Research Report

TOEIC® L&R 英語能力評価における公平性の検証：フルタイム就業者はフルタイム学生に対して優位性があるのか？

翻訳：印南　洋　　中央大学

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 RR-24-06

June 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 Signor Editorship is named in honor of Dr. Daniel R. Eignor, who from2001 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

TOEIC L&R 英語能力評価における公平性の検証：フルタイム就業者はフルタイム学生に対して優位性があるのか?

ジョナサン・シュミッドゴル、イエン・ホウ、ジェイミー・シド、ユーホア・ウェイ
ETS Research Institute, Princeton, New Jersey, United States

翻訳：印南　洋　　中央大学

連絡先：ジョナサン・シュミッドゴル（Jonathan Schmidgall）、メールアドレス：jschmidgall@ets.org

本研究における内容レビューにご協力いただき、有益なご意見を賜りました一般財団法人国際ビジネスコミュニケーション協会（IIBC）の研究開発部門の皆様に、深く感謝申し上げます。

要旨

伝統的に考えられてきたテストにおける公平性の原則とは、バイアスが存在しないこと、すなわち測定が公平に行われ、受験者集団間で不公平な有利・不利が生じないことを意味する。汎用的な目的で実施される言語テストでは、受験者の背景知識は言語熟達度の測定に関係しないと考えられてきた。そのため、仮に受験者間で背景知識に体系的な違いがあったとしても、その差が不公平な有利・不利につながってはならない。TOEIC® Listening & Reading Test は、日常生活や国際的な職場環境における英語使用を想定した汎用的な目的のテストであり、英語を第二言語とする（L2）使用者のリスニングおよびリーディングの理解力を測定するように設計されている。本研究では、職場経験の多い受験者グループ（フルタイム就業者）が、職場経験の少ない受験者グループ（フルタイム学生）に比べて、不公平に有利になるかを調査した。9 つのテスト版（計 1,800 項目）を用いて特異項目機能（DIF）分析を行った結果、統計的に差があると判断された項目は 18 問（全体の 1.0%）であった。専門家による精査の結果、これらの項目において就業者または学生のいずれかに有利なバイアスが明確に認められるものはなかった。得点の公平性評価を用いた分析からも、TOEIC スコアがフルタイム就業者（対フルタイム学生）を不公平に有利にするものではないことが示された。加えて、2 つの専門家委員会がテスト内容を検証し、フルタイム学生に不利になることなく、職場志向の内容をテストに取り入れていることを例示した。これらの分析結果は、TOEIC Listening & Reading Test のスコアが高等教育段階の受験者にとっても公平であることを裏付けており、背景知識と公平性の点から、汎用的な目的での言語テストに関する研究の一助となるものである。

キーワード：特定目的の英語（ESP）、妥当性論証、TOEIC®テスト、評価使用の論証、背景知識、読解力、聴解力、フルタイム労働者、フルタイム就業者、公平性に関する主張、国際的な職場
doi:10.1002/ets2.12380

はじめに

公平性（fairness）は、評価における中核的な原則であり、公正性（equity）、妥当性、信頼性と密接に関連している（Bachman & Palmer, 2010; Kunnan, 2018）。この原則に基づき、すべての受験者は、人種的背景、社会経済的地位、障がいの有無など、アイデンティティや個人のバックグラウンドに起因するバイアスや不利益を受けることなく、自らの知識や技能を平等に発揮できる機会を与えられるべきだとされている（American Educational Research Association [AERA] ほか, 2014）。もしテスト開発者が、公平性に関する懸念に対して、テスト開発や品質管理の過程で適切に対応できなかった場合、テストスコアの意味が損なわれ、受験者の知識や能力に関する偏った解釈を招く可能性がある。したがって、テストにおける公平性は、倫理的な義務にとどまらず、テストの質およびスコアの信憑性を担保する上でも極めて重要な要素である。

公平性に対する認識や定義は、時代とともに変化するものである（Sireci & Randall, 2021 を参照）。とはいえ、評価における公平性とは一般に、テストの質、テストの実施方法の質、テストの得点の質が、異なる受験者集団に対して一貫しており、かつ妥当である度合いと定義されることが多い（ETS, 2015; Xi, 2010）。この定義は、異なる受験者集団間でテストの難易度に差が生じないことを意味するものではない。実際、受験者間に熟達度の正当な違いが存在する場合、難易度に差が生じることはあり得る。重要なのは、受験者の背景やアイデンティティに関する要素のうち、評価対象である知識や技能と直接関係のないものによって、有利または不利に扱われるべきではないという点である。

公平性の原則と密接に関連しているのが、テストにおける偏りがないこと、中立であることが挙げられる（AERA ほか, 2014; Bachman & Palmer, 2010; Kunnan, 2007, 2018; Stoynoff, 2013）。テストの内容、実施方法、採点のいずれかが体系的かつ不適切に、特定の受験者集団を他の集団よりも有利にする場合、バイアス（偏り）が存在するとされる（Camilli, 2006）。したがって、測定対象であるスキルとは無関係な背景特性に基づき、特定の集団にとってテスト内容が有利に働く場合、バイアス―不公平な優位性―が生じる可能性がある。言語評価においては、バイアスの可能性を検討する際、受験者のアイデンティティや背景要因（性別、母語（第一言語：L1）、年齢、専攻分野など）に焦点が当てられることが多い（Kunnan, 2018, p. 173）。

評価において公平性は中核的な原則であり、テストの質にも不可欠であることから、テスト開発者は、公平性に関する懸念にどのように対応しているかを、テスト利用の妥当性論証全体の中で明確に説明すべきである（Kunnan, 2018 ; Xi, 2010）。本研究では、英語のリーディングおよびリスニング能力を測定する汎用的なテストの妥当性論証における、ある特定の主張―すなわち、学生と就業者の受験者間における得点解釈の中立性（または公平性）―に焦点を当てる。まず、言語評価における公平性の観点から、受験者の背景知識について整理し、第二言語（L2）のリーディングおよびリスニング評価において、背景知識が果たしうる役割について検討する。こうした理論的検討を通じて、TOEIC® Listening & Reading Test（TOEIC L&R）のスコア解釈が、学生と就業者受験者の双方に対して中立であるという主張の妥当性を検証するための理論的根拠を提示する。この主張を実証的に検証するために、本研究では混合研究法を用いた調査を実施する。

汎用目的テストと言語使用目的に特化したテストにおける公平性

言語テストにおける公平性を考えるうえで重要な観点の一つは、測定対象とする構成概念（construct）の定義に関わるものである。構成概念の定義には通常、測定対象である知識、技能、能力（KSAs）と、それらが測定される文脈の記述が含まれる。この文脈は、目標言語使用（target language use [TLU]）領域として特徴付けられ（Bachman & Palmer, 2010）、汎用的なもの（例：職場で使われる英語）から特定の用途に特化したもの（例：航空業界で使われる英語）まで、連続体上に位置づけることができる（Douglas, 2000）。特定目的での言語使用を専門とする研究者は、学術目的のための言語と職業目的のための言語という 2 つの主要な下位分野において、この連続体に沿って TLU 領域を概念化し定義することに取り組んでいる（Knoch & Macqueen, 2016）。

職業目的の言語使用では、より特定的な TLU 領域は、例えば航空業界でパイロットや航空管制官として働くために必要とされる英語のように、限定された専門的な対象に焦点を当てている。一方で、より一般的な TLU 領域は、特定の産業や職種に限定されない、より広範な言語使用状況を対象としている。このような違いにより、特定的な評価のための TLU 領域の定義（例：航空英語）は、実社会での特定の場面（例：パイロットと航空管制官の間の通信に用いられる英語）と非常に強く関連している。一方、より汎用的な領域（例：職場で使われる英語）は抽象化された言語使用状況に基づいており、その特徴は、関連する複数の実社会での場面（例：メール対応、会議、プレゼンテーション、出張など）に広く一般化できることが想定され、多様な産業や職種において適用可能である必要がある。

TLU 領域の特定性が高まることで生じることの一つに、受験者の背景知識や内容知識が、構成概念とどの程度関連するかということがある（Douglas, 2000）。この問題は、公平性やバイアスとも深く関係している。一般的に、より汎用的な領域に基づく構成概念には、背景知識は含まれないが、より特化された領域では、背景知識が構成概念の一部として不可欠であると見なされる。汎用的な領域では、タスクの文脈的特徴は広く定義され、さまざまなサブドメインや特定のコミュニケーション状況に共通して適用できることが想定される。たとえば、「職場で使われる英語」といった汎用的な領域に関係する話題は、多様な業種や職歴の言語使用者にとって馴染みのある内容であるべきである。しかし、トピック（またはその他の文脈上の特徴）が、その領域内のあるサブグループ（例：マーケティング・マネジャー）にとって他のサブグループよりも有利に働く場合、それは潜在的なバイアスの原因となり、スコア解釈の公平性や中立性を脅かすことになる（Bachman & Palmer, 2010）。一方で、領域がより特定的である場合、背景知識は構成概念に不可欠な要素と見なされ、公平性に対する同様の脅威とはならない。

L2 リーディング・リスニング評価におけるバイアスの要因としての背景知識

研究者は一般的に、背景知識を言語知識・技能・熟達度とは区別して捉えている（例：Douglas, 2000；Bachman & Palmer, 1996, 2010）。背景知識とは、受験者が既にもっている知識、すなわちリーディングやリスニングの課題で扱われる話題に対する馴染みの有無などを指す。背景知識には、テスト課題のインプットに関わる文脈的要素―たとえば状況、目的、語調、社会的規範、ジャンルなど―に対する親しみや理解を含む場合もある（Douglas, 2000）。

背景知識は、言語産出と理解の多くのモデルにおいて、重要な構成要素の一つとされている。状況モデル（Kintsch, 1998）は、外国語リーディングモデルにも影響を与えており（例：Grabe, 2009；

Grabe & Yamashita, 2022)、読者はテキストの主要な命題と自身の背景知識と統合することで、メンタルモデルを構築するとされる。したがって、言語使用者が理解活動において背景知識を用いることは避けられず、これは不可欠な認知プロセスの一部であると考えられる。

Banerjee（2019）は博士論文において、テスト受験者の背景知識（または先行知識、内容知識、トピック知識）の性質を表すために使用されてきた用語について、包括的なレビューを行っている。「先行知識（prior knowledge)」と「背景知識（background knowledge)」という用語は、一般的に、より領域全般的な知識、または「個人が知っていることの総和」を指す（Alexander, Schallert, & Hare, 1991, p. 333, as cited in Banerjee, 2019)。一方、「内容知識（content knowledge)」と「トピック知識（topical knowledge)」という用語は、より領域固有の知識を指す。他の研究者も、背景知識をドメイン一般知識とドメイン固有知識に分ける慣例に沿っている（Cai, 2013; Cai & Kunnan, 2018)。L2 のリーディングおよびリスニングの評価においては、テストの特定性および受験者の言語熟達度によって、背景知識がパフォーマンスに与える影響が異なることが示されている。複数の研究により、「二重の閾値効果（two threshold effect)」が支持されている。この効果によれば、言語熟達度の低い受験者は、自身の背景知識を効果的に活用することが難しい一方で、言語熟達度の高い受験者は、背景知識が十分でなくても、言語知識を活用して補うことができるとされる（Cai, 2013；Cai & Kunnan, 2018；Chung & Berry, 2000；Clapham, 1996；Ridgway, 1997)。

L2 の理解テスト（あるいはテキスト）において専門性の異なる複数の条件を用いた研究では、背景知識は、より専門的なテストにおいて成績をより的確に予測することが明らかにされている（Chung & Berry, 2000)。このことから、研究者は「テキスト内容がより専門的になるほど、背景知識がより重要になる」と主張している（Chung & Berry, 2000, p. 208)。Jensen and Hansen（1995）は、リスニングテストにおいて、専門的な講義と非専門的な講義の間で、背景知識の効果に差が見られたと結論づけており、専門的な内容においてその効果がより顕著であったと報告している。また別の研究では、経済学専攻の学生が専門目的のリーディングテスト（専門性の度合いが異なる 3 種類）を受験した際、テストの専門性が高まるにつれて、専門目的テストと汎用的なリーディングテストとの間の得点の相関が低下することが示された（Tarlani-Aliabadi et al., 2022)。

その他の研究では、背景知識が汎用的な L2 理解テストの成績に影響を及ぼす可能性があることが示唆されているが、その実際的な影響は限定的であるとされている。たとえば、汎用的な言語能力評価（TOEFL®）において、Hale（1988）は、受験者が自分の専攻分野（例：人文・社会科学系 vs. 自然・生物科学系）に関連するリーディングパッセージで、より高い成績を示す傾向があることを報告したが、その効果量は極めて小さく、スケールスコア（237〜677 点）のうち約 3 点の差に相当するに過ぎなかった。Hill and Liu（2012）では、TOEFL のリーディング理解テストにおいて、全 70 項目中 2 項目（約 3%）が背景知識に基づく統計的バイアス（C レベル DIF）を示したと報告したが、パッセージ全体でのバイアスは見られなかった。また、Jensen and Hansen（1995）は、リスニングテストにおいて、一部の（より専門的な）パッセージで先行知識が統計的に有意な主効果を持つことを報告したが、効果量は小さかった（pr^2 = .03〜.09)。Karami & Alavi（2012）は、文法・語彙・リーディングを含む汎用的なアカデミック英語テストにおいて、約 9%の項目が背景知識に基づく大きな DIF（Differential Item Functioning）を示したと報告しているが、全体のスコアにはバイアスは認められなかった。さらに、Lee（2011）は、TOEFL iBT®のリーディングパッセージに関するストラテジー使用の分析において、受験

者がより馴染みのある（vs. 馴染みのない）アカデミックパッセージを読んだ際にも、使用されたストラテジーに差は見られなかったと報告している。つまり、背景知識がストラテジー使用に影響を与えることはなかった。ただし、受験者はより馴染みのあるパッセージの方が、心理的に快適で自信をもって取り組めると感じていたと述べている。

以上をまとめると、L2 のリーディングおよびリスニングにおける背景知識の影響に関する研究では、次のような傾向が示されている。すなわち、テストの専門性が高い場合（＝より特定の目的に基づくテスト）や、受験者が一定の言語熟達度の閾値を超えている場合には、背景知識がパフォーマンスに与える影響が大きくなる。

本研究の背景：TOEIC リスニング＆リーディングテストの妥当性論証における公平性

TOEIC L&R は、英語を母語としない人の日常生活や職場での英語のリスニングとリーディングの能力を評価するために設計されている（ETS, 2022a)。TOEIC L&R が参照する TLU 領域は、より一般的な職場領域の一例とされている（Knoch & Macqueen, 2016, p. 293)。この TLU 領域は、特定の産業や職業に特化したものではなく、テスト内容は、日常的および職場での多様な文脈や話題を含んでおり、幅広い若年層および成人の L2 英語使用者にとって親しみのある内容となるよう設計されている。具体的には、TOEIC L&R には、日常生活に関連する状況（旅行、娯楽、健康、外食など）や、一般的な職場環境に関する状況（企業の人材育成、財務、製造、オフィス業務、人事、購買など）が含まれている（ETS, 2022a, pp. 3–4)。これらの文脈におけるトピック内容は、業界特有の専門的な語彙や談話ジャンルの使用を最小限に抑えるように工夫されている。

したがって、より汎用的な目的の言語テストとして、TOEIC L&R が測定する能力には、職場における専門的な背景知識は含まれないことが意図されている。つまり、そうした専門的知識―通常は実務経験を通じて得られるもの―が、受験者にとって有利にも不利にも働くべきではないということである。この点は、TOEIC L&R の妥当性論証において提示されている、スコアの解釈の公平性や中立性に関する複数の主張に影響するものである。

Bachman and Palmer（2010）による評価使用の論証（Assessment Use Argument: AUA）は、妥当性論証を構築するための包括的な枠組みであり、TOEIC プログラムのテストにも適用されている（例：Schmidgall, 2017；Schmidgall et al., 2021)。AUA は、テストスコアの特性、スコアの解釈、意図された使用目的、ならびにテスト使用の結果に関する一連の相互に関連する主張（クレーム）で構成される。スコアの解釈に関する重要な特性の一つは、それがすべての受験者グループに対して中立的であることである。このスコアの中立性に関する主張は、以下の補足的な主張によって精緻化できる：

(1) TOEIC L&R の項目は、特定の受験者に有利または不利に働く回答形式や内容を含まない。

(2) TOEIC L&R スコアに基づくリスニングおよびリーディング理解力に関する解釈は、異なる受験者グループ間であっても等しい意味をもつ。

妥当性論証においては、各主張（クレーム）がテスト設計、品質管理に関する文書や手続き、そして実証的研究から得られる証拠によって裏付けられることが求められる（Bachman & Palmer, 2010)。たとえば、上記のステートメント 1 は、テスト項目が本試験として使用される前に実施される多段階の内容および公正性レビューの手続きを通じて裏づけられる。これらの手続きは、特定の受験者グループ（例：就業者 vs. 学生、年長者 vs. 若年者、男性 vs. 女性）に対して不公平に有利となる内容が項目に

含まれないようにすることを目的としている。すべての項目は、ETS の品質および公平性に関する基準（ETS, 2015）のトレーニングを受けたアセスメント開発スタッフによる公平性レビュー（Zieky, 2013 を参照）を受けている。さらに、DIF 分析と呼ばれる統計的手法が、テスト実施後の予備的項目分析の一環として、性別を対象に実施されており、性別によってバイアスのある項目が存在する場合には、それらを特定し、必要に応じて削除する措置が取られている。また、テスト内容に懸念がある場合は、ETS に直接連絡する方法についての情報も受験者に提供されている。また、前述のステートメント 2 についても、実証的な裏付けがある。Yoo and Manna（2017）は、TOEIC L&R における相関 2 因子モデルの因子不変性を、5 つの受験者属性（性別、年齢、就業状況、英語学習時間、英語圏での滞在経験）にわたって検証した。その結果、厳密な測定不変性および構造的不変性がすべてのグループ間で確認され、構成概念の構造がグループ間で一貫していることが示された。さらに Yoo ら（2019）はスコア等価性評価（score equity assessment）（Dorans, 2004）を用いて、サブグループへの所属（性別、年齢、学歴、言語環境、過去のテスト受験経験）が、スコア算出に使用される統計的・心理測定的手法の結果に与える影響を検証した。その結果、調査対象としたサブグループ間でスコアの比較可能性と意味の一貫性が確保されていることが分かった。

学生と就業者の間で職場に関する背景知識の違いが存在する可能性を踏まえ、TOEIC L&R のスコア解釈が両者に対して本当に中立であるかどうかについて疑問を抱く利害関係者がいても不思議ではない。そのような関係者にとっては、前述のステートメント 1 および 2 を支持するさらなる証拠が求められる可能性がある。L2 のリスニングおよびリーディング理解における背景知識の役割に関する先行研究では、背景知識がテストパフォーマンスに影響を与えることがあると示唆されている。実際、全体的な傾向として、フルタイム就業者は、フルタイム学生よりも TOEIC L&R で平均点が高い（例：ETS, 2023, p. 8）。こうした平均点の差が、両グループ間に実際の英語力の差が存在することによるのであれば、問題とはならない。しかし、英語力が等しいにもかかわらず、フルタイム学生とフルタイム就業者の間で同じスコアを得る確率に差があるとすれば（おそらく背景知識の差によって）、それはスコア解釈の公平性に関する主張を損なうものであり、TOEIC L&R が背景知識をどの程度測定しているのかについての再検討が必要かもしれない。

TOEIC L&R の妥当性論証における公平性に関するこれらの主張をさらに検証するために、以下の研究課題を設定した：

1. TOEIC L&R の項目には、就業状況（フルタイム就業者とフルタイム学生）に基づくバイアスが認められるか。
2. 就業状況に関連するバイアスは、TOEIC L&R のスコアに影響するか。

方法

題材

TOEIC L&R は、紙ベースまたはコンピュータベースで実施される試験であり、リスニングとリーディングの 2 つのセクションから構成されている。各セクションには 100 問が含まれ、それぞれ時間を区切って実施される。リスニングセクションの所要時間は 45 分で、音声録音によって進行が管理される。一方、リーディングセクションは受験者自身のペースで進めることができ、最大 75 分間で完了す

るよう設計されている。テストの内容、実施手順、スコアリングおよびスコアの解釈、測定の質、そして想定される利用目的についての概要は、『TOEIC L&R 受験者ハンドブック』（ETS, 2022a）および『スコア利用者ガイド』（ETS, 2022b）に詳しく記載されている。

本研究における分析には、過去に実施された 9 つの TOEIC L&R テストフォームの内容および受験者データが用いられた。これらのうち、同一年に実施された最初の 5 つのフォームは、本研究において A1～A5 と番号が振られた。残りの 4 つのフォームは別の年に実施され、本研究では B1～B4 とされた。A1～A5 のフォームでは、すべての項目が実際のスコア計算に使用された。一方、B1～B4 のフォームでは、品質管理手続きにより意図した機能を果たさないことが確認された 4 問が、スコア計算から除外された。実施スケジュールに関しては、A1～A3 のフォームは同日の午前セッションで実施され、A4 および A5 は午後セッションで実施された。同様に、B1 および B2 は午前セッション、B3 および B4 は午後セッションで実施された。各テストフォームにおけるリスニングおよびリーディングセクションの信頼性（すなわち、同一あるいは同等フォームの繰り返し実施においてスコアが一貫している程度）は、内部一貫性法に基づく KR-20 信頼性係数（Kuder & Richardson, 1937）により推定された。すべてのフォームにおいて、リスニングおよびリーディングセクションの信頼性推定値は十分に高く、0.91～0.94 の範囲にあった。

TOEIC 受験者の人口統計的属性を特定するために、TOEIC 背景質問票（Background Questionnaire: BQ）を用いた。BQ は、受験者の学歴、職務経験、英語の使用および学習状況、TOEIC 受験経験に関する情報を収集する質問票である。この質問票は、TOEIC L&R の実施中、リスニングセクションを開始する前に受験者に対して配布・実施された。

参加者

9 つの TOEIC L&R テストフォームそれぞれにおける受験者数は、1 万人から 2 万人の間であった。すべての受験者がリスニングおよびリーディングの両セクションを受験したため、両セクションにおけるサンプルサイズは等しい。受験者の人口統計的特性（性別、年齢、職業）は、フォーム A1 から A5、フォーム B1 から B4 について表 1 に示す。

各テストフォームの受験者全体のサンプルは、受験者の就業状況に基づいてサブサンプルに分割された。TOEIC の背景質問票（BQ）における質問 3 では、「現在のご自身の状況として最も適切なものを以下からお選びください」（ETS, 2022a, p. 22）と尋ねている。この質問において、選択肢 1 は「フルタイムで働いている（自営業を含む）」、選択肢 2 は「パートタイムで働いている、またはパートタイムで学んでいる」、選択肢 3 は「就業していない」、選択肢 4 は「フルタイムの学生である」となっている。本分析では、選択肢 1 または選択肢 4 を選んだ受験者のみを対象とし、フルタイム就業者（基準グループ）またはフルタイム学生（焦点グループ）として分類した。一方、選択肢 2 または選択肢 3 を選んだ受験者は、就業状況に基づく分析からは除外された。また、年齢に基づく別の DIF 分析においては、就業状況に関係なくすべての受験者を対象とし、22 歳未満の受験者と 22 歳以上の受験者の 2 つの年齢グループに分類した。22 歳未満のグループを焦点グループ、22 歳以上のグループを基準グループとした。

表 1 に示されているように、同一セッション内の受験者は、性別、年齢、就業状況の構成において、同一試験実施内で別のセッションを受験した受験者よりも類似していた。一般的に、午前のセッション（A1～A3、B1 および B2）は、午後のセッション（A4 および A5、B3 および B4）と比較して、フル

タイム就業者および 22 歳以上の受験者の割合が高かった。さらに、午前・午後のどちらのセッションにおいても、男性受験者の割合が女性受験者より高く、22 歳以上の受験者が多数を占めていた。これらの傾向は、2 つの試験実施（A および B）において一貫しており、TOEIC の試験運営において一般的に見られるものである。

表 1. 各フォームにおける受験者の構成比（%）

		性別		年齢		就業状況 [a]	
セッション時間	フォーム	女性	男性	22 歳未満	22 歳以上	フルタイム学生	フルタイム就業者
午前	A1	36%	64%	27%	73%	39%	53%
	A2	43%	57%	29%	71%	43%	47%
	A3	38%	62%	27%	73%	40%	51%
午後	A4	40%	60%	34%	66%	51%	38%
	A5	42%	58%	34%	66%	51%	38%
午前	B1	40%	60%	27%	73%	38%	53%
	B2	40%	60%	25%	75%	35%	55%
午後	B3	43%	57%	35%	65%	54%	35%
	B4	43%	57%	35%	65%	52%	37%

[a] 就業状況の割合は、フルタイム学生およびフルタイム就業者のみを対象とし、合計しても 100%にはならない。

研究手順とデータ分析

TOEIC L&R が就業者受験者に対して不公平に有利に働いていないかを検証するため、本研究では項目レベル（個々の項目のパフォーマンス）およびスコアレベル（テストセクション全体のパフォーマンス）の両方で分析を実施した。項目レベルでは、DIF 分析を用いて、英語能力が同等のサブグループ間でテスト項目に差が見られるかどうかを評価した。また、DIF パネルとして構成された内容領域の専門家による項目内容のレビューも実施した。スコアレベルでは、フルタイム就業者とフルタイム学生の間でのスコアの比較可能性（スコアの一貫性）を検証した。この心理測定学的アプローチ（詳細は次節「パート 2」で説明）は、各サブグループのテストスケールスコアと、全体のスケールスコアとの比較に基づいて行われるものである。

項目レベルおよびスコアレベルでの分析を補足するために、TOEIC L&R の 1 つのフォームに対する独立した内容レビューを実施し、職場関連の内容（例：状況、トピック、語彙）がどのように項目に取り入れられているかについてのさらなる知見を得ることを目指した。この内容分析は、本研究のパート 1 およびパート 2 が完了した後に構想され、実施されたものである。内容分析の目的は、パート 1 およびパート 2 で行われた統計的分析を補足することであり、職場関連の内容を含む「典型的な」項目の具体例を提示することで、バイアスが認められない項目について理解を深めることにあった。

パート 1：項目レベルの DIF 分析

DIF 分析は、各項目の心理測定学的特性が焦点グループと基準グループの間でどの程度異なるかを評価するものであり、項目が焦点グループに有利に働いているかどうかに関する統計的証拠を提供する。DIF 分析において関心のあるグループは焦点グループと呼ばれ、比較対象となるグループは基準グループと呼ばれる。本研究では、ETS で用いられている DIF 分析手法に従い、Mantel-Haenszel デルタ差（MH D-DIF）統計量およびその統計的有意性を用いて各項目の DIF を評価した。この手法では、項目を A、B、C の 3 つのカテゴリに分類し、DIF の程度が A から C にかけて大きくなる（Zwick, 2012）。Zieky（1993, p. 342）の分類によれば：

A カテゴリの項目：DIF が無視できる程度、あるいは統計的に非有意。

B カテゴリの項目：軽度から中程度の DIF。

C カテゴリの項目：中程度から大きな DIF。

さらに、B および C カテゴリの項目は DIF の方向性（どちらのグループに有利か）によって細分化される：

B+／C+：項目が焦点グループ（本研究では学生）に有利。

B-／C-：項目が基準グループ（本研究では就業者）に有利。

本研究では、SAS 統計解析ソフトウェアを用いて、Hao（2013）の MH D-DIF 分析用 SAS マクロプログラムを適用し、DIF 分析を実施した。ここで、フルタイム学生を焦点グループ、フルタイム就業者を基準グループとして設定している。なお、これらの DIF 分析は、本研究の第 2 著者、第 3 著者、第 4 著者によって実施された。

C レベル[1]（C-または C+）に分類された項目については、DIF パネル（Zieky, 2016）によるレビューが行われた。この DIF パネルは、ETS に所属する 5 名のアセスメント開発の専門家で構成されており、分析によってフラグが立てられた各項目を個別に検討し、その項目が焦点グループ（学生）または基準グループ（就業者）に対してバイアスを示しているかどうかを評価した。パネルの評価において、ある項目が明確にバイアスを示していると判断された場合、その項目はスコア算出から除外すべきとされる。これは、テストスコアにおけるバイアスを最小限に抑え、公平性の原則を促進するためである。なお、この DIF パネルによるレビューの過程は、本研究の第 1 著者が観察者として立ち会った。

パート 2：テストレベルの分析

項目レベルでのバイアスの存在は懸念されるが、テスト全体のスコアにおけるバイアスが認められる場合には、スコア解釈の公平性や、特定の受験者グループに対する適切なテスト利用に関して、より深刻な懸念が生じる。DIF 分析は、バイアスの可能性がある個々のテスト項目を特定するために一般的に用いられる統計手法であるが、DIF 分析ではテストスコア全体におけるバイアスの程度を定量的に示すことはできない。このような課題に対して、スコア等価性評価（Score Equity Assessment: SEA）は、公平性の観点から構成概念妥当性を検討するための手法である（Dorans, 2004）。スコア等価性評価

[1] MH D-DIF の絶対値が統計的に有意に 1 を超えており、かつ絶対値が少なくとも 1.5 以上である。

の中心的な考え方は、等化におけるサブグループ不変性の要件を通じて、テストスコアがサブグループと全体グループの間で同じ構成概念を測定しているかどうかを評価することである。つまり、受験者の属性（サブグループ所属）がスコアに影響を与えないのであれば、サブグループから導出されたスコアは、受験者の背景にかかわらず、全体グループのスコアと同等に比較可能であるべきである。より具体的には、各サブグループから導出される等化関数がほとんど差異を示さない場合、そのテストはすべてのサブグループにおいて同じ構成概念を測定していると考えることができる。

等化（equating）分析の一般的な手順は、項目反応理論（IRT: Item Response Theory）に基づいている（Lord & Novick, 1968）。本研究で使用された IRT モデルは、2 パラメータ・ロジスティックモデル（2PL モデル）であり、このモデルでは能力値に加えて、項目識別力と項目難易度のパラメータを推定する。IRT を用いた等化手続きは、以下の 3 つのステップで構成される。第 1 ステップでは、R の mirt パッケージ（Chalmers, 2012）を使用し、複数グループ IRT キャリブレーションによって、フルタイム学生グループとフルタイム就業者グループそれぞれにおける IRT 項目パラメータを推定した。第 2 ステップでは、推定された IRT 項目パラメータを基準用のテストフォームのスケールに変換した。これにより、スコアの等化のための基準とすることができる。この変換には、Stocking-Lord 法（Stocking & Lord, 1983）という手法が用いられた。第 3 ステップとして、変換後の項目パラメータを用いて、IRT を用いた真値の等化が実施され、各サブグループ（学生、就業者）ごとに新しいフォームのスコアが等化された。これらの項目キャリブレーション、スケール変換、およびスコア等化の手続きは、それぞれのサブグループに対して実施されたのと同様に、各テストフォームの全体グループ（全受験者）に対しても実施された。したがって、本研究では、全体グループとサブグループの等化比較を評価するために、以下の 3 つの受験者グループに対して等化関係が算出された：フルタイム学生サブグループ、フルタイム就業者サブグループ、全体グループ（すべての受験者）。

各テストフォームにおいて、2 つの受験者サブグループ（フルタイム学生とフルタイム就業者）に基づくスコア変換の差異は、全体グループに基づくスコア変換と比較された。これらの差異は、Dorans ら（1994）によって提案された意味のある差（difference that matters: DTM）の基準に従って評価された。DTM はスコア単位の半分として定義されており、TOEIC L&R セクションスコアのスコア単位が 5 点であることから、本研究では DTM を 2.5 点と設定した（非四捨五入のスケールスコアにおいて）。この基準値を下回る差は、「無視できる差（negligible）」と見なされた。なお、この DTM の 2.5 点という値は、TOEIC L&R 各セクションの標準測定誤差（standard error of measurement）である 25 点（ETS, 2022b, p. 19）を大きく下回っている。すべてのスコアレベルにおける分析は、本研究の第 2 著者、第 3 著者、および第 4 著者によって実施された。

パート 3：内容レビュー

職場を想定した内容（例：職場の状況設定や職場に関連する言語表現）を含むテスト項目の例を特定し、就業者あるいは学生の受験者にとって不公平な利点を与える可能性があるかどうかを評価するために、内容レビューが実施された。ただし、このレビューでは実際のテスト項目の内容を本稿に開示することになるため、パート 1 およびパート 2 で使用されたテストフォームと特性が類似している、過去に実施された別の TOEIC L&R テストフォームを選定した。内容分析に先立ち、パート 1 で説明した手続きと同様の方法で DIF 分析を実施した。C レベルの DIF が認められた項目については、DIF 評価委員

会によるレビューが行われた。

内容レビューは、二段階に分けて二つの専門家グループによって実施された。第一段階では、TOEIC L&R のアセスメント開発に携わる二名のコンテンツリードが、刺激素材および項目を含むテストフォームを精査し、職場を想定した内容を含む TOEIC L&R テスト項目の代表例を特定した。代表例と判断された各項目については、職場関連の内容がどのように取り入れられているか（例：状況設定の性質、トピック、使用または暗示された語彙）に関するコメントが付され、また、その内容が就業者（対学生）にとって不公平な利点となりうるかどうかについても評価がなされた。第二段階では、第一段階で特定されたすべての項目を、外部の評価パネルが検討・議論した。この段階では、対象項目がすでに職場の文脈に位置づけられていると判断されていたため、各項目に含まれる内容が、就業者あるいは学生のいずれかに不公平な利点を与える可能性があるかどうかに焦点を当てて評価が行われた。この評価パネルは、TOEIC プログラムの日本国内パートナーである国際ビジネスコミュニケーション協会（IIBC）の職員三名で構成されていた。参加した IIBC 職員は、TOEIC L&R の内容に精通しており、英語の運用能力を有し、また受験者層（フルタイム学生、フルタイム就業者）に関する知識を有していた。

結果

パート 1：項目レベルの DIF 分析

表 2 および表 3 には、DIF 分析で用いた 9 つのテストフォームにおけるリスニングセクションおよびリーディングセクションそれぞれについて、DIF の検出結果、ならびに各サブグループのスケールスコアの平均値および標準偏差が示されている。各セクションには、分析対象として 100 問の項目が含まれている。表中に示されている検出数は、それぞれの DIF レベルで検出された項目数を表している。運用上のスコアリング（採点）から除外された項目が 4 問存在したが、これらも DIF 分析には含められ、すべての実施済み項目に対して就業状況による DIF を調査した。なお、この 4 問を DIF 分析から除外しても、本研究における結果は本質的に同一であった。

表 2. フォーム A1～A5 および B1～B4 におけるリスニングセクションの就業状況別 DIF 検出結果（フラグ数）

フォーム	焦点グループ（学生）		基準グループ（就業者）		DIF 検出結果（リスニング）					
	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			

注：DIF フラグの（+）は焦点グループ（フルタイム学生）に有利な項目を、(−) は基準グループ（フルタイム就業者）に有利な項目を示す。

表 3. フォーム A1〜A5 および B1〜B4 におけるリーディングセクションの就業状況別 DIF 検出結果（フラグ数）

フォーム	焦点グループ（学生）		基準グループ（就業者）		DIF 検出結果（リーディング）					
	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		

注：DIF フラグの（+）は焦点グループ（フルタイム学生）に有利な項目を、(−) は基準グループ（フルタイム就業者）に有利な項目を示す。

すべてのテストフォームおよびセクションにおいて、フルタイム就業者グループのスケールスコア平均は、フルタイム学生グループの平均よりも高かった。本稿では掲載していないが、フルタイム就業者グループは素点の平均値においてもフルタイム学生グループを上回っており、独立サンプルによる Welch の t 検定により両グループ間の平均値の差は統計的に有意であることが確認された（$p < .0001$）。なお、Welch の t 検定を採用したのは、分散の等質性の仮定が成り立たなかったためである。

2 グループ間の平均値の比較をより適切に把握するため、平均値の差の大きさを評価する目的で効果量を算出した。効果量の指標として、サンプルサイズの不均等性を考慮できる Hedges の g 統計量（Hedges, 1981）を用いた。リスニングおよびリーディングセクションの素点に対して、フルタイム学生グループとフルタイム就業者グループの間で算出された Hedges の g による効果量は 0.2〜0.4 の範囲に収まり、両グループ間の差は比較的小さいものであると解釈される（Cohen, 1992）。

C レベルの DIF が検出された項目は少数であった。全体として、各テストセクションおよびフォームごとに、就業状況に関連する C レベルの DIF が検出された項目数は 0〜3 問であった。全フォーム・セクションを通じて、C レベルの DIF が検出された項目は合計 18 問であり、全体の 1.0%に相当する。リーディングセクションよりもリスニングセクションで多くの項目が C レベル DIF としてフラグされており、その数はリスニングが 11 問（1.2%）、リーディングが 7 問（0.8%）であった。また、C レベ

ルの DIF が検出された項目のうち、フルタイム学生に有利な項目（C+）よりも、フルタイム就業者に有利な項目（C−）の方が多く見られた。

すべての C レベル DIF 項目について、DIF パネルによる内容レビューが実施され、フルタイム就業者またはフルタイム学生のいずれかに有利とみなされる可能性のある内容を含んでいるかどうかがさらに検討された。パネルの結論によれば、18 項目中 9 項目において、焦点グループであるフルタイム学生に不利なバイアスの可能性が認められた。これらの項目では、「colleague（同僚）」のような、フルタイム学生には馴染みの薄い職場関連の語彙が使用されている点が指摘された。一方、5 項目については、いずれのグループに対しても明確または潜在的なバイアスが存在するという証拠は認められなかった。最終的に、パネルは、いずれの項目も明確にバイアスがあるとは判断されないと結論づけた。バイアスの可能性があるとされた項目についても、それらは構成概念に関連する内容を含んでいると判断された。さらにパネルは、バイアスの背後にある要因として、年齢が潜在的に影響している可能性があると指摘した。実際、2 つの項目においては、若年層よりも高年齢層の受験者に有利に働く可能性があると評価された。たとえば、ある項目のレビューでは、「pharmacy（薬局）」という語やその概念が若年層の受験者にとっては十分に知られていない可能性があるため、年齢が影響因子となっている可能性が示唆された。

就業状況と年齢に基づくサブグループは、必ずしも完全に区別されるわけではないことに留意する必要がある。日本における大学卒業年齢は通常 22 歳であるため、22 歳未満の受験者の多くはフルタイム学生に分類され、フルタイム就業者の多くは少なくとも 22 歳以上であった。しかし、22 歳以上の受験者のうち約 20%～40%は、自身をフルタイム学生と申告しており、逆に 22 歳未満でフルタイム就業者とされた受験者は 1%未満であった。このような状況から、年齢が分析結果に及ぼす潜在的な影響を検討することは妥当であると考えられた。したがって、年齢に基づく補足的な DIF 分析も実施された。その結果は付録 A に報告されている。年齢に基づく DIF 分析では、C レベルの DIF として検出された項目はごく少数であった（リスニング項目では 4 問、すなわち全リスニング項目の 0.4%、リーディング項目では 3 問、すなわち全リーディング項目の 0.3%）。年齢ベースの DIF 分析で検出されたすべての項目は、就業状況に基づく DIF 分析でも同様に検出されていた。ただし、就業状況に基づいて検出されたすべての項目が年齢ベースでも検出されたわけではなかった。年齢と就業状況による分類は相互排他的ではないため、各項目における受験者のパフォーマンス傾向は、これら 2 つの属性に関連したグループ特性が複雑に絡み合っていると考えられる。

パート 2：テストレベルの分析

運用上の採点から項目が除外されていない 5 つのフォームのうち、A2（午前セクション）および A5（午後セクション）のフォームについては、サブグループ間におけるスケールスコアの比較可能性の観点から評価が行われた。他のフォームでは、DIF が検出されなかった項目のうち 2 項目が IRT キャリブレーション後にサブグループ間で適切に機能しなかったため、テストレベルにおける DIF 項目の影響を示す分析には A2 および A5 のフォームのみを用いた。図 1 は、フォーム A2 のリスニングセクションにおける、フルタイム学生グループと全体グループの非四捨五入換算スコアの差、ならびにフルタイム就業者グループと全体グループの差を示している。図 2 は、同じくフォーム A2 におけるリーディングセクションの非四捨五入換算スコアの比較結果を示している。図 3 および図 4 は、それぞれフォーム A5

におけるリスニングセクションおよびリーディングセクションの非四捨五入換算スコアの比較結果を示している。

図 1〜図 4 に示されているように、横軸にはリスニングまたはリーディングの素点（0〜100 点）が表示されている。縦軸は、サブグループと全体グループとの間で算出されたスケールスコアの差を表している。各図において、2.5 および−2.5 は、「重要とされる差異（DTM）」の正負の範囲を示しており、スケールスコアの差が正方向・負方向のいずれにも生じうることを意味する。リスニングセクションにおける偶然的な正答（選択肢をランダムに選んだ場合の想定正答数：27 点）、およびリーディングセクションにおける同様の点数（25 点）は、図中で縦線により示されている。2 本の変動する折れ線は、それぞれ全体グループと比較して、フルタイム学生グループおよびフルタイム就業者グループにおけるスケールスコアの差を、素点に応じて表している。これらの図は、各サブグループと全体グループとの間で得られたスケールスコアの違いを可視化し、比較を可能にするものである。スケールスコアにおける大きな差異は、各サブグループにおいて、全体グループとの比較可能性が低いことを示唆する可能性がある。しかし、すべての図において、サブグループと全体グループとの換算スコアには軽微な差異しか認められなかった。これらの差異の程度はセクションやフォームによって異なり、換算スコアの比較において一貫したパターンは確認されなかった。

図 1. フォーム A2 におけるリスニングセクションの全体グループおよびサブグループ間の換算スコア比較

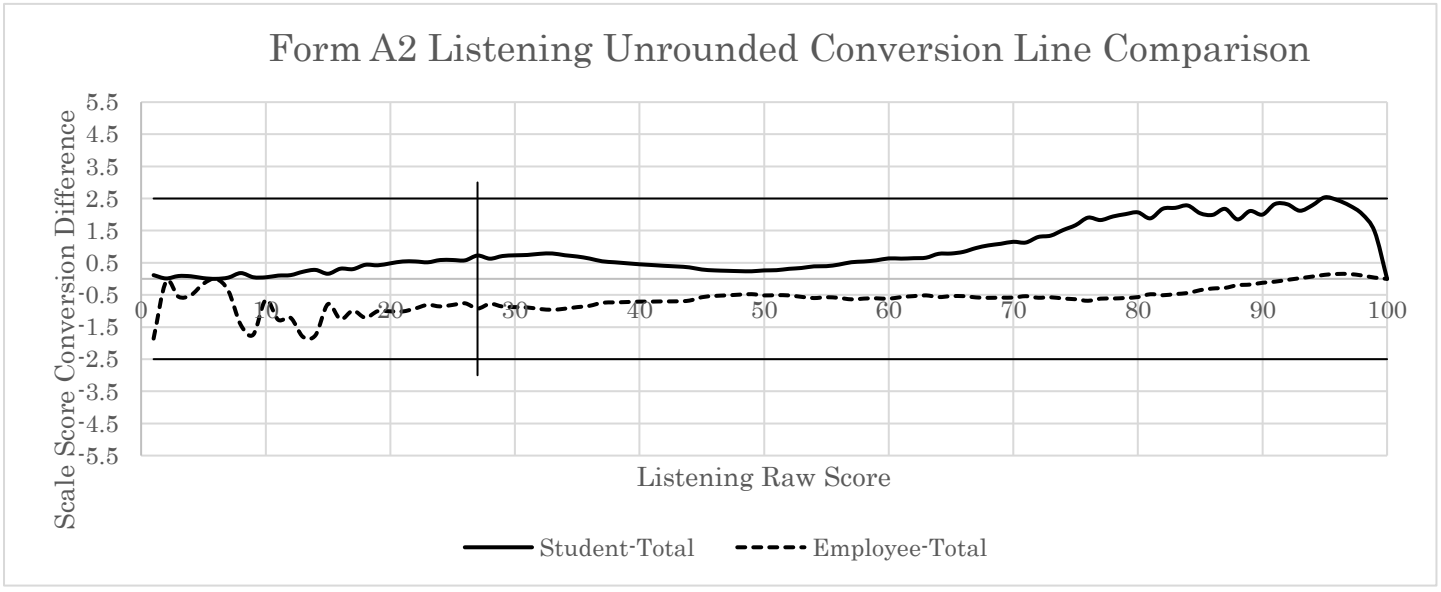


図 2. フォーム A2 におけるリーディングセクションの全体グループおよびサブグループ間の換算スコア比較

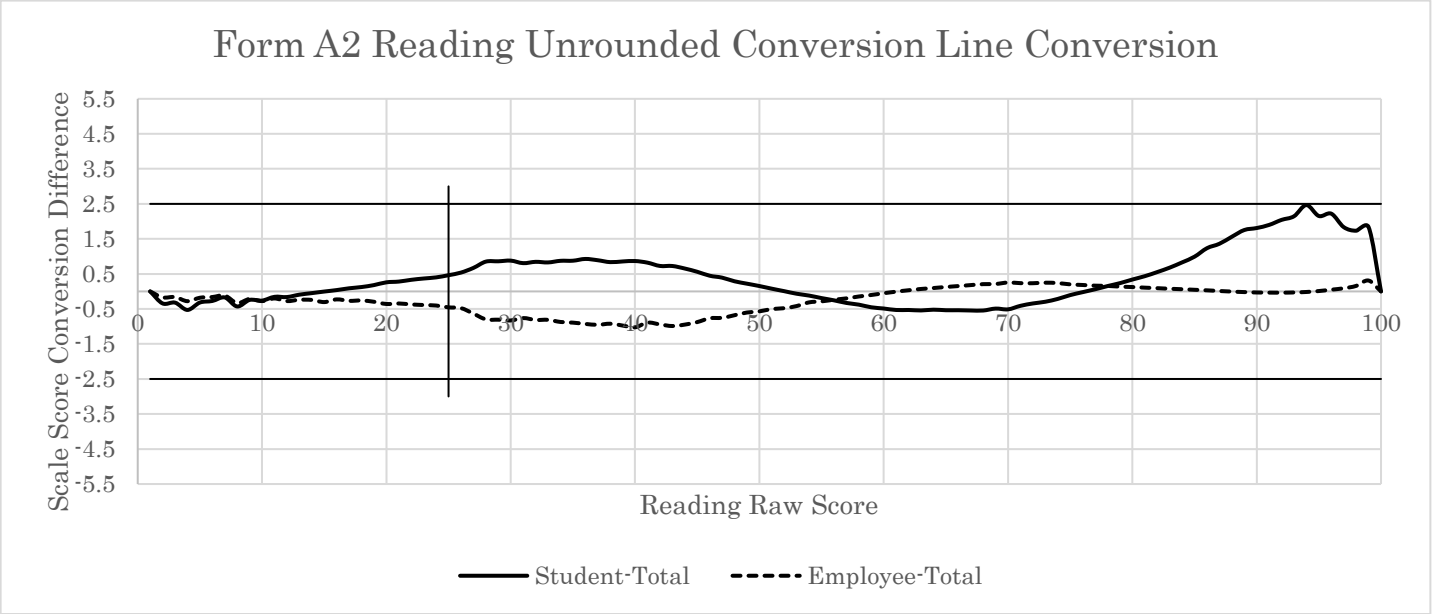


図 3. フォーム A5 におけるリスニングセクションの全体グループおよびサブグループ間の換算スコア比較

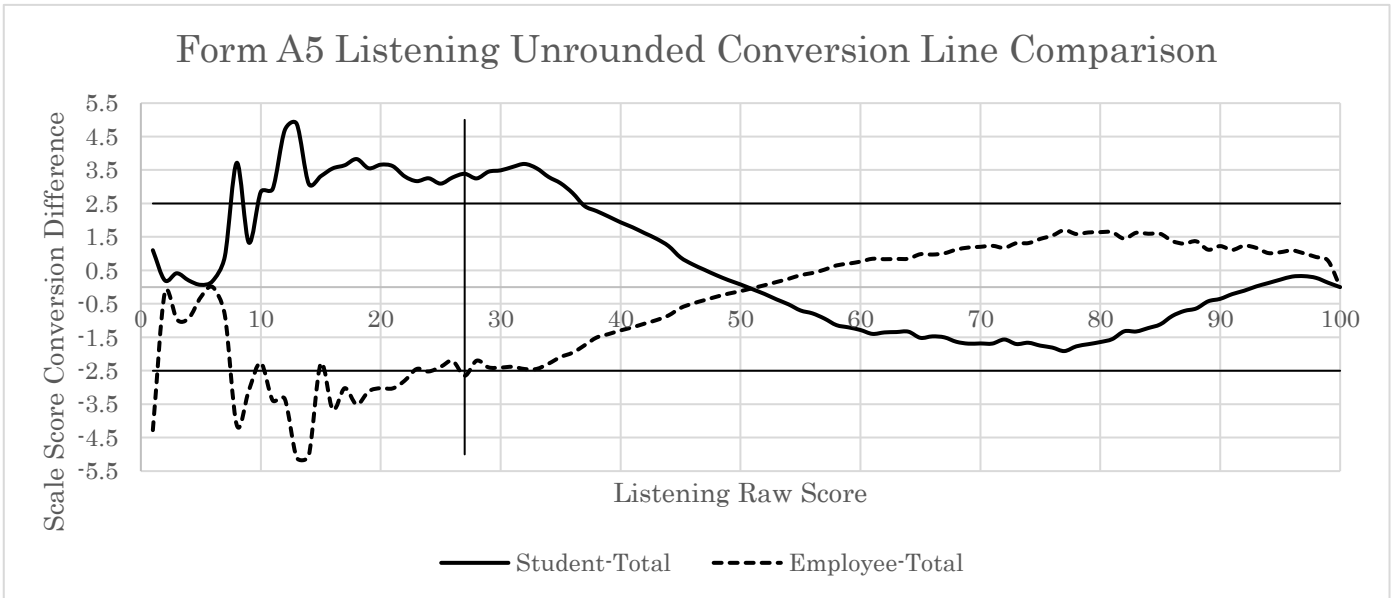
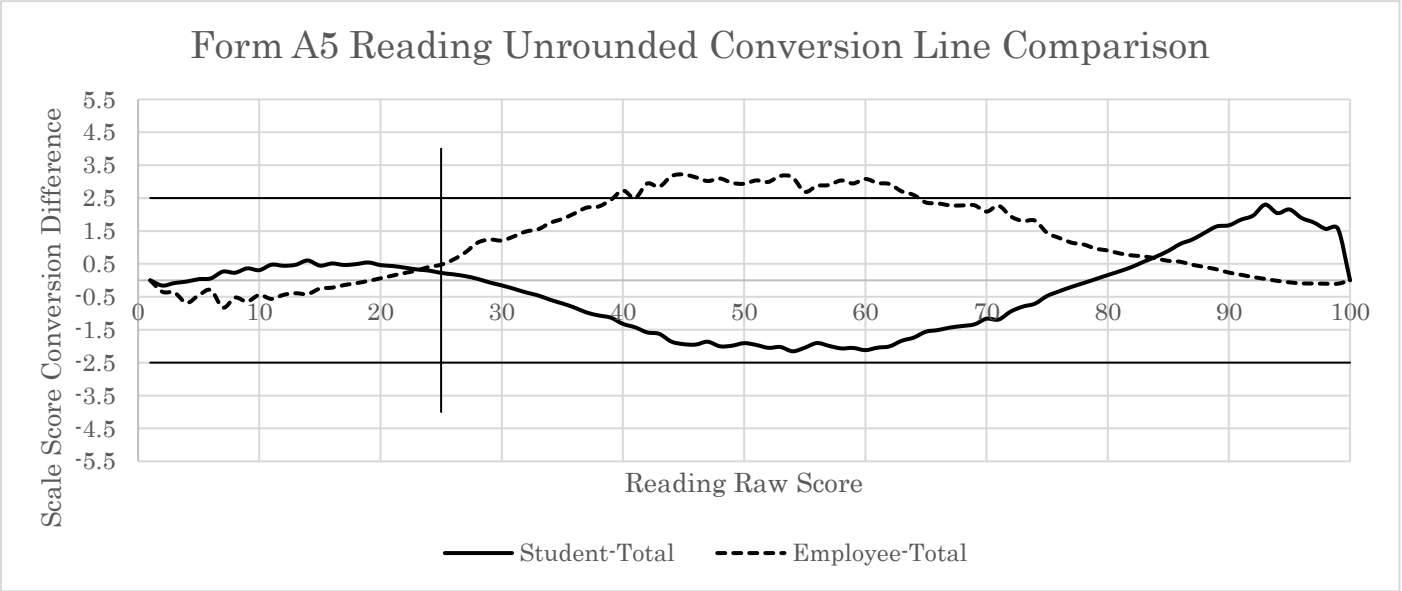


図 4. フォーム A5 におけるリーディングセクションの全体グループおよびサブグループ間の換算スコア比較



結果として、すべてのフォームにおいて、偶然に正答する水準を上回る素点の大部分において、サブグループと全体グループとの換算スコアの差は DTM の範囲（±2.5 点）内に収まっていた。ただし、フォーム A5 のリスニングセクションにおいては、フルタイム学生グループにおいて、素点 25 点から 36 点の範囲で DTM を超える差異が確認された。また、同じくフォーム A5 のリーディングセクションでは、フルタイム就業者グループにおいて、素点 40 点から 64 点の範囲で DTM を超える差異が見られた。これらの差異は DTM の閾値を超えてはいるものの、その乖離の程度は大きくない。

総じて、スコアレベルの分析においては、一方のサブグループから得られた換算スコアが、他方のサブグループの換算スコアに比べて体系的に高くまたは低くなるという傾向は認められなかった。同様に、全体グループの換算スコアとサブグループの換算スコアとの間の差異も、リスニングおよびリーディングのいずれにおいても、素点の大部分においてはごく僅かなものであった。換算スコアの比較結果は、TOEIC L&R テストにおける現在の等化手法が、全体グループを基準とする換算において妥当であることを示している。フルタイム学生およびフルタイム就業者といったサブグループごとの換算は、各グループの特性をある程度反映する可能性はあるものの、サブグループに基づく等化による差異は、スコア報告の観点から見て実質的な影響はないと判断できる。

パート 3：内容レビュー

内容レビューに使用された TOEIC L&R テストフォームに対して実施された DIF 分析では、2 項目（リスニング 1 項目、リーディング 1 項目）が C レベル DIF として検出された。これら 2 項目について、DIF パネルによるレビューが行われた結果、いずれの項目にもフルタイム学生またはフルタイム就業者に明確に不公平に有利になる内容は含まれていないと判断された。したがって、第 1 段階の内容レビューにおいては、すべてのテスト素材および項目がそのまま保持された。

ETS の内容レビュー担当者は、TOEIC L&R の全パートにわたり、職場に関連した内容（場面

設定、語彙など）を含む代表的な項目を特定した。リスニングセクションにおいては、以下の項目が該当した：写真描写問題（Photograph）3 問、応答問題（Question-Response）1 問、会話問題（Conversation）1 セット（1 問）、説明文問題（Talk）2 セット（合計 6 問）。リーディングセクションにおいては、以下の項目が該当した：短文穴埋め問題（Incomplete Sentence）3 問、長文穴埋め問題（Text Completion）2 セット（合計 8 問）、1 つの文書問題（Single Passage）2 セット（合計 4 問）。これらのリスニングおよびリーディング項目・セットにおける場面設定としては、以下のようなものが含まれていた：一般的なオフィス環境（例：会議、電話、ビデオ会議、オリエンテーション）、非オフィス系の職場環境（例：倉庫、工業系の現場）、公共空間（例：博物館、官公庁、顧客向けウェブページ）。

両方の内容レビュー・パネルにおいて、すべてのテスト項目は専門的な背景知識を必要とせずに正答可能であるという点について、概ね一致した見解が得られた。加えて、両パネルは、職場環境に対する（フルタイム就業者のほうが一般的に高いと推測される）親しみやすさに基づいて、フルタイム就業者がフルタイム学生に対して不公平に優位になることはないと結論づけた。

職場に関連した内容がテスト項目にどのように取り入れられているかを示すために、内容レビューで取り上げられたリスニングセクションおよびリーディングセクションから、それぞれ 3 問（または 3 セット）の項目を再掲し、パネルによるコメントを付して紹介する。なお、掲載にあたっては、項目の形式を一部変更しており、項目の指示文や説明は省略されている（TOEIC L&R のサンプルテストについては、ETS［2019］を参照のこと）。

リスニングセクション：写真描写問題の例

写真描写問題（Photograph item）では、受験者は 1 枚の写真に関する 4 つの英文音声を聞く。英文は問題冊子上に印刷されていない。受験者は、写真の内容を最も適切に描写している文を選び、マークシート形式（紙ベースの場合）では A〜D の中から 1 つの選択肢を選んでマークし、コンピュータ形式では画面上で 1 つの選択肢を選ぶ。写真描写問題は、以下の能力を測定することを目的として設計されている：(a) 話の要点または中心的な内容を理解する能力、または (b) 明白な詳細情報を理解する能力。

Test takers see:

Test takers hear:

(Woman, British voice):

(A) One of the women is throwing away some files.

(B) One of the women is holding a piece of paper.

(C) The man is taking off his glasses.

(D) The man is typing on a laptop.

このリスニングの項目は、明白な詳細情報を理解する能力を測定することを目的として設計されている。内容レビュー担当者は、本項目がオフィス環境を想定した設定に基づく一例であると指摘した。音声で流れる選択肢には写真に写る人物達が何をしているかという説明が含まれるが、その説明には「files（書類）」「paper（紙）」「laptop（ノートパソコン）」といった一般的（非専門的）かつ多くの状況に適用可能な語彙が用いられている。正答（選択肢 B、太字で強調されている）は、職場での経験がなくても選ぶことが可能であり、フルタイム就業者が学生に対して有利となることはないと判断された。

リスニングセクション：応答問題の例

応答問題（Question-Response item）では、受験者は 1 つの質問または発言と、それに続く 3 つの応答（いずれも英語で読み上げられる）を聞く。これらのやり取りは 2 人の話者によって行われる。質問（または発言）および応答文はいずれも問題冊子上に印刷されていない。受験者は、質問（または発言）に対する最も適切な応答を選び、該当する選択肢をマークする。この項目形式は、受験者が短いやり

取りの中で以下の能力を測定することを目的として設計されている：(a) 話の要点、目的、基本的な状況を理解する能力、(b) 詳細情報を理解する能力、または (c) 含意（暗示されている意味）を理解する能力。

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.

この項目は、短い音声テキストにおける含意を理解する能力（語用論的知識）を測定することを目的として設計されている。内容レビュー担当者は、本項目の場面が会議やスケジュールに関するものであり、これはリスニングセクションでよく見られる一般的な場面（電話対応、クライアントとのビデオ会議、スケジュール管理など）であると指摘した。このタイプの項目は、正答に含まれる推論を理解する必要があるため、より難易度が高いことを意図している。今回の項目においては、最も適切な応答（選択肢 B、太字で強調されている）は、女性が会議に関する情報を何も持っていないことを暗示しており、その結果として質問に明確に答えることができないという意味合いを含んでいる。日本人の内容レビュー担当者は、第一言語の背景がこの項目の難易度に影響を及ぼす可能性があると推測した。というのも、日本語では「〜を持っている（所有）」と「会議がある（出来事の存在）」に対して異なる動詞を使用するため、英語の "have" の用法が一義的に理解されない可能性があるからである。とはいえ、両グループの内容レビュー担当者は、本項目で取り上げられているタスクや概念は、理解にあたって職業経験を必要としないと判断した。

リスニングセクション：説明文問題の例

説明文問題（Talk question set）では、受験者は 1 人の話者によるスピーチを聞く。その後、話の内容に関する 3 つの項目に答える。各項目は音声で読み上げられるが、紙ベースの試験では問題冊子にも印刷され、コンピュータベースの形式では画面にも表示される。受験者は、最も適切な選択肢を 1 つ選ぶ。このタイプの項目は、受験者が以下の能力を有しているかどうかを評価することを目的としている：(a) 話の要点（主旨、文脈）を推測する能力、(b) 詳細情報を理解する能力、または (c) 話者の意図や発話に含まれる含意を理解する能力。

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.

このトークは新入社員向けのオリエンテーションという文脈に位置づけられており、これはオフィス環境の一例である。こうした状況は、学生にとっても（たとえば学校初日など）類似の経験があると考えられ、親しみやすい内容である。トークおよび項目内で使用されている語彙は、「lockers（ロッカー）」「break room（休憩室）」「factory（工場）」「lunches（昼食）」などの専門的でなく一般的な語彙である。項目は典型的な構成であり、以下に関する事柄の理解が必要になる：トークの要点または想定される文脈（製造工場、選択肢 A）、トーク中に言及される詳細情報（個人用ロッカーの利用可能状況、選択肢 C）、女性の発話に含まれる語用論的含意（選択肢 B）。両グループの内容レビュー担当者は、本トークにおける一般的な状況や語彙は、フルタイム学生およびフルタイム就業者の双方にとって十分に理解可能であると判断した。

リーディングセクション：短文穴埋め問題の例

短文穴埋め問題（Incomplete Sentence item）では、一部が空欄になっている文が提示される。受験者は、その空欄を最も適切に補完する語または語句を 4 つの選択肢から選ぶ。この項目形式では、文レベルの文法知識または語彙知識のいずれかが問われる。

Test takers read:

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

(A) high

(B) higher

(C) highly

(D) highest

この項目は、形容詞に関する文法項目を測定している。内容レビュー担当者は、本項目が扱う話題にはビジネス的な要素が含まれており、「sales figures（売上実績）」のような語彙は学生にとってなじみが薄い可能性があると指摘した。しかし、この語彙の知識は項目に正答するために必須ではない。つまり、受験者は項目に含まれるビジネス的な内容を理解していなくても、正解を選ぶことが可能である。

リーディングセクション：長文穴埋め問題の例

長文穴埋め問題（Text Completion set）では、受験者はさまざまな形式の短い文章を読む。各文章には、語、句、または主要な文など、4 か所の欠落箇所が含まれている。受験者は、それぞれの空欄に最も適切な語、句、または文を 4 つの選択肢の中から選び、補完する。この項目形式では、文法力、語彙力、および短い文章内で情報をつなぐ力が問われる。

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

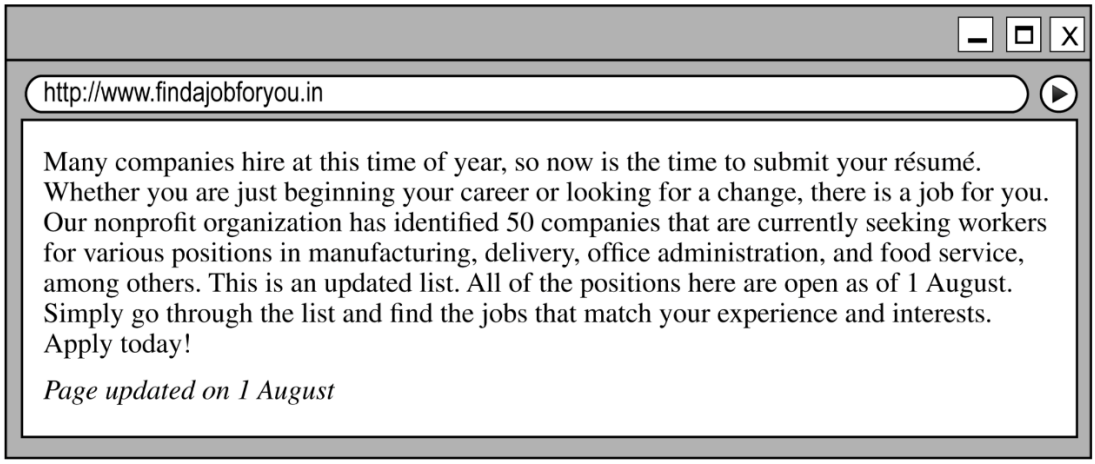
Key: 135 = A; 136 = C; 137 = C; 138 = A.

このセットは、会計業務における顧客対応であり、顧客を意識したテキストを含んでいる。内容レビュー担当者は、「reception area（受付）」「consultation rooms（相談室）」「clients（顧客）」といった表現について、専門的な語彙ではなく、どのような職場環境でも一般的に見られるものであると指摘した。この項目では、非専門的な語彙を用いて、語彙力および文法力が問われている。そのため、たとえ本テキストがビジネスサービスに関連する文脈に基づいていたとしても、内容レビュー担当者は、本テキストおよび項目は多様な受験者層にとって広く理解可能であり、ビジネスに関する知識を必要とせず、またそれがアドバンテージとなることもないと結論づけた。

リーディングセクション：１つの文書問題の例

１つの文書問題（Single Passage set）は、リーディングセクション内のリーディング問題に含まれており、受験者は日常的または職場関連のテキストを読む。各テキストに付随する項目では、以下のような能力が求められる：主旨の把握、明示された詳細情報の特定、含意の理解（たとえば、文脈、筆者の意図）、テキスト内または複数のテキスト間における情報の統合。

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.

このリーディングテキストは、求人ウェブサイトという場面に位置づけられており、職場に関連する状況として、成人および若年成人のいずれにも広く親しみのある内容である。内容レビュー担当者は、テキストに含まれる言語が非専門的であり、就業を求める側・提供する側の誰にとっても理解可能な語彙で構成されていると指摘した。求人広告というジャンル自体が、学生およびフルタイム就業者のいずれにもアクセス可能なものであり、「hire（雇用する）」「submit your resume（履歴書を提出する）」「nonprofit organization（非営利団体）」といった表現も、さまざまな職場場面において広く適用可能な語彙であると指摘された。

考察

汎用的な目的で実施される言語テストでは、多くの場合、背景知識や話題に関する知識を構成概念とは無関係な変動要因として扱うことが多い。本研究では、TOEIC L&R における構成概念の定義およびその運用が、TLU 領域である国際的な職場に関する背景知識に基づくスコア解釈の公平性に関して、どのような主張を可能にするかを検討した。具体的には、職場に関する知識や経験が豊富であると想定されるフルタイム就業者が、職場経験が限られるフルタイム学生に対して不公平に優位になるかどうかを調べるために、項目レベルおよびスコア全体レベルの分析を実施した。項目レベルにおける DIF 分析の結果、DIF が統計的に検出された項目は全体の 1%程度と非常に少なく、バイアス評価パネルによって

DIF が明確に存在すると判断された項目はなかった。

スコアレベル分析の結果は、就業状況に基づく DIF 分析から導かれた結論を、統計的に補強するものである。ごく一部の項目が DIF の対象として検出されたものの、TOEIC L&R は、日常生活および一般的な職場環境における英語運用能力を測定するという枠組みのもと、フルタイム学生とフルタイム就業者という異なる受験者グループ間で一貫して機能していた。さらに、スコア等化の手法は、異なる属性や特徴をもつサブグループに対するテストの公平性を、心理測定的な観点から担保する役割を果たしている。将来的な研究においては、より多くのテストフォームを用いて、スコア等価性評価の枠組みに基づいた追加的な等化分析を実施することが望ましい。

本レポートで例示したテスト項目は、TOEIC L&R において職場の文脈や話題がどのように取り入れられているかを示しており、それらが専門的な職場知識や経験を必要とせず、またそれによって有利・不利が生じることのないように設計されていることを明らかにしている。本テストに含まれる職場の文脈やトピックは、多様な背景をもつ受験者にとって広くアクセス可能で、なじみのある内容となるよう意図されている。また、やや特殊な文脈や話題が含まれている場合でも、言語知識こそが測定過程においてより重要な役割を果たすよう設計されている。このことは、汎用的な目的で実施される言語テストにおける背景知識の役割に関する先行研究（例：Hill & Liu, 2012）とも一致している。

したがって、本研究の結果は、TOEIC L&R スコアが受験者の背景知識（すなわち、学生と就業者の違い）に対して公平であるという主張に対し、実証的な裏付けを提供するものである。また、本研究の結果は、TOEIC L&R スコアが性別に関しても公平であることを示した先行研究の結果を補強するものでもある。

ある意味において、たとえその定義が広範であったとしても、特定の TLU 領域に文脈化された言語テストが、そのドメインの内部者（例：フルタイム就業者）に有利に働かないという結果は、意外であると捉えられるかもしれない。本研究においては、フルタイム就業者が学生よりも有意に高い成績を示すという証拠は得られなかったが、両グループの受験者がテストを受ける際の体験には、多少の違い（たとえば、心理的な快適さや自信の程度）がある可能性がある。L2 のリーディングおよびリスニングテストにおける背景知識の効果を扱った研究の多くは、パフォーマンスの差異に焦点を当てているが、受験体験に注目した研究では、背景知識の多い受験者の方が、テスト中により高い快適さや自信を感じると報告する傾向があることが示されている（例：Lee, 2011）。

本研究における一つの限界点は、背景知識の概念化のあり方に関するものである。本研究では、背景知識を、より一般化された職場における TLU 領域に関する一般的な知識として捉えた。具体的には、関連するコミュニケーションタスク、状況、話題に関する知識やなじみの度合いを含むものとした。このような捉え方は、背景知識を「よりドメイン汎用的な知識（すなわち、個人が有する知識全体）」として定義する立場と一致しており、「特定の話題に関する専門的・限定的な知識」とは区別される（Banerjee, 2019）。本研究では、背景知識の中でも最も広義にあたる、職場という領域に属する者（本研究におけるフルタイム就業者）が経験を通じて獲得しうる、領域一般知識に焦点を当てた。先行研究においては、このような領域一般知識が、領域一般的なリーディング力を予測する可能性が示されている（Cai & Kunnan, 2018）。しかし、領域があまりに広く定義される場合には、そのような知識による有利性はなくなる可能性もある。

慎重に設計された汎用目的の L2 リーディングまたはリスニングテストにおいて、背景知識が優

位性をもたらすかどうかを検討する際には、「優位性（advantage）」という語の意味内容が重要となる可能性がある。本研究では、学生と就業者のいずれかが、個々の項目あるいは全体のスコアにおいて成績面で優位になるかどうかに焦点を当てた。一方で、内容レビューに参加した１名のパネルからは、特定の項目において職場を題材とした内容により、学生がその文脈や語彙に不慣れであるために、理解や正答に至るまでにより多くの時間を要する可能性があるとの指摘もあった。本研究では、こうした受験体験上の差異（親しみやすさが安心感や不安に与える影響）を検討することはできていない。つまり、背景知識によって快適さや処理速度に違いが生じたとしても、それがスコアの優劣として表れるとは限らないという点に、本研究の限界がある。本研究から分かることは、職場経験（および関連する背景知識）がテストパフォーマンスに与える全体的な影響についての、一般的な結論である。したがって、今後の研究においては、受験者のパフォーマンスのみならず、受験中の体験や認知的・感情的な反応に焦点を当てた検討が求められる。

参考文献

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

付録

表 A1 および表 A2 は、年齢に基づく DIF 分析において対象となった 9 つのテストフォームにおけるリスニングセクションおよびリーディングセクションのスケールスコアの平均、標準偏差、および DIF 検出の結果を示している。

表 A1. フォーム A1～A5 および B1～B4 におけるリスニングセクションの年齢別 DIF 検出結果（フラグ数）

フォーム	焦点グループ（22 歳未満）		基準グループ（22 歳以上）		DIF 検出結果（リスニング）					
	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				

注：DIF フラグの（+）は焦点グループ（22 歳未満）に有利な項目を、（−）は基準グループ（22 歳以上）に有利な項目を示す。

表 A2. フォーム A1～A5 および B1～B4 におけるリーディングセクションの年齢別 DIF 検出結果（フラグ数）

フォーム	焦点グループ（22 歳未満）		基準グループ（22 歳以上）		DIF 検出結果（リーディング）					
	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		

注：DIF フラグの（+）は焦点グループ（22 歳未満）に有利な項目を、（−）は基準グループ（22 歳以上）に有利な項目を示す。

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

アクションエディター：Larry Davis
レビュアー: Ru Lu, Saerhim Oh, Jennifer Sakano, and Satoko Shimoyama
日本語版レビュアー：Renka Ohta

日本語版発行日：2025 年 10 月
日本語版発行：一般財団法人 国際ビジネスコミュニケーション協会
（The Institute for International Business Communication; IIBC）
〒164-0001　東京都中野区中野 4-10-2　中野セントラルパークサウス 5F
公式サイト https://www.iibc-global.org