異なる形式のTOEIC L&Rテストのスコアの非同等性について

異なる形式の TOEIC L&R テストのスコアの非同等性について
Non-equivalence of TOEIC L&R Scores Across Different Test Versions

橋本 将*
Masashi HASHIMOTO

概要

英語客観試験として広く使用される TOEIC Listening & Reading テストは，従来，受験者共通の問題 200 問を紙と鉛筆を使って解答させるマークシート方式のテストのみが提供されていたが，2020 年 4 月から，受験者の能力に応じて出題される問題が変わるアダプティブなオンライン形式のテスト（「TOEIC Listening & Reading IP テスト（オンライン）」）の提供が始まった．これら実施形式が異なる 2 種のテストのスコアは，公式には同じ意味を持つとされているが，IP テスト（オンライン）の方がスコアが高くなるという報告もあり，これまでのところテストスコアの関係ははっきりしていない. 本研究は, 金沢大学で 2019 年と 2021 年に実施した従来の形式の TOEIC テストのスコアと，2020 年に実施した IP テスト（オンライン）のスコアを，2019 年から 2021 年に実施した等化可能な英語期末試験のデータと組み合わせて分析することで，実施形式の異なる TOEIC テストのスコアの関係を調査した．その結果，英語能力がある程度高い場合に，従来の形式のテストよりも IP テスト（オンライン）の方がスコアが高く出ることがわかった．

1. はじめに

英語客観試験として，TOEIC Listening & Reading（以下 TOEIC）テストは現在広く使用されている．TOEIC テストは，従来，紙と鉛筆によるマークシート方式で，2 時間に全受験者共通の 200 問を解く形式のテストだけであったが，2020 年 4 月より，「TOEIC IP テスト（オンライン）」と呼ばれる，オンライン方式で 1 時間に 90 問を解く形式のテストが加わった．この新しいテストは，受験者の能力によって出題される問題が異なるテスト（アダプティブテスト）である．本研究では，この 2 種類のテストのスコアが同等であるかを検証した．

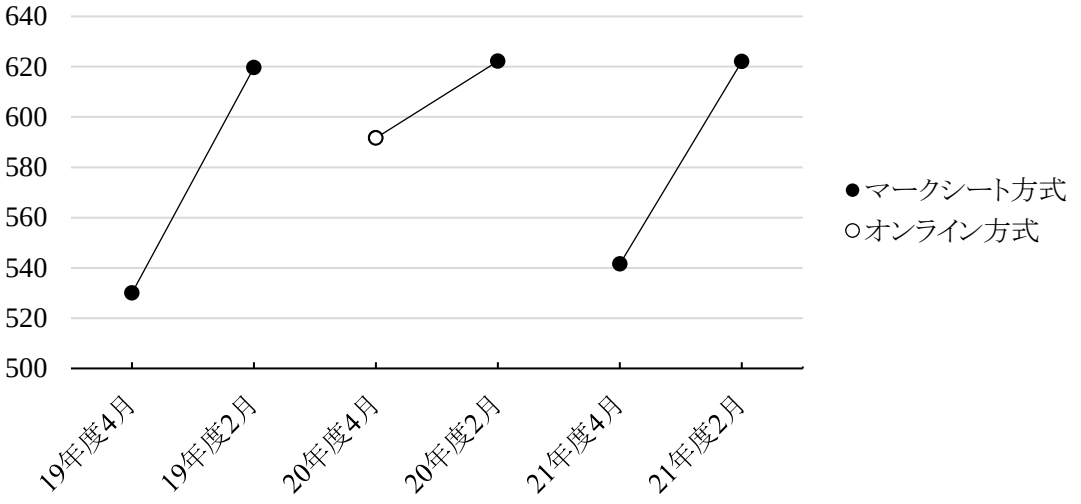
金沢大学では，1 年生の英語力を客観的に測定するために，2016 年度からマークシート方式の TOEIC テストをほぼ全 1 年生を対象に 2 月に実施している．また，2019 年度からは総合教育部の 1 年生（約 130 名）を対象に 4 月にも TOEIC テストを実施することで，1 年次の英語教育の 1 年間の効果を（全学ではないが）測定できるようにしている．しかし，2020 年には，新型コロナウイルス感染症の流行が始まり，4〜5 月は大学が登学禁止となったため，2020 年度の 4 月 TOEIC テストは，マークシート方式の従来のテストから，サービスが開始されたばかりの TOEIC IP テスト（オンライン）に切り替えて実施された．その後，大

* 金沢大学国際基幹教育院

学の対面授業が再開されると, 再び従来の方式の TOEIC テストが実施されるようになった.

図 1 は，このようにして実施された 2019～2021 年度の総合教育部入学者の 4 月と 2 月の TOEIC テストのトータルスコア平均の推移である．この期間に，2 月 TOEIC テストのトータルスコア平均は約 3 点しか変化しておらずほぼ一定だが，4 月 TOEIC テストのトータルスコア平均を見ると，2020 年度は 2019 年度と比べ約 62 点も上昇し，その翌年 2022 年度は約 50 点下降と，変動が激しい．そのため，4 月から 2 月までの約 1 年間のトータルスコア平均の伸びは， 2019 年度は約 90 点であったのが，2020 年度は約 31 点に大幅に減少し，それから 2021 年度は約 81 点となって 2019 年度とほぼ同じ程度に戻っている.

図 1. 2019～21 年度総合教育部入学者の 4 月・2 月 TOEIC テストのトータルスコア平均



TOEIC テストを実施している ETS Global（2022）や国際ビジネスコミュニケーション協会（IIBC）(2020）は，従来のマークシート方式の TOEIC テストと新しいオンライン方式の TOEIC テストのスコアの意味は同じだとしているが，それが正しいとすると，金沢大学の総合教育部の 2020 年度入学者の英語能力は，前年度・翌年度入学者と比べて 4 月入学時には顕著に高く，1 年間の伸びは顕著に小さかったということになるだろう.

しかし，新型コロナウイルス感染症対策で特に 2020 年度前期の授業形態が対面から遠隔に変わっていたとはいえ, 1 年間のスコアの伸びが約 90 点から約 30 点に減少するほど英語教育の内容・効果が変わっていたというのは非常に考えにくい. それよりは，2020 年度の 4 月の TOEIC IP テスト（オンライン）のスコアが，従来の方式のテストのスコアよりも（平均 5，60 点程度）高く出る傾向にあったと考える方がもっともらしいのではないだろうか.

そこで，本研究では，2 つの異なる実施形式の TOEIC テストのスコアを同等とみなしてよいか検証を行った. 2 種類のテストのスコアの関係の研究には，同一の受験者グループを対象にそれらのテストを間をあまり置かずに実施するのが理想的であるが，それは費用等の理由で困難であったので，本研究では，実施年の異なる 2 種類の TOEIC IP テスト間の比較を，TOEIC テストに準拠した等化可能な期末試験を利用することで行うことを試みた.

以下，本論文の第 2 節では，現行の TOEIC テストの実施形式について説明したのち，2 種類の TOEIC テストのスコアの関係についての先行研究を検討する．次に，第 3 節で，分析の手法と扱うデータについて説明したのち，分析結果を提示する．第 4 節で分析結果の考察を行い，第 5 節で結論を述べる．

2. TOEIC テストの実施形式とスコア

2.1. 解答方式とテスト構成が異なる TOEIC テスト

TOEIC テストは，運営・管理する団体の違いによって，TOEIC テストを開発している Educational Testing Service (ETS) またはそのパートナー（日本では IIBC）が運営・管理する「公開テスト」と，学校や企業などの団体が運営・管理する「Institutional Program（IP）テスト」の 2 種類に分けられる．従来は，公開テストと IP テストのいずれも，2 時間でリスニング問題 100 問とリーディング問題 100 問の合計 200 問の解答をマークシートに鉛筆でマークする，マークシート方式のテストしかなかった．しかし，日本では 2020 年 4 月から，IP テストに限って，1 時間でリスニング問題 45 問とリーディング問題 45 問の合計 90 問を PC 上で解答する，オンライン方式のテストも提供されるようになった．「IP テスト（オンライン）」と呼ばれるこのテストは，リスニング問題とリーディング問題のそれぞれのセクションが 2 つのステージ（ユニット）に分けられており，どちらのセクションも，最初のステージでは全受験者に共通の問題が出題されるが，2 番目のステージでは，最初のステージの結果に基づいて受験者ごとにその能力に合った問題が出題される[1]．表 1 に，これらのテストの解答方式とテスト構成を示す．

表 1. TOEIC テストの実施形式

<table>
<tr><th></th><th colspan="2">公開テスト</th><th>IP テスト（オンライン）</th></tr>
<tr><td>解答方式</td><td colspan="2">マークシート方式</td><td>オンライン方式</td></tr>
<tr><td rowspan="4">テスト構成</td><td colspan="2" rowspan="2">リスニングセクション
（全受験者共通）</td><td>リスニングセクション
Unit One（全受験者共通）</td></tr>
<tr><td>リスニングセクション
Unit Two（受験者により異なる）</td></tr>
<tr><td colspan="2" rowspan="2">リーディングセクション
（全受験者共通）</td><td>リーディングセクション
Unit One（全受験者共通）</td></tr>
<tr><td>リーディングセクション
Unit Two（受験者により異なる）</td></tr>
</table>

[1] 海外では，公開テストをオンラインで受験できる受験地もある．ただし，オンラインの公開テストは，アダプティブテストではなく，従来と同じように全受験者共通の 200 問を 2 時間で解くリニアテストであり，解答方法だけがマークシート用紙と鉛筆から PC に変わったものである．アダプティブテストを受験できるのがオンラインの IP テストだけであるのは海外でも同じである．

2.2. TOEIC IP テスト（オンライン）のスコアの意味

公式には，TOEIC IP テスト（オンライン）のスコアと従来の TOEIC IP テストのスコアは同等であるとされている．ETS が公開している *TOEIC Score User Guide* には，“The TOEIC IP Online tests are identical to the paper-and-pencil versions in terms of scoring, timing, content, and test format”（ETS, 2022, p. 8）とあり，IIBC のニュースレター（IIBC, 2020, p. 5）でも，「評価やスコアの意味合いは，公開テストや従来の IP テストと同様で，スコアが同じであれば，英語力も同等です」と述べられている[2]．ただし，現在のところ，TOEIC IP テスト（オンライン）と従来の TOEIC IP テストのスコアの同等性についての ETS の研究結果は公開されていない．

2.3. TOEIC IP テスト（オンライン）のスコアに関する先行研究とその検討

TOEIC IP テスト（オンライン）のスコアと従来のテストのスコアの関係を考察した研究論文は現在のところほとんどないが，希な例外として，岡山大学のグループの研究（寺西ら，2021）がある．その分析によると，IP テスト（オンライン）のスコアは従来のテストのスコアと同等ではなく，高くなりやすいという．本節ではこの研究について，簡単な説明と検討を行う．

岡山大学では，新入生の入学時に英語の全学統一試験が実施されている．この全学統一試験は，2014 年から 2017 年まではマークシート方式の TOEIC IP テストであったが，2018 年からは GTEC Academic 2 技能テスト[3,4]（以下，GTEC テスト）に変更になった（2022 年まで）．2014 年–2017 年実施のマークシート方式の TOEIC IP テストの全学トータルスコア平均は約 472 点であった．この全学統一試験に加えて，寺西らは，2021 年 4 月に 92 名[5]の入学者を対象に TOEIC IP テスト（オンライン）を実施した．そのトータルスコア平均は 540.92 点であった．

2021 年 4 月実施の GTEC テストのトータルスコアの全学平均は 500 点満点中 243.96 点であり，TOEIC IP テスト（オンライン）も受験した 92 名に限った平均は 243.53 点で，全学平均とほぼ同じであったので，寺西らは，TOEIC IP テスト（オンライン）を受験した 92 名のグループは全学新入生の平均的なサンプルであると指摘した．そして，そのことから，2021

---

[2] ETS Global のサイトにも同様の記述がある（“The level of difficulty, scoring scale and corresponding CEFR levels are identical”（ETS Global, 2022））．

[3] GTEC Academic テストは大学生向けのテストであり，リスニング能力とリーディング能力を測定する 2 技能テストと，リスニング能力，リーディング能力，ライティング能力，スピーキング能力を測定する 4 技能テストがある．どちらも，TOEIC IP テスト（オンライン）と同じように，PC 上で行うアダプティブテストである．

[4] 2018，2019 年度はキャンパス内 PC での実施，2020，2021 年度は自宅 PC での実施であった．

[5] TOEIC IP テスト（オンライン）を受験させたのは 96 名であったが，その内 4 名は（ここでの議論とは無関係な）アンケートに回答しなかったという理由で分析から除外されている．

年度の全学の新入生が TOEIC テスト（オンライン）を受験していたら，そのトータルスコアの平均は，2021 年 4 月に GTEC テストと TOEIC IP テスト（オンライン）の両方を受験した 92 名の TOEIC トータルスコア平均（約 540 点）とほぼ同じであっただろうと寺西らは推測した．このスコアは，2014–2017 年のマークシート方式の TOEIC テストのトータルスコア平均（約 472 点）と比べて 70 点近くも高い．

寺西らは，この 2 つの TOEIC トータルスコア平均（2014–2017 年のマークシート方式のテストの全学平均と，2021 年のオンライン方式のテストの全学平均推定値）から，従来の TOEIC IP テストと TOEIC IP テスト（オンライン）のスコアの関係を明らかにしようと試みている．ただし，2014 年から 2017 年までの 4 月 TOEIC IP テストと 2021 年 4 月の TOEIC IP テスト（オンライン）の受験者は異なっているので，トータルスコア平均を単純に比較することはできない．そのため，寺西らは 2018 年から 2021 年までの 4 月実施 GTEC テストスコアデータを援用して，マークシート方式の TOEIC IP テストを実施していた時期から TOEIC IP テスト（オンライン）を実施した 2021 年まで，新入生の英語力に大きな変化はないという推測を行った．その根拠として寺西らが挙げたのは，2018 年と 2019 年に実施された全学 GTEC テストのトータルスコア平均が約 241 点であったのに対し，2020 年と 2021 年に実施された全学 GTEC テストのトータルスコア平均が約 244 点であったことである．「2018–19 年と 2020–21 年の GTEC の平均スコアの差が，……僅か 3 点であることから，近年の新入生の英語力に大きな変動はないものと推測される」（寺西ら，2021，p. 13）．この，テスト受験者の英語力に大きな変動はないという推測に基づいて，寺西らは，従来のマークシート方式の TOEIC IP テストよりも，TOEIC IP テスト（オンライン）の方が高いトータルスコアが出るとした．

また，寺西らは，トータルスコアだけでなく，リスニングスコア，リーディングスコアについても平均点を比較し，いずれにおいても TOEIC IP テスト（オンライン）の方が従来のマークシート方式のテストよりも高いことを見出しており，「オンライン試験の方がマークシート式より点数が出やすいということは確認できたと言えるだろう」（寺西ら，2022，p. 14）と結論付けた．

この岡山大学のグループの分析結果は，1 節で見た 2019–2020 年度の金沢大学総合教育部新入生の 4 月 TOEIC IP トータルスコア平均の推移（図 1）とも合うと考えられる．金沢大学総合教育部新入生の 2020 年 4 月の TOEIC IP テスト（オンライン）のトータルスコア平均が，従来の TOEIC IP テストと比べ（岡山大で 70 点近く高く出たように）約 60 点高く出たのであれば，2020 年度は 2019 年度とほぼ同じ英語力の学生が入学し，1 年間の英語力の伸びの大きさもほぼ同じであったということになるので，2020 年度に急に高い英語力の学生が入学し，1 年間の伸びは 3 分の 1 に急減したという不自然な想定をしなくて済む．

ただ，岡山大グループのこの研究には，不十分な点もいくつか挙げられる．まず，この研究では，2018–2021 年の GTEC テストのスコアに大きな変動が見られないことから，TOEIC テストのスコアも大きな変動をしていないはずだとされている．しかし，GTEC テストと TOEIC テストは異なる英語力を測定するテストである．そのトータルスコア間の相関は

0.518（寺西ら，2021，p. 10）であり，これは TOEIC テストのリスニングスコアとリーディングスコアの相関（約 0.8[6]）や，トータルスコアとリスニングスコア，リーディングスコアの相関（0.9 よりも大きい）と比べるとかなり低い．そのため，GTEC テストのスコアに大きな変動が見られないからといって，TOEIC テストのスコアが大きな変動をしていないとは言い切れない．

また，GTEC テストのスコアに大きな変動が見られないという主張についても，GTEC テストのスコアの毎年の変動を見るのではなく，2018 年・2019 年の平均と 2020 年・2021 年の平均を比べているのは問題であろう．2 年分の平均の比較では，年変動があっても見過ごされてしまうからである．

更に，岡山大グループが主張するように GTEC テストのスコアに大きな変動が実際なかったとしても，そのデータからわかることは，2018 年から 2021 年の間に英語力に大きな変動がなかったということである．岡山大学でマークシート方式の TOEIC テストを全学統一試験として実施していた 2014–17 年から 2021 年まで英語力に大きな変動がなかったと考えるのは外挿による推測であり，裏付けとなるデータが提示されているものではない．

以上の点を考えると，TOEIC IP テスト（オンライン）は従来の TOEIC テストよりも高いスコアが出やすいという岡山大グループの主張は，データにより十分に裏付けられているとは言いがたく，更にデータを収集して検討する必要があると考えられる．

3. 等化した推定能力値を説明変数として用いた TOEIC IP テストスコアの分析

3.1. 分析の手法

本研究では，2019 年から 2021 年に金沢大学で実施した，4 月 TOEIC IP テスト（2020 年のみオンライン）のスコアデータを用いて，従来のマークシート方式の TOEIC IP テストと TOEIC IP テスト（オンライン）のスコアの意味が同じかを調べた．ただし，そのスコアデータは，実施年により受験者が異なるため，単純な比較はできない．そこで，同一尺度での受験者の英語能力を知るために，ほぼ全 1 年生を対象に実施している第 2 クォーターの英語科目「TOEIC 準備 II」の期末試験（以下，Q2 期末試験）のデータを参照した．Q2 期末試験は，TOEIC テストのリーディングセクションに準拠したテストであり，等化可能なように作成してあるため，異なる年度の受験者でも同一の尺度で英語能力を測定できる．4 月 TOEIC IP テスト受験時から Q2 期末試験時までの英語能力の変化が受験者間で大きく違わないと仮定すると，Q2 期末試験の等化済みスコア（能力値）がほぼ同一の受験者について，従来の TOEIC IP テストのスコアと TOEIC IP テスト（オンライン）のスコアを比較することによって，TOEIC IP テストの実施形態によってスコアに差が生じているか否かを調べることができる[7]．統計手法としては，以下で説明するように，Q2 期末試験の等化済み能力値

[6] TOEIC テストのリスニングスコア，リーディングスコア，トータルスコア間の相関については，例えば ETS（2022, p. 25）を参照のこと．

[7] TOEIC IP テストのスコアとしてはトータルスコアを使用する．Q2 期末試験は TOEIC テスト

と 4 月 TOEIC IP テストのトータルスコアを用いた重回帰分析を採用した.

以下では，3.2 節～3.4 節で 4 月 TOEIC IP テスト，Q2 期末試験とその等化，分析の対象者について説明したのち，3.5 節で重回帰分析の結果を示す.

3.2. 4 月 TOEIC IP テスト

金沢大学では，2019 年から第 1 クォーターの授業開始後 2 週目の土曜日に総合教育部入学生を対象に TOEIC IP テストを実施している．表 2 に，その TOEIC IP テストの 2021 年までの実施状況を示す．2020 年のオンライン方式の IP テストのみ実施日が 4 月下旬にずれ込んでいるが，これはその年の 3 月頃から新型コロナウイルス感染症の罹患者数が急増し，新型コロナウイルス対策が強化されていく中で，第 1 クォーターの開始が 4 月 20 日まで遅れたことが原因である.

表 2. 4 月実施 TOEIC テストの実施状況（2019～2021 年）

実施日	実施形式	受験者数（人）
2019 年 4 月 13 日	マークシート方式	129
2020 年 4 月 25 日～5 月 8 日	オンライン方式	130
2021 年 4 月 10 日	マークシート方式	137

3.3. Q2 期末試験とその等化

金沢大学では，2016 年から第 2 クォーターの英語科目「TOEIC 準備 II」の期末試験として，従来のマークシート方式の TOEIC テストのリーディングセクションと同一の内容・形式（100 問）・試験時間（75 分）の共通テストをほぼ全ての 1 年生を対象に実施している. この試験は, 2020 年も従来どおりのマークシート方式で行った. 表 3 に, 2017 年および 2019 年から 2021 年までの Q2 期末試験の実施状況を示す.

表 3. Q2 期末試験の実施状況（2017, 2019～2021 年）

実施日	受験者数（人）
2017 年 8 月 1 日～8 月 4 日	1,673
2019 年 7 月 31 日～8 月 6 日	1,671
2020 年 8 月 12 日～8 月 18 日	1,679
2021 年 8 月 2 日～8 月 6 日	1,682

のリーディングセクションに準拠したテストであり，リスニングセクションを含んでいないので，Q2 期末試験のデータから推定される英語能力は完全にはトータルスコアに対応しないが，ETS（2022, p. 25）によると，TOEIC テストのリーディングセクションのスコアとトータルスコアの相関係数は 0.95 と高いので，大きな問題にはならないと考えられる.

Q2 期末試験では，基本的に毎年異なる試験問題（フォーム）を使用しているが，フォームの作成の際に，いくつか共通の問題項目（アイテム）を入れており，それによってスコアの等化を可能としている．本分析では，2019 年から 2021 年に実施した Q2 期末試験について，2017 年に実施した Q2 期末試験を基準として等化を行った結果を使用した．

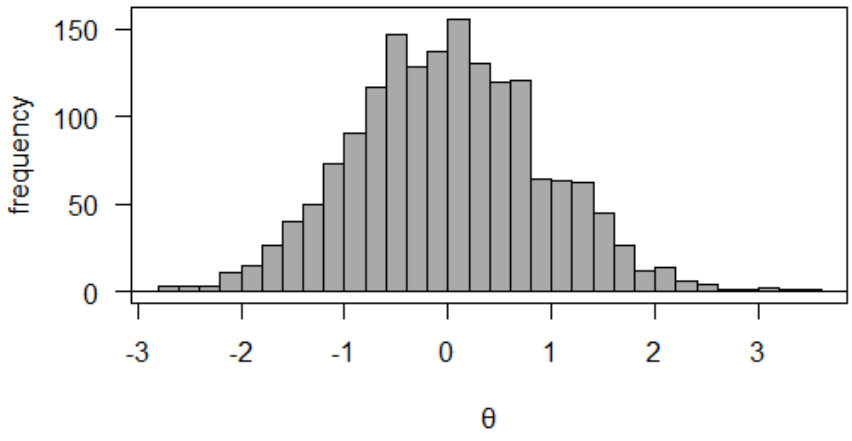
本研究での Q2 期末試験の等化の手順は次のとおりであった．まず，IRTPRO 5（Cai et al., 2020）を使用して，2017，2019～2021 年の試験の解答データの項目分析を行った．そして，(i) 項目困難度（正答率）が 0.1 以下または 0.9 以上であった項目，及び (ii) 項目識別力（項目-テスト得点相関）が 0.1 以下であった項目を，以降の分析の対象から除外した．このスクリーニング後に残った項目数と，その中に含まれる 2017 年実施の Q2 期末試験と共通の項目の数を表 4 に示す．

表 4. Q2 期末試験のスクリーニング後の項目数と共通項目数（2017，2019～2021 年）

試験実施年	スクリーニング後の項目数	2017 年実施 Q2 期末試験との共通項目数
2017 年	71	—
2019 年	85	14
2020 年	74	9
2021 年	76	17

次に，スクリーニング後に残った項目について，項目反応理論（IRT）に基づき，2 パラメーター・ロジスティックモデル（2PLM）を仮定して[8]，パラメーターを IRTPRO 5 で推定した．図 2 に 2017 年実施の Q2 期末試験の全受験者の推定された能力値 θ（平均 0.00，標準偏差 0.92）のヒストグラムを示す．

図 2. 2017 年実施 Q2 期末試験の全受験者の推定能力値 θ の分布



[8] 例えば Şahin and Anıl (2017)によると，30 項目の試験の場合，2PLM で必要な最小標本サイズは 250 である．Q2 期末試験の受験者数は約 1,700 人であるので，それよりも十分大きい．

最後に，R（R Core Team, 2022）の plink パッケージ（Weeks, 2010）を使用して，2017 年実施の Q2 期末試験を基準として Haebara 法による等化を行った．これにより，2019 年～2021 年 Q2 期末試験の受験者の，2017 年 Q2 期末試験を基準として等化した推定能力値を得た．

3.4. 分析の対象者

以下の分析の対象者は，2019～2021 年度の総合教育部入学者で，4 月 TOEIC IP テストと Q2 期末試験の両方を受験した者である[9]．表 5 に，各年度の対象者の人数，4 月 TOEIC IP テストトータルスコアの平均と標準偏差，Q2 期末試験の等化済み推定能力値の平均と標準偏差を示す．

表 5. 回帰分析の対象とした 2019～201 年度総合教育部入学者の TOEIC スコアと能力値

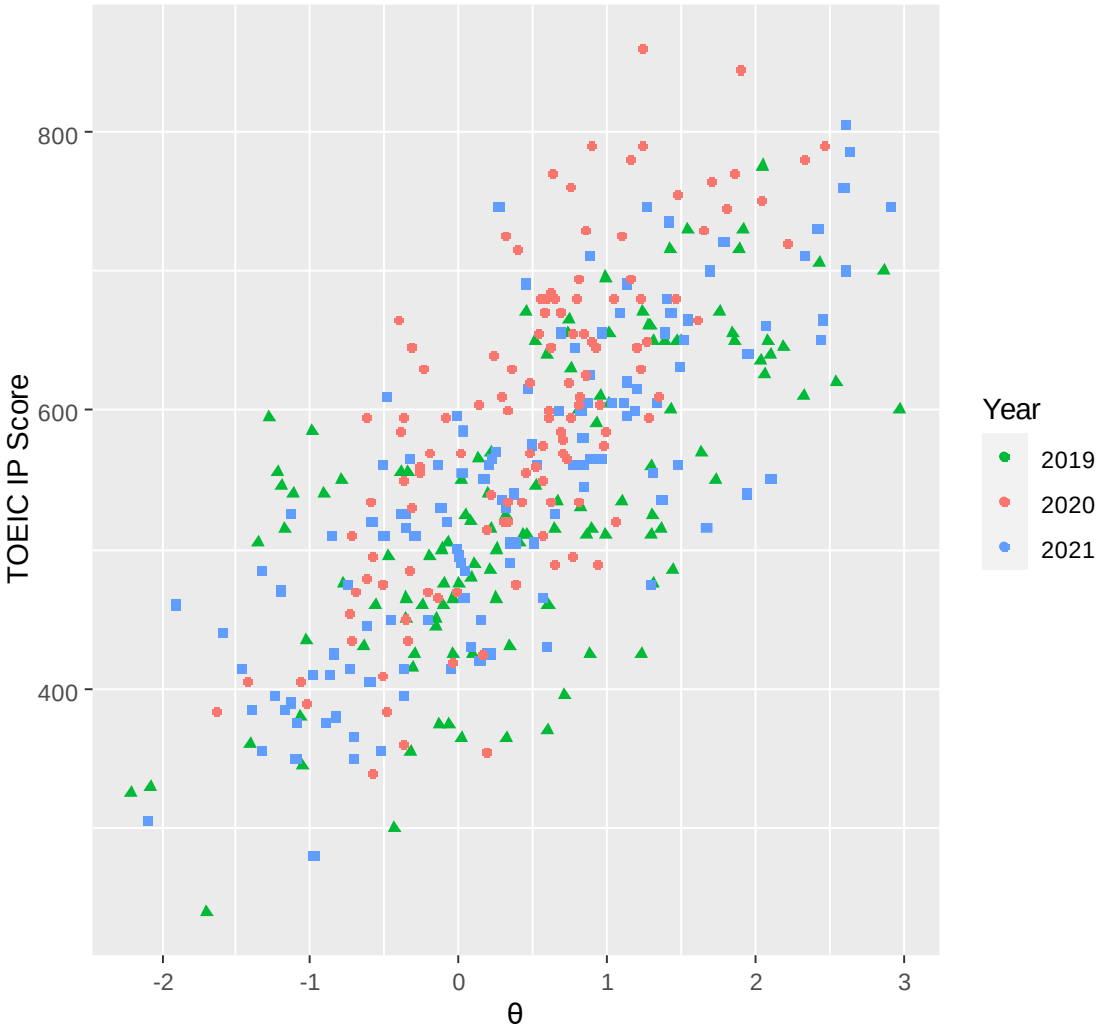
入学年度	人数（人）	4 月 TOEIC IP テスト トータルスコア		Q2 期末試験 等化済み推定能力値	
		平均	標準偏差	平均	標準偏差
2019	120	532.1	105.8	0.48	1.07
2020	122	593.6	112.6	0.46	0.77
2021	129	540.3	111.8	0.37	1.09

3.5. 能力値と TOEIC IP スコアの関係の分析

従来の方式の TOEIC IP テストと TOEIC IP テスト（オンライン）のスコアの差を調べるために，まず，2019～2021 年度の総合教育部入学者の，等化した推定能力値を横軸，4 月 TOEIC IP テストのトータルスコアを縦軸とした散布図を図 3 に示す．図 3 から，従来のマークシート方式の TOEIC IP テストを受験した 2019 年度と 2021 年度の入学者のスコアは同じような分布であるが，TOEIC IP テスト（オンライン）を受験した 2020 年度入学者は能力値 θ が高くなるにつれてスコアが 2019 年度・2021 年度入学者よりも高くなっている傾向がうかがわれる．

[9] ただし，体調不良などで外れ値と見なせるデータは除いた．

図 3. 2019～2021 年度の総合教育部入学者の等化済み能力値 θ と TOEIC IP スコアの分布



次に，従来のマークシート方式の TOEIC IP テスト受験者（2019 年度・2021 年度総合教育部入学者）と，TOEIC IP テスト（オンライン）受験者（2020 年度総合教育部入学者）のそれぞれについて，等化済み能力値 θ を説明変数，TOEIC IP スコア T を目的変数とした回帰式 (1) を用いて単回帰分析を行った．

(1) $$T = \beta_0 + \beta_1 \times \theta$$

図 4 と表 6 は従来のマークシート方式の TOEIC IP テスト受験者の残差分析のプロットと単回帰分析の結果である．図 4 を見ると，データの線形性，それに残差の正規性と均一分散性について特段の問題は見つからないので，回帰分析を適用するための前提をデータが満たしていると仮定してよい．

図 4. 残差分析の結果（従来のマークシート方式の TOEIC IP テスト受験者）

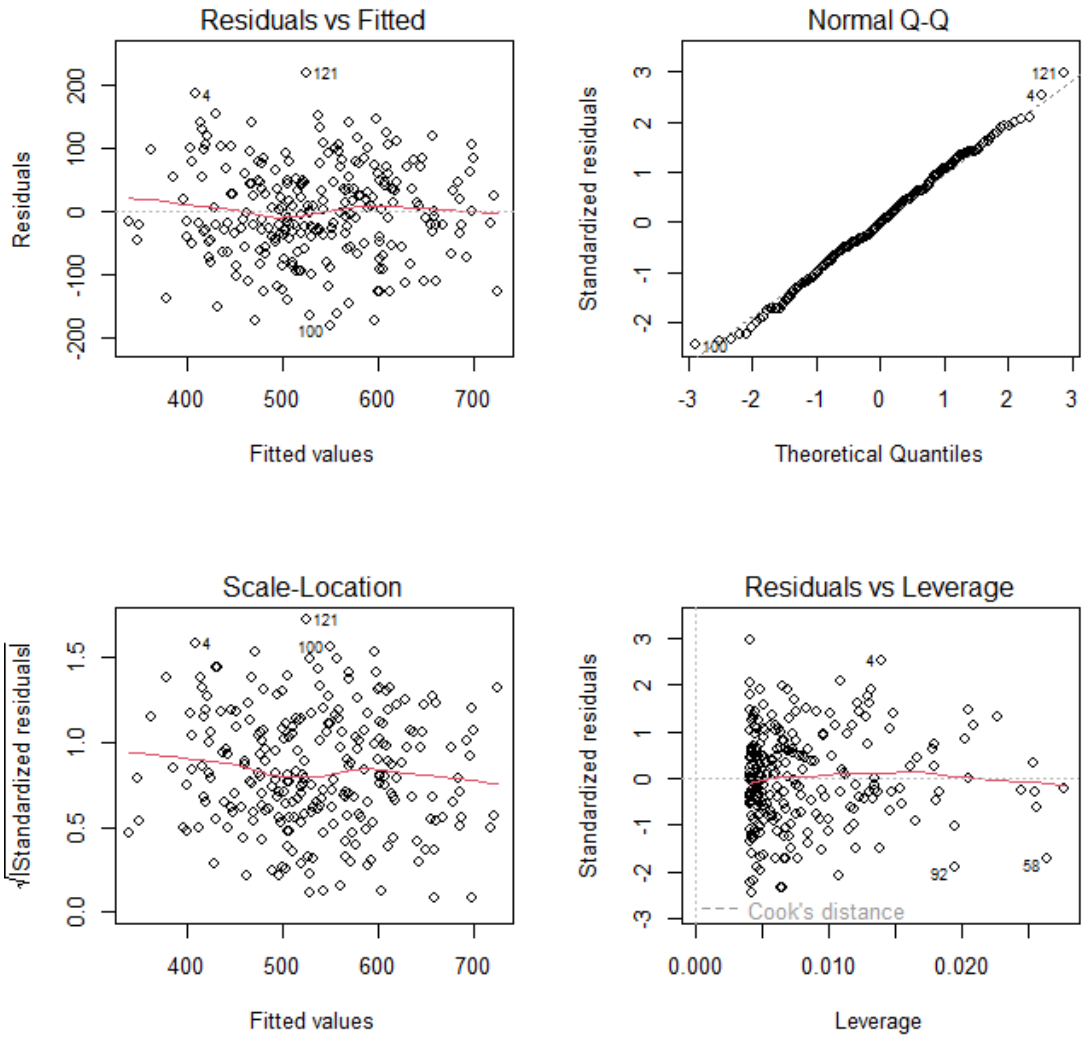


表 6. 単回帰分析の結果（従来の形式の TOEIC IP テスト受験者）

	回帰係数	標準誤差	*p* 値
θ	74.3	4.3	< .001
切片	505.2	5.0	< .001

R^2 = .54

図 5 に，TOEIC IP テスト（オンライン）受験者の残差分析のプロットを示す．図 5 を見ると，データの線形性，残差の正規性には問題がないが，予測値に対する標準化した残差の絶対値の平方根のプロットが凸になっているように見えるので，（予測値の両端近辺のデータが少ないことによるものとも考えられるが）不均一分散に対して頑健な標準誤差を使用した回帰分析を行った．その結果を表 7 に示す[10]．

[10] 念のため，残差の均一分散性を仮定した通常の回帰分析も行ったが，結果はほぼ同じであっ

図 5. 残差分析の結果（TOEIC IP テスト（オンライン）受験者）

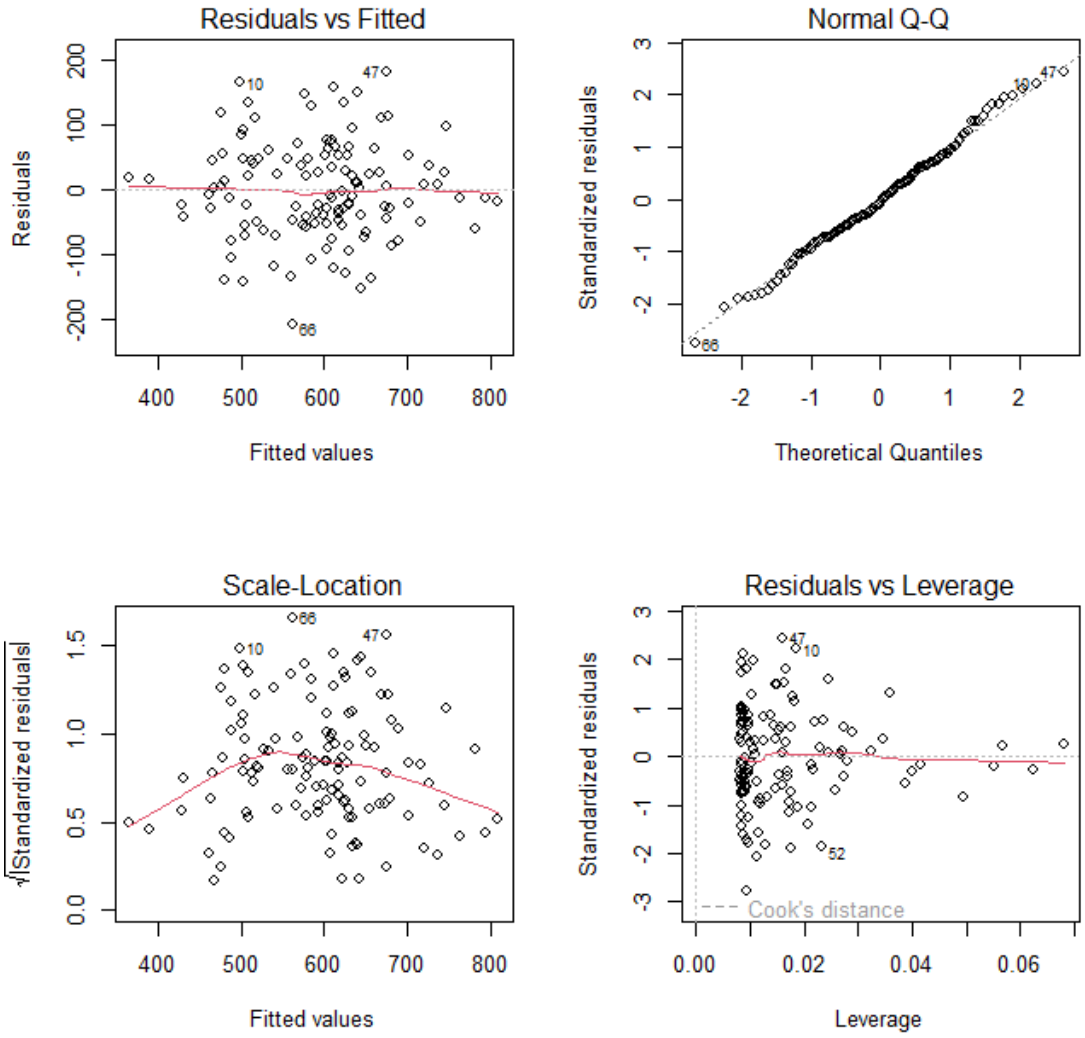


表 7. 単回帰分析の結果（TOEIC IP テスト（オンライン）受験者）

	回帰係数	標準誤差	*p* 値
θ	108.4	7.3	< .001
切片	543.8	7.9	< .001

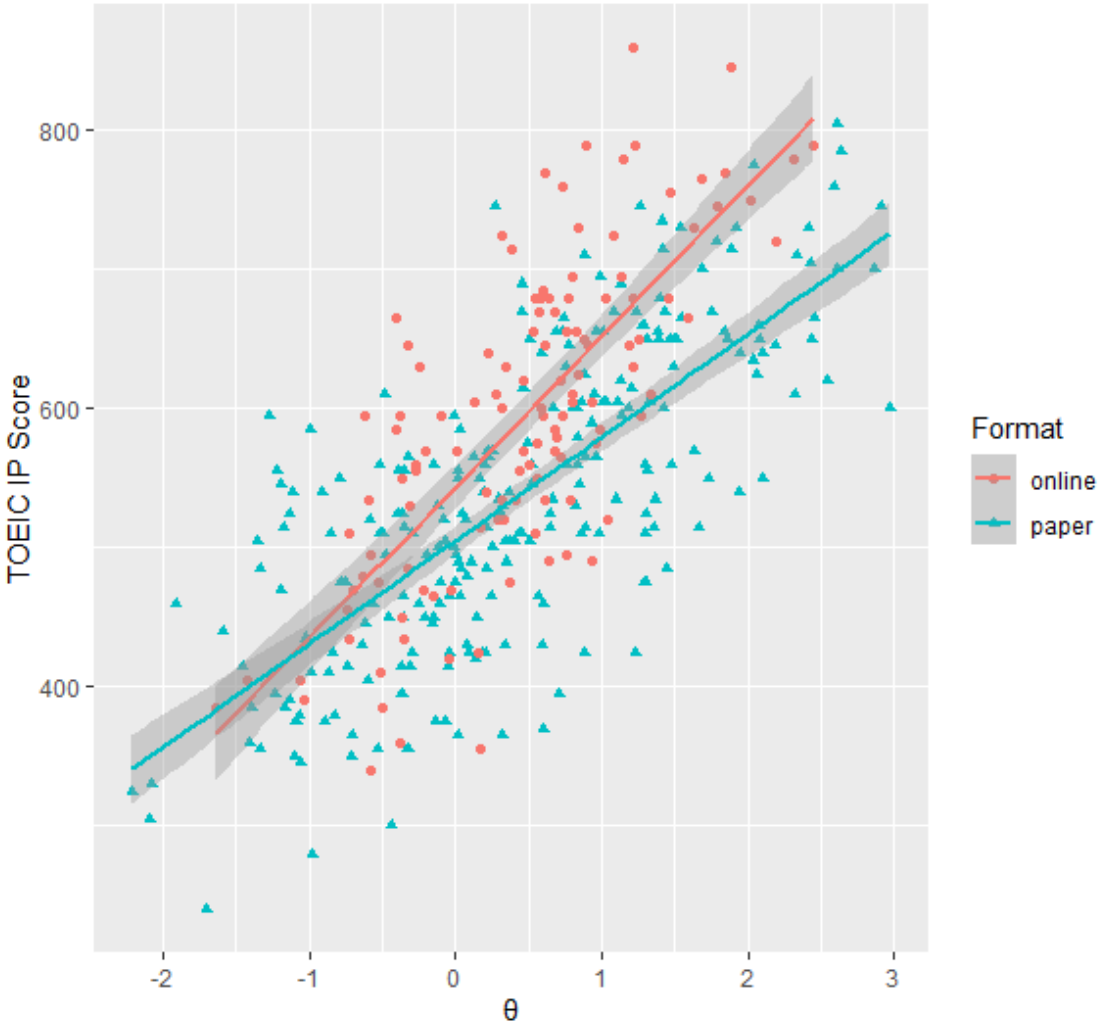
R^2 = .56

以上の分析から得られた回帰直線を，従来のマークシート方式の TOEIC IP テスト受験者と TOEIC IP テスト（オンライン）受験者のデータをプロットした散布図に重ねたグラフが図 6 である．図 6 から，回帰直線の傾きが従来の TOEIC IP テスト受験者と TOEIC IP テスト（オンライン）受験者で大きく異なることが見て取れる．2 つの回帰直線の交点は，(θ, TOEIC IP Score) = (−1.13, 421.1) であるが，それほど能力値 θ が低い受験者は，（図 6 で赤い

た．

丸で示される）2020 年の受験者の中にはほとんどいないということは結果の検討に際して注意が必要である．

図 6. 2019～2021 年度の総合教育部入学者の等化済み能力値 θ と TOEIC IP スコアの分布と回帰直線



注：グレーの範囲は回帰直線の 95%信頼区間を示す．

それから，図 6 から明らかではあるが，能力値 θ の回帰係数の差を検定するために，能力値 θ に加えて，TOEIC IP テストの実施形式を表すダミー変数 Format（従来の TOEIC IP テストを 0，TOEIC IP テスト（オンライン）を 1 とする）と，Format と θ の交互作用項を含む回帰式 (2) による重回帰分析を行った結果を表 8 に示す．Format と θ の交互作用項の偏回帰係数の p 値が 0.001 未満であることから，従来の TOEIC IP テストと TOEIC IP テスト（オンライン）で能力値 θ の（偏）回帰係数が統計的に有意に異なることが確認できた．

(2) $$T = \beta_0 + \beta_1 \times \theta + \beta_2 \times \text{Format} + \beta_3 \times \text{Format} \times \theta$$

表 8. 重回帰分析の結果

	偏回帰係数	標準誤差	95%信頼区間	*p* 値
θ	74.3	4.3	[65.9, 82.8]	< .001
Format	38.6	9.4	[20.1, 57.1]	< .001
Format × θ	34.1	8.5	[17.4, 50.7]	< .001
切片	505.2	5.1	[495.1, 515.4]	< .001

R^2 = .57

これらの結果を総合すると，従来のマークシート方式の TOEIC IP テストでトータルスコアが平均[11]約 420 点よりも高い英語能力の受験者は，英語能力が高いほど，TOEIC IP テスト（オンライン）の方が従来の TOEIC IP テストよりもスコアが上昇する傾向にあったことがわかった．それよりも英語能力が低い受験者については，データが少ないため確かなことはいえない．

4. 考察

前節で，従来の TOEIC IP テストで約 420 点よりも高い平均トータルスコアを取る英語能力がある受験者は，TOEIC IP テスト（オンライン）の方が従来のマークシート方式の TOEIC IP テストよりもスコアが高くなる傾向があることを示した．目安として，2020 年度の総合教育部入学者の等化済み推定能力値 θ の平均（θ＝0.46[12]）を用いて，その能力値を前節の回帰分析で得られた回帰式に当てはめてみると，2020 年度の総合教育部入学者の中で平均的な英語能力を持っていた受験者は，TOEIC IP テスト（オンライン）では約 594 点前後を取っているが，従来のマークシート方式の TOEIC IP テストでは約 540 点前後を取っていることがわかる．英語能力がこの程度の受験者は，TOEIC IP テスト（オンライン）の方が従来のテストよりも平均して約 5，60 点程度高いスコアが得られたということになる．この結果は，1 節で説明した，金沢大学総合教育部の 2019 年度から 2021 年度入学者の 4 月・2 月 TOEIC テストのトータルスコア平均の不自然な推移（図 1）をよく説明できる．つまり，2020 年 4 月の平均スコア 592 点は，2019 年・2021 年のスコアから推測される予想値（例えば 2019 年 4 月 TOEIC テストの平均スコア 530 点と 2021 年 4 月 TOEIC テストの平均スコア 542 点の平均である 536 点）よりも 5，60 点程度高いが，これは，2020 年度入学者の 4 月の英語能力が前年度・翌年度と比べ大幅に高かった（また，それにもかかわらず，2020 年度入学者は大学 1 年次の 1 年間の英語能力の伸びが非常に小さかったので，2 月 TOEIC IP テストスコアには変化が見られなかった）のではない．2020 年度入学者の英語能力は前年

[11] ここでいう平均は，受験者が複数回テストを受験したときのスコアの平均である．受験者が 1 回しかテストを受験しない場合は，それよりも高くなる場合もあれば低くなる場合もある（TOEIC テストの測定誤差については，例えば ETS（2005）を参照）．

[12] 表 5 を参照のこと．

度・翌年度入学者と大きく変わらなかったが[13]，2020 年 4 月は TOEIC IP テスト（オンライン）を実施したために，従来のマークシート方式の TOEIC IP テストよりも平均 5，60 点程度高いスコアが得られたのだという説明ができるのである．

5. まとめ

本研究は，従来のマークシート方式の TOEIC IP テストと TOEIC IP テスト（オンライン）のスコアの関係を調べた．2 種類のテストのスコアの関係の研究のためには，2 種類のテストを，同一の受験者グループを対象に，かつ間をあまり置かずに実施することが理想的ではあるが，それを行うことは難しかったので，本研究では，実施年の異なる 2 種類の TOEIC IP テスト間の比較を，TOEIC テストに準拠した等化可能な期末試験を利用することで行った．

その結果，従来の TOEIC IP テストで約 420 点よりも高い平均トータルスコアを取れる受験者は，TOEIC IP テスト（オンライン）の方がスコアが高くなる傾向にあることが明らかになった．また，英語能力が（データを収集できた範囲内で）高いほど，TOEIC IP テスト（オンライン）のスコアと従来の TOEIC IP テストのスコアの差が大きくなる傾向があることも明らかになった．

参考文献

Cai, L., Thissen, D., & du Toit, S. H. C. (2020). *IRTPRO 5: Flexible, multidimensional, multiple categorical IRT modeling* (Version 5.20). Vector Psychometric Group.

ETS. (2005). TOEIC® technical manual. http://www.toeic.cl/down/toeic_tech_man.pdf

ETS. (2022). TOEIC® score user guide: TOEIC® Listening & Reading test & TOEIC® Listening & Reading test: MSA format. https://etswebsiteprod.cdn.prismic.io/etswebsiteprod/0d46bc04-e14d-4a9f-aadb-d487440318f5_MAN020_TOEIC-LR_TOEIC-LRMSA_SCOR_IP_HOUS_PBT_OBT_31012022.PDF

ETS Global. (2022). TOEIC® Listening and Reading test. https://www.etsglobal.org/fr/en/test-type-family/toeic-listening-and-reading-test

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, *22*, 144–149.

橋本将 (2019).「金沢大学 1 年生の英語学力の変化（Ⅰ）:「TOEIC 準備Ⅱ」科目の共通期末試験の共通項目による等化」『外国語教育フォーラム』*13,* 51–57.

国際ビジネスコミュニケーション協会（IIBC）(2020).「場所と時間を問わずに活用できる IIBC のオンラインプログラム」*IIBC Newsletter, 141,* 2–9. https://www.iibc-global.org/hubfs/library/newsletter/data/pdf/iibc_newsletter-141.pdf

[13] 実際，表 5 からわかるように，Q2 期末試験の等化済み推定能力値の平均は 2019 年度入学者と 2020 年度入学者でほぼ同一であった．

R Core Team. (2022). *R: A language and environment for statistical computing* (Version 4.2.2). R Foundation for Statistical Computing. https://www.R-project.org/

Şahin, A., & Anıl, D. (2017). The effects of test length and sample size on item parameters in item response theory. *Educational Sciences: Theory & Practice, 17,* 321–335.

寺西雅子，大年順子，劔持淑，荻野勝 (2021).「TOEIC L&R オンライン試験と GTEC テストの相関関係に関する一考察」『岡山大学全学教育・学生支援機構教育研究紀要』*6,* 8–19.

Weeks, J. P. (2010). plink: An R package for linking mixed-format tests using IRT-based methods. *Journal of Statistical Software 35*(12), 1–33. https://www.jstatsoft.org/v35/i12/