational Status DIF Flag Results for Listening in Forms A1 to A5, and B1 to B4 (Number of Flags)

	Focal group (students)		Reference group (employed)		DIF flag results (listening)					
Form	*M*	*SD*	*M*	*SD*	A+	A−	B+	B−	C+	C−
A1	315	78	340	83	50	45		2	1	2
A2	309	78	342	84	45	50	3	1		1
A3	314	78	337	82	45	46	3	3		3
A4	308	79	342	84	51	43	4	1		1
A5	310	78	338	85	49	44	3	3		1
B1	322	78	342	83	43	52	2	2	1	
B2	320	81	345	84	48	46	3	2		1
B3	315	80	339	86	51	45	1	3		
B4	314	79	345	84	42	54	4			

Note. DIF = differential item functioning. DIF flags with (+) favor the focal group (full-time students) and DIF flags with (−) favor the reference group (full-time employees).

Table 3 Occupational Status DIF Flag Results for Reading in Forms A1 to A5, and B1 to B4 (Number of Flags)

	Focal group (students)		Reference group (employed)		DIF flag results (reading)					
Form	*M*	*SD*	*M*	*SD*	A+	A−	B+	B−	C+	C−
A1	261	86	285	92	49	41	3	5		2
A2	260	87	286	91	53	39	2	5		1
A3	264	86	286	91	50	38	3	7	1	1
A4	258	89	291	95	47	51		2		
A5	256	85	284	93	54	43		2		1
B1	270	90	290	96	51	39	4	6		
B2	266	92	291	98	53	41	1	4		1
B3	261	90	284	97	54	46				
B4	265	92	294	99	52	47		1		

Note. DIF = differential item functioning. DIF flags with (+) favor the focal group (full-time students) and DIF flags with (−) favor the reference group (full-time employees).

The numbers of flags presented in the tables are the numbers of items flagged by different DIF levels. Although four items were excluded from operational scoring, they were included in the DIF analyses to investigate DIF by occupational status for all the items that were administered. Excluding these four items from the DIF analysis produced essentially the same results in this study.

Across all test forms and test sections, the full-time employee group had higher scale score means than the full-time student group. Although not presented in this paper, the full0time employee group also had higher raw score means than the full-time study group, and these group mean differences are statistically significant ($p < .0001$), based on independent group Welch's t-tests for the full-time student and full-time employee groups. The Welch's t-tests were used because the assumption of the homogenous variances is not held. To better understand the comparison of two groups' means, we also calculated the effect size of the mean differences to evaluate the magnitude of the differences. Hedges' *g* statistic computes effect size by taking account of unequal sample sizes (Hedges, 1981). The effect sizes computed by Hedges' *g* for raw scores in the listening and reading sections by the subgroups of full-time students and full-time employees ranged from 0.2 to 0.4, indicating relatively small effect sizes of the differences between these two subgroups (Cohen, 1992).

The total number of items flagged as C level DIF is small. Overall, for each test section and form, zero to three items were flagged for C level DIF associated with occupational status. In total, 18 items (1.0%) were flagged for C level DIF. More listening items were flagged than reading items (11 vs. seven, or 1.2% vs. 0.8%). More of the items flagged as C level DIF appear to favor full-time employees (C-) than full-time students (C+).

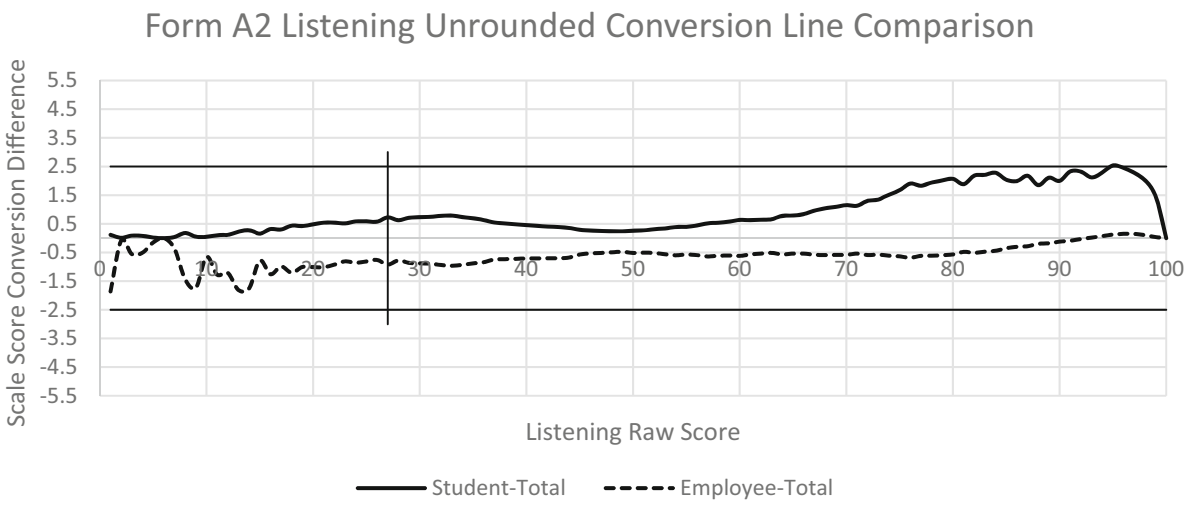


Figure 1 Conversion Comparisons for the Total and Subgroups in the Listening Section in Form A2

All C level DIF items were assigned for content review by the DIF panel to further investigate whether items had content that could be deemed to favor full-time employees or students. The panel concluded that nine of 18 items demonstrated a potential bias against the focal group of full-time students, noting the use of workplace-oriented language that could be less familiar to students (e.g., "colleague"). Five items did not show evidence of a discernable or potential bias in favor or against either group. Ultimately, the panel believed that none of the items should be clearly identified as biased; items with possible bias were perceived to include content that was construct-relevant. The panel noted that age may be a hidden factor driving bias in some cases, as two of the items were perceived to demonstrate a potential advantage for older test takers (versus young test takers). For example, in their review of one of these items, the panel noted that age could be an underlying factor if younger test takers do not know "pharmacy" (the word and the place).

It is worth noting that subgroups based on occupational status and age are not entirely distinct. As the usual age for college graduation in Japan is 22, most test takers under 22 years of age were identified as full-time students and most full-time employees were at least 22 years old. However, approximately 20% to 40% of test takers who were at least 22 years old identified themselves as full-time students. Conversely, less than 1% of full-time employees were under 22 years old. It appeared to be reasonable to explore the possible impact of age on the analyses. Therefore, a supplementary DIF analysis based on age was also conducted. The results of this analysis are reported in Appendix A. The age-based DIF analysis produced a small number of items flagged as C level DIF (four listening items, or 0.4% of the listening items, and three reading items, or 0.3% of the reading items were flagged). All items flagged by the age-based DIF analysis were also previously flagged by the DIF analysis for occupational status, although not all items flagged with status DIF were detected by age DIF. Because classification by age and occupational status is not mutually exclusive, item performance for different subgroups were expected to be intertwined by occupational and age-related group characteristics.

Part 2: Test-Level Analysis

Among the five forms that did not have items excluded from operational scoring, forms A2 and A5 (from the morning and afternoon sections, respectively) were evaluated in terms of the scale score comparability between the subgroups. In the other forms, two non-DIF items did not function well in the subgroups after IRT calibration. Therefore, only forms A2 and A5 were used to illustrate the impact of the DIF items at the test level. Figure 1 depicts the differences between the unrounded conversions for the full-time student group and the total group, and it also shows the differences between the unrounded conversions for the full-time employee group and the total group in the listening section in form A2. Figure 2 presents the comparison results for the unrounded conversion in the reading section in form A2. Figures 3 and 4 present the comparison results for the unrounded conversion in the listening section and reading section of form A5.

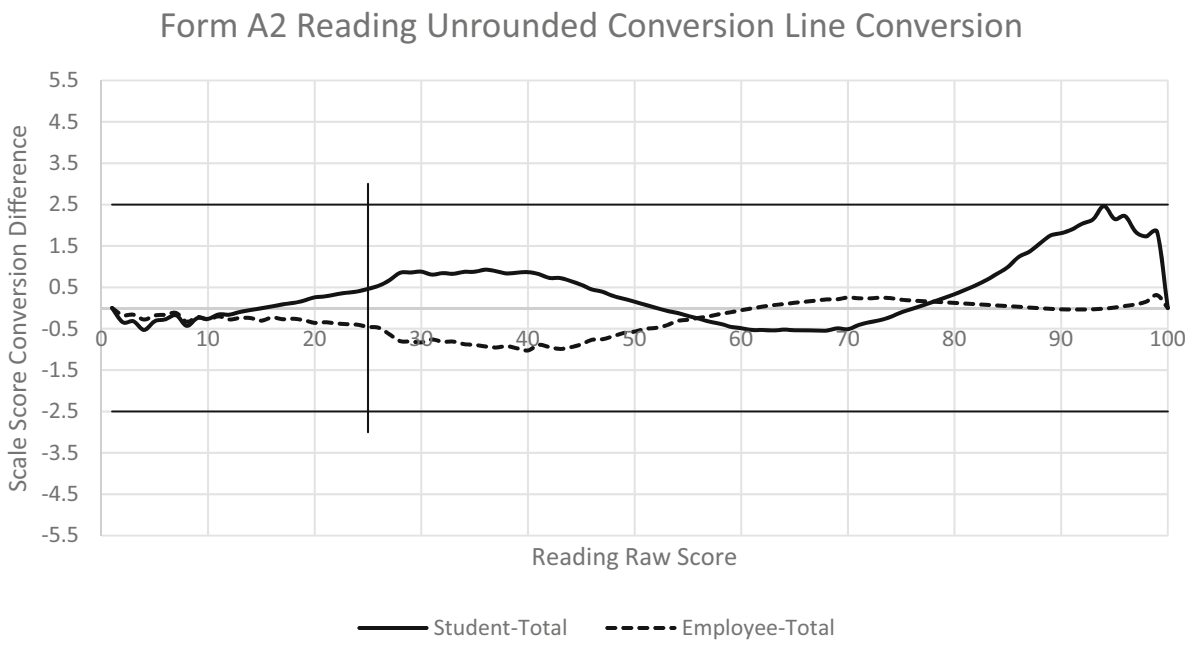


Figure 2 Conversion Comparisons for the Total and Subgroups in the Reading Section in Form A2

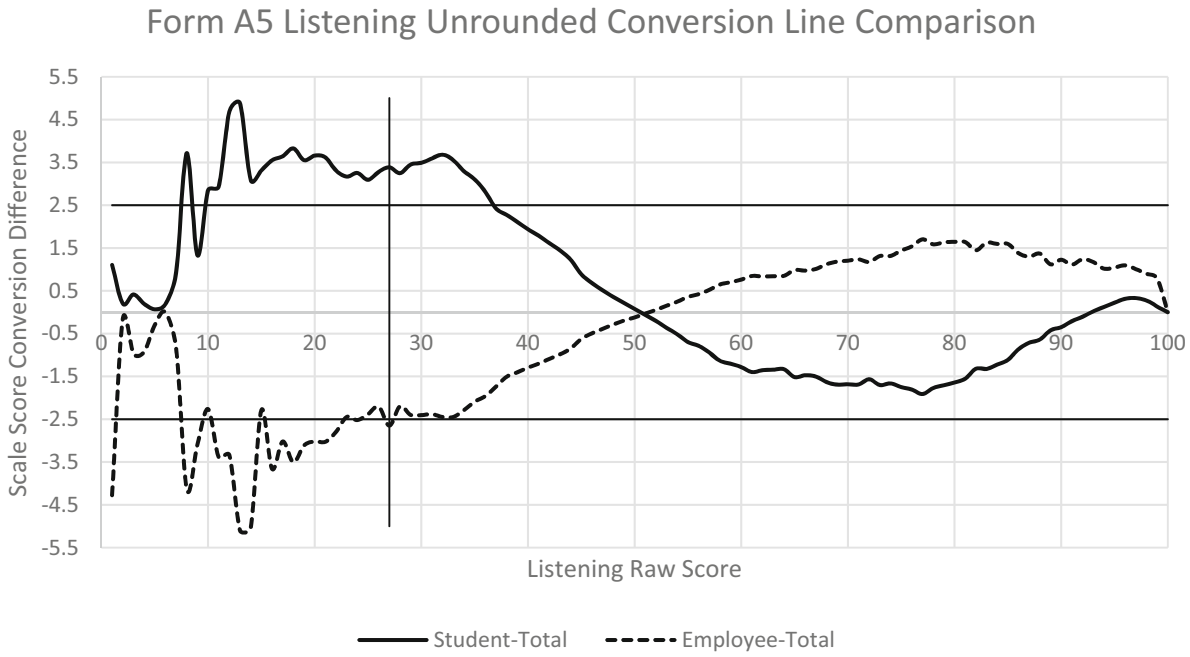


Figure 3 Conversion Comparisons for the Total and Subgroups in the Listening Section in Form A5

As shown in Figures 1 through 4, the horizontal axis shows the raw test scores for listening or reading, from 0 to 100. The vertical axis represents the scale score differences computed between the subgroup and total group. In the figures, 2.5 and -2.5 indicate the DTM can be positive and negative. The chance level of performance (i.e., guessing at random [27 for the listening section, 25 for the reading section]) is indicated in the plots by a vertical line. The two fluctuating lines depict the scale score differences computed from the total group to the subgroups of the full-time students and full-time employees, respectively, along the scale of raw scores. The figures visualize comparisons by displaying the differences of the scale test scores derived from each subgroup and from the total group. Large differences may indicate that scale scores

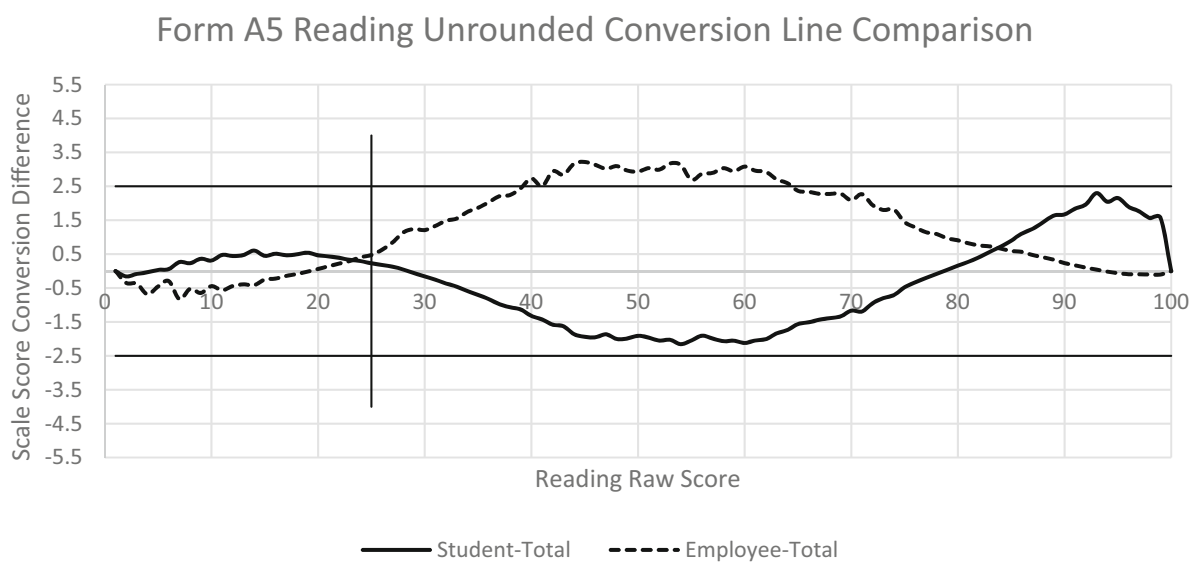


Figure 4 Conversion Comparisons for the Total and Subgroups in the Reading Section in Form A5

are not that comparable across the subgroups with respect to the total group. All plots show minor discrepancies between conversions derived from the subgroups in comparison with that from the total group. The degree of discrepancies varies among sections and forms and no specific patterns exist for the comparison differences of conversions.

The results show that conversion differences between the subgroups and total group are within the DTM for all forms on most raw points above the chance level. The listening section of Form A5 produced differences above the DTM on the raw score range from 25 to 36 for the subgroup of full-time students. The reading section of Form A5 produced differences above the DTM on the raw score range from 40 to 64 for the subgroup of full-time employees. Although these differences are beyond the DTM, their discrepancies are not large.

In summary, the score-level analyses do not indicate that the conversions derived from one subgroup are systematically higher or lower than the conversions from the other subgroup. Similarly, differences between the total group conversion and the subgroups' conversions showed negligible differences across most of the raw score range for both listening and reading. Conversion comparison indicates that the current equating practice for TOEIC L&R tests is solid in deriving conversions from the total group. Although the separate conversions from the subgroups of the full-time student and full-time employees may capture the unique characteristics of the subgroups to some extent, the differences due to the equating based on subgroups have no significant impacts from the perspective of score reporting.

Part 3: Content Review

The DIF analysis performed on the TOEIC L&R test form used for the content review flagged two items for C level DIF (one listening item, one reading item). A DIF panel reviewed the two flagged items and concluded that the items do not contain content that would provide an obvious unfair advantage for full-time students or full-time employees. Consequently, all test content and items were retained for the first phase of the content review.

ETS content reviewers identified exemplar items with workplace-oriented content (settings, vocabulary) across all parts of the TOEIC L&R test. For the Listening test section, this included three Photograph questions, one Question-Response question, one Conversation set (with one question), and two Talk sets (containing six questions in total). For the Reading test section, this included three Incomplete Sentence items, two Text Completion sets (containing eight questions in total), and two Single Passage sets (containing four questions in total). The settings of the reviewed Listening and Reading test questions and sets included general office (e.g., meetings, calls, videoconferencing, orientations), nonoffice workplaces (e.g., warehouse, industrial settings), and public spaces (e.g., museum, government office, customer-oriented webpages).

There was general agreement across both content review panels that all test questions could be answered correctly without any specialized background knowledge. In addition, both panels concluded that full-time employees should not

have an unfair advantage over full-time students based on the employees' (presumably) greater familiarity with workplace settings.

To illustrate how workplace-oriented content is incorporated into test items, we have reproduced three Listening test section and three Reading test section items (or sets) from the review, with commentary from the panels. The format of these items has been slightly altered for publication purposes, omitting directions and instructions (for a sample TOEIC L&R test, please see ETS, 2019).

Listening Test Section

Photograph Item Example. In Photograph items, test takers hear four statements about a picture, which are not printed. Test takers must select one statement that best describes the picture by marking one option (A through D) on an answer sheet (paper-based) or by selecting one option on a computer screen (computer-based). Photograph items are designed to contribute evidence of test taker's ability to (a) understand the gist or central idea or (b) understand obvious details.

Test takers see:

Test takers hear:
(Woman, British voice):

(A) One of the women is throwing away some f les.
(B) **One of the women is holding a piece of paper.**
(C) The man is taking off his glasses.
(D) T he man is typing on a laptop.

This listening comprehension item is designed to measure a test taker's ability to understand obvious details. The content reviewers noted that this is an example of an item situated in an office setting. The spoken language in the response options provide possible explanations of what the people in the photo are doing, using language that is common

(non-technical) and applicable across many settings (e.g., "files," "paper," "laptop"). The item does not require workplace experience to choose the correct answer (Option B, bolded) and should not advantage employed test takers over students.

Question-Response Item Example. In Question-Response items, test takers hear a question or statement followed by three potential responses spoken in English. The question and response involve two speakers. Neither the question (or statement) nor responses are printed. Test takers must select the best response to the question (or statement) by marking the appropriate option. These items are designed to provide evidence of the test taker's ability to (a) understand gist, purpose, and basic context in short spoken texts; (b) understand details in short spoken texts; or (c) understand implied meaning in short spoken texts.

Test takers hear:

(Man, Australian voice): Aren't we having a videoconference with Ms. Lobo at ten?
(Woman, American voice):

(A) I have one more than I need.
(B) **It's not on my calendar.**
(C) Those microphones over there.

This item is designed to measure a test taker's ability to understand implied meaning in short spoken texts (i.e., pragmatic knowledge). The content reviewers noted that the topical context of this item is meetings and schedules and is a common context (along with calls, videoconferences with clients, and timelines) in the listening section. This type of item is intended to be more challenging as it requires understanding the inference in the correct response (Option B, bolded). In this case, the most appropriate response implies that the woman does not have any information about the meeting; consequently, she is unable to respond clearly to the question. Japanese content reviewers speculated that L1 background could potentially have an influence on the difficulty of this item, as Japanese has different verbs for "having something" (i.e., possession) versus "having a meeting" (i.e., existence or occurrence of an event). Regardless, both groups of content reviewers noted that the tasks and concepts referenced in this item should not require any work experience to understand.

Talk Item Example. In Talk question sets, test takers listen to talks given by a single speaker. After hearing each talk, test takers answer three questions about what the speaker says. Each question is spoken aloud and printed in the test book (paper-based) or shown on the computer screen (computer-based). Test takers select the best response option. These questions evaluate the test taker's ability to (a) infer gist (main idea, context), (b) understand details, or (c) understand a speaker's purpose or implied meaning in a phrase or sentence.

Test takers hear:

(Woman, British voice): Welcome to the new-employee orientation here at Parton Manufacturing. Before we begin, let me point out that we have lockers in the break room. A locker is available for each factory worker, so feel free to keep your jackets or lunches there. OK, first up, paperwork. You'll find company safety policies inside your employee packet. Please sign them, and turn them in at the end of the meeting. There's a basket on the table in the back of the room. Let me know if you have any questions.

(Narrator): Where does the talk most likely take place?

(A) **At a manufacturing plant**
(B) At a medical facility
(C) At a hardware store
(D) At an employment agency

(Narrator): According to the speaker, what is available to the listeners?

(A) Identification badges
(B) Dining facilities
(C) **Personal lockers**
(D) Parking permits

(Narrator): What does the speaker mean when she says, (Woman, British voice: "There's a basket on the table in the back of the room?")

(A) Someone forgot to take a basket home.
(B) **Forms should be put in the basket.**
(C) A room has not been cleaned yet.
(D) Some snacks are now available.

This talk is situated in the context of a new employee orientation, an office context that should also be easily familiar to students who have likely experienced similar situations (e.g., on the first day of school). The vocabulary used in the talk and in the questions is nontechnical (e.g., "lockers," "break room," "factory," "lunches"). The questions are typical and require understanding the gist or likely context for the talk (a manufacturing plant, Option A), a detail brought up in the talk (availability of personal lockers, Option C), and the pragmatic implication of a phrase used by the woman during the talk (Option B). Both groups of content reviewers noted that the general context and vocabulary should be accessible to both students and full-time employees.

Reading Test Section

Incomplete Sentence Item Example. In Incomplete Sentence items, test takers are presented with a sentence that has a missing word or phrase. Test takers must then select the word or phrase, from among four options, that best completes the sentence. These questions test either grammar or vocabulary at the sentence level.

Test takers read:

In July, the Martine Electronics Company achieved its – – – – monthly sales figures ever.

(A) high
(B) higher
(C) highly
(D) **highest**

This question is testing a grammar point (adjectives). The content reviewers noted that the topical aspect of this question is business-oriented and includes vocabulary that could potentially be less familiar to students (e.g., "sales figures"), but knowledge of this vocabulary is not necessary to answer the item correctly. Quite simply, test takers do not need to understand the business aspect of the item to answer it correctly.

Text Completion Item Example. In Text Completion sets, test takers read short texts in a variety of formats. Each short text is missing four elements such as words, phrases, or key sentences. Test takers must correctly identify each missing element by selecting the appropriate word, phrase, or sentence from among four options. These items test grammar, vocabulary, and the ability to connect information within a short text.

Test takers read:

To: Undisclosed Recipients
From: Drucker and Lowe Accounting
Date: January 3
Subject: New Office Space

Dear Clients:

We are pleased to announce that Drucker and Lowe Accounting has ------- **135.** . Our new facilities in the Lambert Office Building offer a more spacious reception area and additional consultation rooms. ------- **136.** .

Our street address remains the same, but our suite is now on the fifth floor. You will see the entrance ------- **137.** in front of you when you step out of the elevator. Our phone number and e-mail address will remain unchanged.

We would like to use this ------- **138.** to express our appreciation to all of our clients. We hope to see you soon.

Sincerely,

Christine Drucker and Ron Lowe

135. (A) relocated
(B) consolidated
(C) promoted
(D) registered

136. (A) We will review the lease and notify you of any necessary revisions.
(B) This transition exceeded our budget by 20 percent.
(C) These changes will enable our staff to serve you more effectively.
(D) After January 30, we will be taking a leave of absence.

137. (A) directs
(B) directed
(C) directly
(D) directness

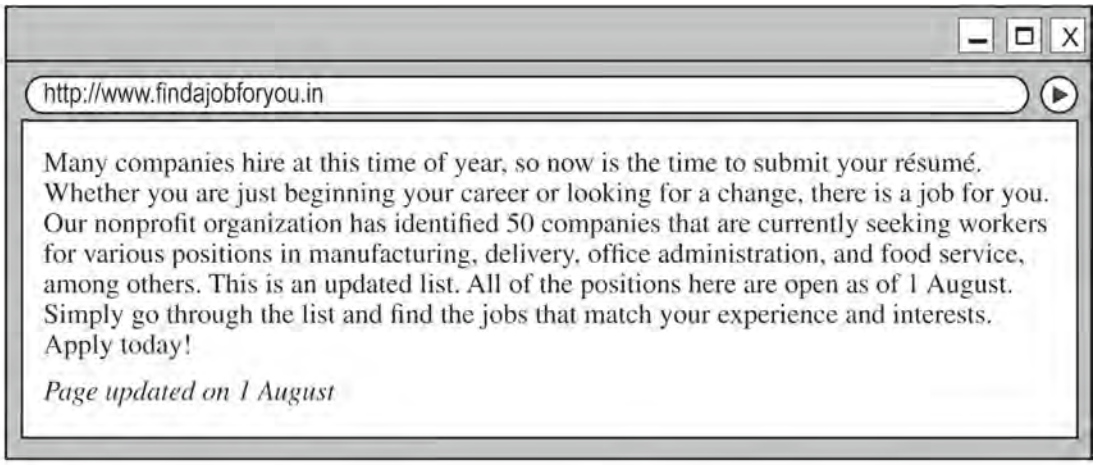
138. (A) opportunity
(B) approach
(C) ability
(D) event

Key: 135 = A; 136 = C; 137 = C; 138 = A.

This set includes a customer-focused text, situated in the context of customer-oriented communication for an accounting business. Content reviewers noted that references to "reception area," "consultation rooms," and "clients" are not technical and should be familiar in any workplace context. The questions test vocabulary and grammar points using non-technical language. Thus, although the item is situated in a business services context, content reviewers concluded that the text should be broadly accessible to different groups of test takers and neither the text nor the questions should require (or provide an advantage to those with) business knowledge.

Single Passage Item Example. Single Passage sets appear in the Reading Comprehension Passages part of the Reading test section and require test takers to read everyday and workplace-oriented texts. The items accompanying each text may require the test taker to identify the main idea, identify stated details, infer implied meanings (e.g., the context, the writer's purpose), or connect information within or across texts.

Test takers read:



149. What is the purpose of the Web page?
(A) To promote a new product
(B) To advertise a new company
(C) To help people looking for a job
(D) To explain new policies to employees

150. What has recently been revised?
(A) A delivery route
(B) A list of companies
(C) A résumé requirement
(D) A set of hiring guidelines

Key: 149 = C; 150 = B.

This reading text is situated in the context of an employment website, a workplace-oriented context that should be broadly familiar to adults and young adults. The content reviewers noted that the text contains language that is nontechnical and should be familiar to anyone looking for employees or employment. The genre of the job advertisement should be accessible to both students and full-time employees, and references to "hire," "submit your resume," and "nonprofit organization" are broadly applicable across workplace contexts.

Discussion

The construct definitions of more general-purposes language assessments often result in test designs where background or topical knowledge is treated as construct-irrelevant variance. In this study, we focused on how construct definition and its operationalization in the TOEIC L&R test led to specific claims about the impartiality of score interpretations based on background knowledge of the TLU domain: the international workplace. We conducted analyses at the item and total score level to investigate whether those with more knowledge of and experience with the workplace (full-time employees) had an unfair advantage over those with less experience in the workplace (full-time students). The results of the item-level DIF analysis indicated that very few items (1%) were flagged for statistical DIF, and none of the items flagged were judged to clearly exhibit DIF by a bias review panel.

The findings of the score-level analysis provide additional statistical support for the conclusions drawn from the DIF analysis for occupational status. Even though a very small proportion of items were flagged for DIF, the TOEIC L&R tests functioned consistently overall across the subgroups of full-time students and full-time employees within the framework of testing English proficiency in daily life and general workplace contexts. The equating practice further safeguards the psychometric practice in terms of test fairness for subgroups with different membership and characteristics. In future studies, additional equating analyses under the SEA framework could be conducted using more test forms.

The example test questions illustrate how workplace contexts and topical content are typically represented in the TOEIC L&R test without requiring (or providing an advantage for) specialized workplace knowledge or experience. The workplace contexts and topics included in the test are designed to be broadly accessible and familiar to test takers of varying

backgrounds, and when a more idiosyncratic context or topic is included, a test taker's linguistic knowledge is expected to play a much more important role in the measurement process. This observation is consistent with previous research on the role of background knowledge in more general-purposes language testing (e.g., Hill & Liu, 2012).

Thus, the results of this study contribute empirical support for claims about the impartiality of TOEIC L&R test scores with respect to test takers' background knowledge (i.e., students vs. employees). These results complement prior empirical research that has provided evidence of the impartiality of TOEIC L&R test scores with respect to gender.

In some sense, it may be surprising that a language test that is contextualized in a TLU domain—regardless of how broadly-defined it is—does not measure proficiency in a manner that benefits domain insiders (employees) over domain outsiders (students). Although this study did not find any evidence for a performance advantage for employees, it is possible that the subgroups in this study experience the test in slightly different ways (e.g., with more or less comfort and confidence). Although most studies of the effect of background knowledge on L2 reading and listening test performance focus on performance, those that have focused on the test taker experience have found some evidence that those with more background knowledge report more comfort and confidence (e.g., Lee, 2011).

One limitation of this study was how background knowledge was conceptualized. In our study, background knowledge was conceptualized as general knowledge of a more generalized workplace TLU domain, including knowledge of and familiarity with associated communicative tasks, contexts, and topics. This corresponds to a definition of background knowledge as more domain-general knowledge (or the totality of what an individual knows), and can be differentiated from more domain-specific knowledge such as knowledge of a particular topic (Banerjee, 2019). Our study focused on the broadest connotation of background knowledge, the type of domain-general knowledge that may be acquired through experience by domain-insiders (in our study, full-time employees). Although prior research has shown that domain-general knowledge can predict domain-general reading comprehension (Cai & Kunnan, 2018), perhaps a more general-purposes domain can be defined so broadly that this advantage disappears.

When considering the issue of whether more background knowledge provides an advantage on a carefully designed general-purposes L2 reading or listening assessment, what one means by "advantage" may matter. In this study, we focused on whether students or employees had a performance advantage on individual items and in overall test scores. One of the content review panels noted that for some test questions, workplace-oriented content may lead to students having to take longer to comprehend a text (or answer a question correctly) due to less familiarity with workplace contexts or vocabulary. This study is unable to address that possibility: a potential difference in the overall test taking experience, wherein familiarity impacts comfort or anxiety but does not necessarily translate into a performance advantage or disadvantage. Rather, this study allows us to draw general conclusions about the overall impact workplace experience (and relevant background knowledge) may have on performance. Future research in this area would be wise to include a focus on test taker experiences and perceptions, not simply test taker performance.

Note

1 MH D-DIF absolute value is significantly greater than 1 and its absolute value is at least 1.5

References

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

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford. https://doi.org/10.1016/B978-0-08-044894-7.00263-3

Banerjee, H. L. (2019). Investigating the construct of topical knowledge in a scenario-based assessment designed to simulate real-life second language use. *Language Assessment Quarterly*, *16*(2), pp. 133–160. https://doi.org/10.1080/15434303.2019.1628237

Cai, Y. (2013). *Modeling ESP ability in reading: A focus on interaction among grammatical knowledge, background knowledge and strategic competence* [Unpublished doctoral dissertation]. The University of Hong Kong.

Cai, Y., & Kunnan, A. J. (2018). Examining the inseparability of content knowledge from LSP reading ability: An approach combining bifactor-multidimensional item response theory and structural equation modeling. *Language Assessment Quarterly*, *15*(2), 109–129. https://doi.org/10.1080/15434303.2018.1451532

Camilli, G. (2006). Test fairness. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 221–256). American Council on Education and Praeger.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1 – 29. https://doi.org/10.18637/jss.v048.i06

Chung, T., & Berry, V. (2000). The influence of subject knowledge and second language proficiency on the reading comprehension of scientific and technical discourse. *Hong Kong Journal of Applied Linguistics*, *5*(1), 187 – 225.

Clapham, C. (1996). *Studies in Language Testing: Vol. 4. The development of IELTS: A study of the effect of background knowledge on reading comprehension*. Cambridge University Press.

Cohen, J. (1992) A power primer. *Psychological Bulletin*, *112*(1), 155 – 159. https://doi.org/10.1037//0033-2909.112.1.155

Dorans, N. J. (2004). Using subpopulation invariance to assess test score equity. *Journal of Educational Measurement*, *41*(1), 43 – 68. https://doi.org/10.1111/j.1745-3984.2004.tb01158.x

Dorans, N. J., Feigenbaum, M. D., Feryok, N. J., Lawrence, I. M., Schmitt, A. P., & Wright, N. K. (1994). *Technical issues related to the introduction of the new SAT and PSAT/NMSQT* (Research Memorandum RM-94-10). ETS.

Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732911

ETS. (2015). *ETS standards for quality and fairness*. https://www.ets.org/pdfs/about/standards-quality-fairness.pdf

ETS. (2019). *Sample Tests: TOEIC® Listening and Reading test*. https://www.ets.org/pdfs/toeic/toeic-listening-reading-sample-test.pdf

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

ETS. (2022b). *TOEIC® Listening and Reading test: Score user guide*.

ETS. (2023). *TOEIC® Listening & Reading test: 2022 report on test takers worldwide*. https://www.ets.org/pdfs/toeic/toeic-listening-reading-report-test-takers-worldwide.pdf

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. Cambridge University Press. https://doi.org/10.1017/CBO9781139150484

Grabe, W., & Yamashita, J. (2022). *Reading in a second language: Moving from theory to practice* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/9781108878944

Hale, G. A. (1988). Student major field and text content: Interactive effects on reading comprehension in the Test of English as a Foreign Language. *Language Testing*, *5*(1), 49 – 61. https://doi.org/10.1177/026553228800500104

Hao, S. (2013). Two SAS macros for differential item functioning analysis. *Applied Psychological Measurement*, *38*(1), 81 – 82. https://doi.org/10.1177/0146621613493164

Hedges, L. V. (1981), Distribution theory for Glass's estimator of effect size and related estimators, *Journal of Educational and Behavioral Statistics*, *6*(2), 107 – 128. https://doi.org/10.3102/10769986006002107

Hill, Y. Z., & Liu, O. L. (2012). *Is there any interaction between background knowledge and language proficiency that affects TOEFL iBT® Reading performance?* (TOEFL iBT Research Report No. 18). ETS. https://doi.org/10.1002/j.2333-8504.2012.tb02304.x

Jensen, C., & Hansen, C. (1995). The effect of prior knowledge on EAP listening-test performance. *Language Testing*, *12*(1), 99 – 119. https://doi.org/10.1177/026553229501200106

Karami, H., & Alavi, S. M. (2012). Examining background knowledge bias in a high stake general academic language test: A differential item functioning analysis. *English Language Assessment*, *7*, 25 – 41. http://kelta.kr/bbs/board.php?bo_table=articla&wr_id=34&page=12

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.

Knoch, U., & Macqueen, S. (2016). Language assessment for the workplace. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 291 – 308). De Gruyter Mouton. https://doi.org/10.1515/9781614513827-020

Kolen, M. J., & Brennan, R. J. (2014). *Test equating: Methods and practices* (3rd ed.). Springer.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*, 151 – 160. https://doi.org/10.1007/BF02288391

Kunnan, A. J. (2007). Test fairness, test bias, and DIF. *Language Assessment Quarterly*, *4*(2), 109 – 112. https://doi.org/10.1080/15434300701375865

Kunnan, A. J. (2018). *Evaluating language assessments*. Routledge. https://doi.org/10.4324/9780203803554

Lee, J.-Y. (2011). *Second language reading topic familiarity and test score: Test-taking strategies for multiple-choice comprehension questions* [Unpublished doctoral dissertation]. University of Iowa.

Lord, F. M., & Novick, M. R. (with Birnbaum, A.). (1968). *Statistical theories of mental test scores*. Information Age Publishing.

Ridgway, T. (1997). Thresholds of the background knowledge effect in foreign language reading. *Reading in a Foreign Language*, *11*(1), 151 – 168. https://nflrc.hawaii.edu/rfl/item/28

Schmidgall, J. E. (2017). *Articulating and evaluating validity arguments for the TOEIC® tests* (Research Report No. RR-17-51). ETS. https://doi.org/10.1002/ets2.12182

Schmidgall, J., Cid, J., Carter Grissom, E., & Li, L. (2021). *Making the case for the quality and use of a new language proficiency assessment: Validity argument for the redesigned TOEIC Bridge® tests* (Research Report No. RR-21-20). ETS. https://doi.org/10.1002/ets2.12335

Sireci, S. G., & Randall, J. (2021). Evolving notions of fairness in testing in the United States. In B. E. Clauser & M. B. Bunch (Eds.), *The history of educational measurement: Key advancements in theory, policy, and practice* (pp. 111–135). Routledge. https://doi.org/10.4324/9780367815318-6

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*(2), 201–210. https://doi.org/10.1177/014662168300700208

Stoynoff, S. (2013). Fairness in language assessment. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Blackwell. https://doi.org/10.1002/9781405198431.wbeal0409

Tarlani-Aliabadi, H., Tazik, K., & Azizi, Z. (2022). Exploring the role of language knowledge and background knowledge in reading comprehension of specific-purpose tests in higher education. *Language Testing in Asia*, *12*(1), 1–23. https://doi.org/10.1186/s40468-022-00198-x

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing*, *27*(2), 147–170. https://doi.org/10.1177/0265532209349465

Yoo, H., & Manna, V. F. (2017). Measuring English language workplace proficiency across subgroups: Using CFA models to validate test score interpretation. *Language Testing*, *34*(1), pp. 101–126. https://doi.org/10.1177/0265532215618987

Yoo, H., Manna, V. F., Monfils, L. F., & Oh, H.-J. (2019). Measuring English language proficiency across subgroups: Using score equity assessment to evaluate test fairness. *Language Testing*, *36*(2), pp. 289–309. https://doi.org/10.1177/0265532218776040

Zieky, M. J. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Erlbaum.

Zieky, M. J. (2013). Fairness review in assessment. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology*: *Vol. 1. Test theory and testing and assessment in industrial and organizational psychology* (pp. 293–302). American Psychological Association. https://doi.org/10.1037/14047-017

Zieky, M. J. (2016). Fairness in test design and development. In N. J. Dorans & L. L. Cook (Eds.), *Fairness in educational assessment and measurement* (pp. 9–32). Routledge. https://doi.org/10.4324/9781315774527-3

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

Appendix

Tables A1 and A2 present the scale score means, standard deviations, and the DIF flag results for both the listening and reading sections in the nine forms analyzed for the DIF analysis by age.

Table A1 Age DIF Flag Results for Listening in Forms A1 to A5, and B1 to B4 (Number of Flags)

	Focal group (under 22)		Reference group (at least 22)		DIF flag results (listening)					
Form	*M*	*SD*	*M*	*SD*	A+	A−	B+	B−	C+	C−
A1	311	78	339	83	47	46	2	4	1	
A2	301	79	340	83	48	50	1			1
A3	307	77	337	82	49	45	3	2		1
A4	302	80	337	83	50	47	1	2		
A5	304	79	335	84	50	47	1	2		
B1	311	77	344	83	46	50	2	2		
B2	310	80	346	83	46	51	2			1
B3	302	79	340	84	53	46		1		
B4	305	78	343	84	50	50				

Note. DIF flags with (+) favor the focal group (full-time students) and DIF flags with (−) favor the reference group (full-time employees).

Table A2 Age DIF Flag Results for Reading in Forms A1 to A5, and B1 to B4 (Number of Flags)

	Focal group (under 22)		Reference group (at least 22)		DIF Flag Results (reading)					
Form	*M*	*SD*	*M*	*SD*	A+	A−	B+	B−	C+	C−
A1	254	87	285	91	50	42	2	5		1
A2	248	87	286	90	54	43	1	1		1
A3	253	85	287	91	50	44	3	2		1
A4	248	89	288	93	47	52		1		
A5	246	85	282	91	52	47		1		
B1	254	88	294	95	51	43	2	4		
B2	252	92	293	97	55	41	1	3		
B3	244	87	288	95	51	49				
B4	251	91	295	96	48	52				

Note. DIF flags with (+) favor the focal group (full-time students) and DIF flags with (−) favor the reference group (full-time employees).

Suggested citation:

Schmidgall, J., Huo, Y., Cid, J., & Wei, Y. (2024). *Investigating fairness claims for a general-purposes assessment of English proficiency for the international workplace: Do full-time employees have an unfair advantage over full-time students?* (Research Report No. RR-24-06). ETS. https://doi.org/10.1002/ets2.12380

Action Editor: Larry Davis

Reviewers: Ru Lu, Saerhim Oh, Jennifer Sakano, and Satoka Shimoyama