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

The Change of English Proficiency of Kanazawa University's First-year Students (I): Common-item Equating of Final Examination Scores of the TOEIC Preparation II Course

橋本 将 *
Masashi HASHIMOTO

Abstract

Kanazawa University has made TOEIC Preparation and English for Academic Purposes courses mandatory for its first-year students since academic year 2016/17. In TOEIC Preparation courses, students are required to take a common final examination at the end of Quarters 1 to 3 and a TOEIC L&R IP test at the end of Quarter 4. In this study, the scores of the Quarter 2 (Q2) final examinations administered in 2017 and 2018 were equated by common items using an Item Response Theory-based method. It was found that the examinees of the 2018 Q2 final examination outperformed those of the 2017 Q2 final examination, although the mean of the raw scores of the 2018 Q2 final examination was lower than that of the 2017 Q2 final examination. It indicates that the 2018 Q2 final examination was more difficult than the 2017 Q2 final examination. To reduce the influence of variation of difficulty of examinations on final grades, it would be desirable to develop a grading method which does not heavily rely on unadjusted raw scores of final examinations for the TOEIC Preparation classes.

金沢大学は 2016 年度に「TOEIC 準備」科目と「English for Academic Purposes」科目を 1 年生の英語の必修科目として導入した．「TOEIC 準備」科目では，第 1 クォーターから第 3 クォーターの終わりに共通期末試験を，第 4 クォーターの終わりに TOEIC L&R IP テストを実施している．本研究では，2017 年度と 2018 年度の第 2 クォーターの共通期末試験について，項目反応理論を用いて共通項目による等化を行った．素点は 2018 年度の方が 2017 年度よりも全体に低かったが，等化によって推定された能力値は 2018 年度の方が高かったことがわかった．このことは，2018 年度の第 2 クォーターの共通期末試験は 2017 年度のものよりも問題が難しくなっていたことを示している．現在の最終成績の計算方法では，最終成績は期末試験の素点に大きく依存しているが，試験問題の難易度によって最終成績が左右されることのないように，最終成績の計算方法の改善策の検討が今後必要であろう．

1. はじめに

金沢大学では，2016 年度にクォーター制が採用されるとともに共通教育の英語科目が再編され，共通シラバスを使用する「TOEIC 準備」科目と「English for Academic Purposes」科

* 金沢大学国際基幹教育院外国語教育系

表 1　2017 年度と 2018 年度の Q2 共通試験の実施日，受験者数，平均点，標準偏差

	実施日	受験者数（人）	平均点	標準偏差
2017 年度試験	7 月 28 日–8 月 3 日	1,670	55.8	9.76
2018 年度試験	7 月 31 日–8 月 3 日	1,679	54.2	8.66

目が 1 年生の英語の必修科目となった．「TOEIC 準備」科目は、TOEIC Listening & Reading Test (TOEIC L&R) を利用して，学生が大人として読む・聞く際に困らない英語力を養成する科目で，TOEIC L&R のリスニング・セクションを扱う「TOEIC 準備 I」が第 1 クォーターに，TOEIC L&R のリーディング・セクションを扱う「TOEIC 準備 II」が第 2 クォーターに開講され，第 3・第 4 クォーターにはリスニングとリーディングのどちらのセクションも扱う「TOEIC 準備 III」と「TOEIC 準備 IV」がそれぞれ開講される．また，「TOEIC 準備 I・II・III」では各クォーター末に全学共通の試験を期末試験として実施し，「TOEIC 準備 IV」では第 4 クォーターの授業終了後に TOEIC L&R IP テスト（以下，「Q4 TOEIC-IP 試験」と呼ぶ）を一斉実施して，それらの共通試験の結果を成績評価の 80%に使用することによって客観的な成績評価を行えるようにしている．

本研究では，2017 年度と 2018 年度の第 2 クォーターの全学共通試験（以下，「Q2 共通試験」と呼ぶ）の結果の比較を行った．

表 1 に 2017 年度 Q2 共通試験と 2018 年度 Q2 共通試験の実施日，受験者数，平均点，それに標準偏差を示す．「TOEIC 準備」科目は金沢大学 1 年生のほぼ全員が受講するため，受験者数は約 1,700 人と多い．この表からわかるように，2017 年度から 2018 年度に Q2 共通試験の素点の平均は 55.8 点から 54.2 点に 1.6 点下降した．これは統計的に有意な変化であった（$p < .001$，Hedges' $g = .17$，95%信頼区間 [1.0, 2.3]）．そして，2017 年度と 2018 年度の受験者の素点の分布は図 1 に示す通りであった．この図で 2018 年度（破線）と 2017 年度（実線）を比べると，2018 年度は前年よりも素点が約 60～80 点の受験者が減少し，約 40～60 点の受験者が増加していたことがわかる．

この素点の下降は，必ずしも 2018 年度の受験者の（TOEIC テストで測定される）英語学力が低下したことを意味しない．「TOEIC 準備」科目の共通試験は難易度ができるだけ変化しないように注意を払って作成されているが，難易度の変化をあらかじめ完全に防ぐことはできない．そのため、表 1，図 1 で示した素点の下降について，それが学生の英語学力が低下したために起こったのか，それとも共通試験の難易度が上がったために起こったのかが判別できないのである．

全学共通試験を行うことで，同学年の学生間での英語学力の比較は素点でできるようになったが，異なる学年は基本的に異なる全学共通試験を受験するため，異なる学年の学生間での英語学力の比較については素点ではできないことは変わらない．そこで，表 1，図 1 で示した素点の下降が起こった理由について，上で述べた二つの理由のどちらが正しいのかを明らかにするためには，2017 年度の Q2 共通試験と 2018 年度の Q2 共通試験について等

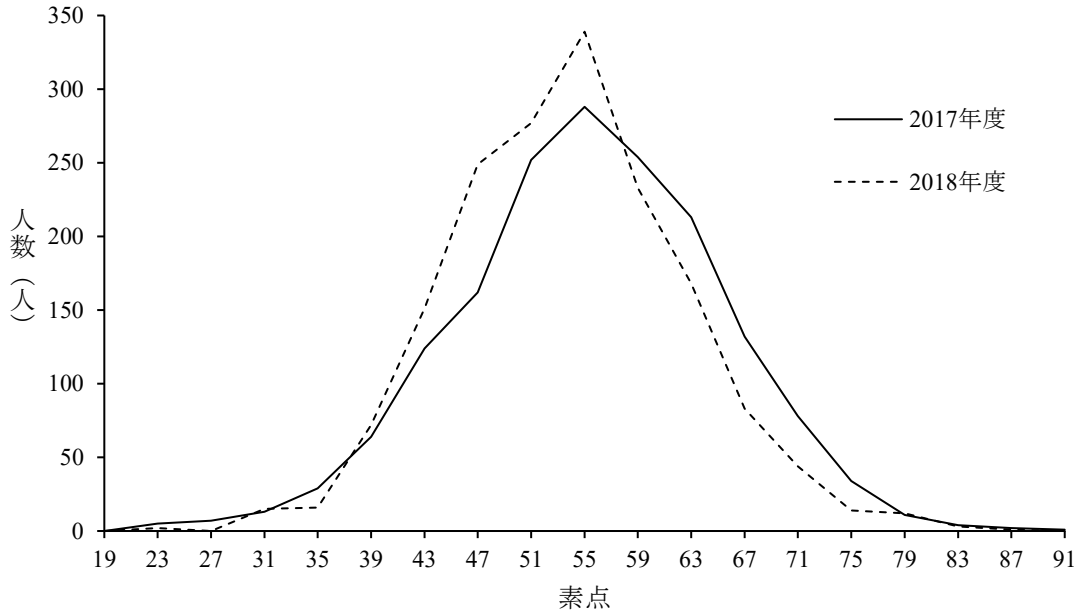

図 1　2017 年度と 2018 年度の Q2 共通試験受験者の素点の分布

化 (equating) を行うことが必要となる．等化とは，同じ仕様の試験について，受験者の能力の比較を可能とするように，受験者の得点（能力値）を一つの尺度上に換算する統計処理のことである．等化には，2 つの試験のどちらも受験した受験者の解答を手掛かりとする方法や 2 つの試験中の共通の項目[1]に対する解答を手掛かりとする方法があり，また，解答データとして素点を用いるもの（等パーセンタイル法など）と解答パターンを用いるもの（項目反応理論）があるが，本研究で分析した 2017 年度の Q2 共通試験と 2018 年度の Q2 共通試験については，共通項目が十分にあったことと，金沢大学 1 年生全体を対象とした試験であるために受験者が約 1,700 人と多いことから，本研究では，これら 2 つの試験について，項目反応理論を用いた共通項目による等化を試みた．

2. 方法

2.1. Q2 共通試験について

第 2 クォーターに開講される「TOEIC 準備 II」の共通期末試験は，TOEIC L&R のリーディング・セクションに準拠しており，項目数は 100 で，試験時間は 75 分間である．TOEIC L&R は，公開テストについては 2016 年 5 月から，IP（団体特別受験制度）テストについては 2017 年 4 月から出題形式が一部変更されたが，「TOEIC 準備」科目の共通期末試験は，2017 年度実施試験は出題形式が変更される前の旧形式で作成され，2018 年度実施試験は新形式で作成されている．厳密には，等化は同じ仕様の試験について行えるものであるが，旧

[1] テスト理論の用語法に従って，ここからは個々の問題を項目と呼ぶことにする．

形式と新形式の違いは大きくないため，本研究ではその違いを無視して等化を行った.

2.2. 項目反応理論を用いた共通項目デザインによる等化

2017 年度 Q2 共通試験と 2018 年度 Q2 共通試験には，共通項目が 100 項目中 22 項目含まれていたので，共通項目デザインによる等化を行った．また，等化には項目反応理論を使用し, 受験者数が約 1,700 人と比較的多いため, 2 パラメーター・ロジスティック・モデル (two-parameter logistic model, 2PLM) を用いることにした.

具体的な手順は，まず，項目分析を行って，正答率が非常に高い（0.9 よりも大きい）項目と非常に低い（0.1 未満）項目を除外し，更に項目合計相関 (item-total correlation) が非常に低い（0.1 未満）項目を除外した．これらの項目を除外するのは，これらの項目をデータに入れたまま項目反応理論に基づくパラメーターの推定を行うと，推定の精度が下がるためである.

次に，2017 年度と 2018 年度のそれぞれの Q2 共通試験について，2PLM の項目パラメーター（各項目の識別力パラメーターと困難度パラメーター）を周辺最尤推定法 (marginal maximum likelihood estimation) を用いて推定した．そして，それによって得られた項目パラメーターの推定値を使って，受験者の能力パラメーターを期待事後 (expected a posteriori, EAP) 推定法によって推定した．これらの推定には，IRTPRO 4.2 (Cai, Thissen, & du Toit, 2017) を使用した.

それから，2017 年度 Q2 共通試験を基準として，2018 年度 Q2 共通試験のデータを等化した．具体的には，まず，2018 年度 Q2 共通試験に含まれる共通項目の項目特性関数 (item characteristic function, ICF) の和が 2017 年度 Q2 共通試験に含まれる共通項目の項目特性関数の和にできるだけ一致するように等化係数の推定を行った（Stocking-Lord 法 (Stocking & Lord, 1983) による等化）後，それによって推定された等化係数を用いて，2018 年度 Q2 共通試験の項目パラメーターとその受験者の能力パラメーターを 2017 年度 Q2 共通試験の尺度に合うように変換した. 以上の等化の計算の実行には，R 3.4.0 (R Core Team, 2017) の plink パッケージ (Weeks, 2010) を使用した.

3. 結果

項目分析の結果に基づいて項目反応理論でのパラメーター推定に不適切な項目をスクリーニングしたところ，総項目数は，100 項目から，2017 年度 Q2 共通試験では 71 項目に，2018 年度 Q2 共通試験では 67 項目にそれぞれ減った．また，両方の試験に含まれる共通項目は 22 項目から 10 項目に減った.

スクリーニング後，2017 年度 Q2 共通試験を基準に 2018 年度 Q2 共通試験を等化をした結果得られた, 各年度の受験者の推定された能力値の分布の平均と標準偏差を表 2 に示す. 表 2 から，2017 年度 Q2 共通試験の受験者よりも，2018 年度 Q2 共通試験の受験者の方が，能力値の平均が高かったことがわかる.

表 2　2017 年度と 2018 年度の受験者の等化後の能力パラメーターの平均と標準偏差

	受験者数	推定された能力値	
	（人）	平均	標準偏差
2017 年度受験者	1,671	0.00	0.92
2018 年度受験者（等化後）	1,679	0.21	0.82

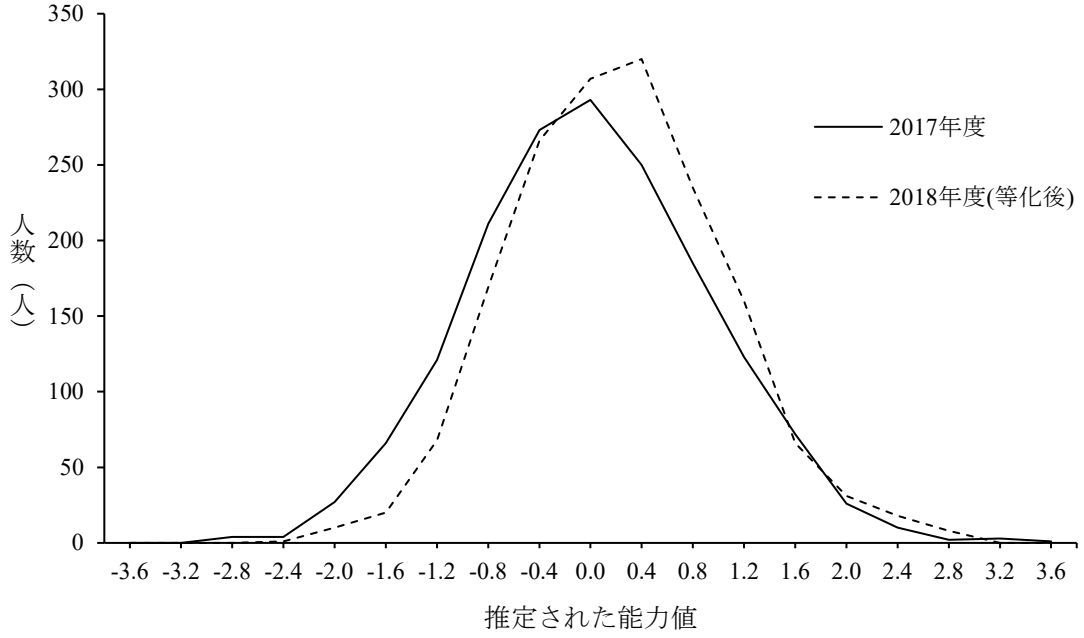

図 2　推定された 2017 年度と 2018 年度の受験者の等化後の能力値の分布

能力値の変化をもう少し詳しく詳しく検討するために，2017 年度 Q2 共通試験の受験者の能力値と 2018 年度 Q2 共通試験の受験者の等化後の能力値の分布を図示したのが図 2 である．図 2 から，2018 年度の受験者（破線）は，その能力値が 2017 年度の受験者（実線）の平均（0.0）より低い受験者が減り，その代わりに能力値が 0.0 から 1.5 あたりまでの受験者が増加していたことがわかる．一方で，能力値が 1.5 よりも大きい受験者の数はあまり増加していない．

4. 考察

素点の平均は 2017 年度 Q2 共通試験よりも 2018 年度 Q2 共通試験の方が 1.6 点低かった（第 1 節参照）が，等化の結果，受験者の能力は 2017 年度よりも 2018 年度の方が上であることがわかった．これは，受験者の能力は上昇したが，試験の難易度がそれ以上に上がってしまったために，受験者の能力の上昇が素点の上昇として反映されなかったことを示して

表 3　Q2 共通試験と Q4 TOEIC IP 試験の両試験の受験者数，Q4 TOEIC IP 試験のリーディングスコアの平均点，標準偏差

	Q2, Q4 両試験 受験者数（人）	平均点	標準偏差
2017 年度	1,596	249.3	60.8
2018 年度	1,587	259.0	59.6

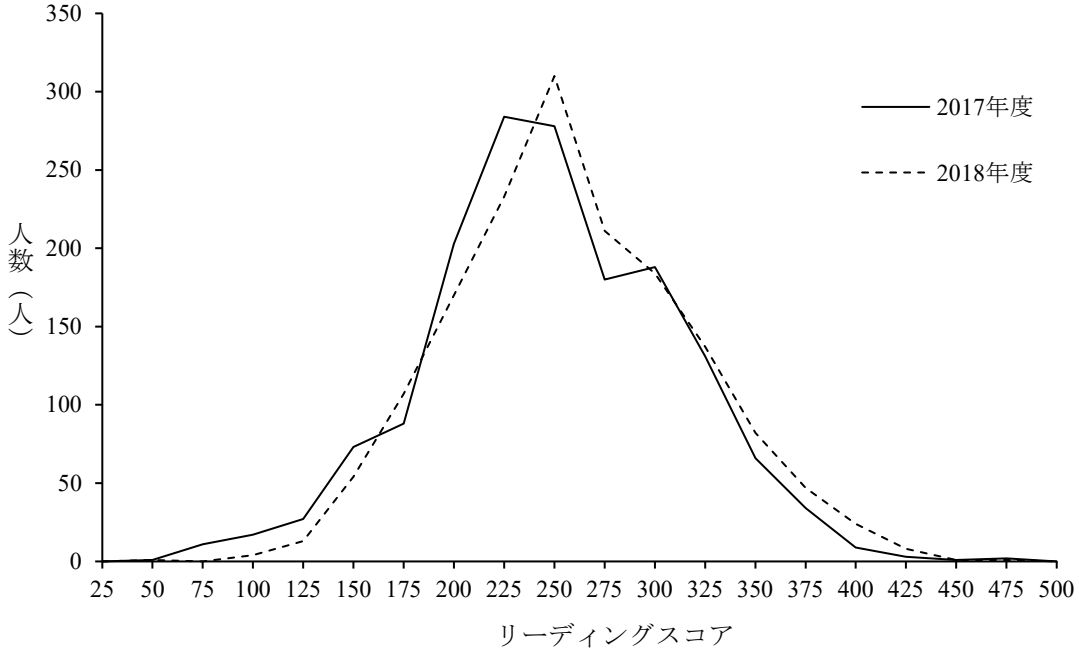

図 3　2017 年度と 2018 年度の Q2 共通試験受験者の Q4 TOEIC IP 試験のリーディングスコアの分布

いる．

参考に，Q4 の終わりに実施される全 1 年生対象の TOEIC IP 試験のスコアの分布を検討した（表 3，図 3）．表 3 は 2017 年度と 2018 年度の Q2 共通試験受験者の内，Q4 TOEIC IP 試験も受験した者の人数と，Q4 TOEIC IP 試験のリーディングスコアの平均点，標準偏差を示したもので，図 3 は，そのリーディングスコアの分布を図示したものである（Q2 共通試験は TOEIC L&R のリーディングセクションと同じ仕様の試験であるため，リーディングスコアを扱った）．これらを見ると，2018 年度の受験者は 2017 年度の受験者よりも Q4 末の時点でも英語能力が高かったことがわかる．

本研究では，Q2 共通試験の素点は 2017 年度と比べて 2018 年度は低下したが，2018 年度の受験者は 2017 年度の受験者よりも英語能力は高かったことを明らかにした．このような

英語能力と素点の対応の破れは，成績の 80%が素点に基づいた絶対評価で決まる現在の TOEIC 準備科目の成績評価方法において問題となる可能性がある．例えば，GPA の良し悪しが影響する選抜の際，難易度が高い共通試験が実施された学年の学生は，そうでない学年の学生よりも現行の成績評価方法の下では TOEIC 準備科目の成績が低くなるため，GPA の比較で不利になってしまう．

項目反応理論を用いて等化を行い，素点の代わりに能力値または真のスコアの推定値を評価に使うようにすればこの問題は解決されるが，共通試験実施から成績入力までに時間がほとんどないクォーターがあるため，実現は難しい．

また，難易度を一定に保つために試験作成者は細心の注意を払っているが，受験者は試験作成者が意図しないところで躓いて間違えることもあり，難易度を試験作成者の想定通りに一定にコントロールすることは難しい．項目バンクを作成して試験の難易度をコントロールするという方策も考えられるが，それにはかなり長い時間（または費用）が掛かってしまう．

現在の成績評価方法の，試験の難易度に依存するという不公平性について，短い時間でそれを減らすには，センター試験で実施されているような得点調整を実施することや，部分的に相対評価を取り入れることなど，評価方法を修正することが効果的であると思われる．

参考文献

Cai, L., Thissen, D., & du Toit, S. H. C. (2017). IRTPRO 4.2 [Computer software]. Skokie, IL: Scientific Software International.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement, third edition* (pp. 221–262). New York, NY: Macmillan.

Partchev, I., Maris, G., & Hattori, T. (2017). irtoys: A collection of functions related to Item Response Theory (IRT). R package version 0.2.1 [Computer software]. Retrieved from https://CRAN.R-project.org/package=irtoys

R Core Team. (2017). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement 7*, 201–210.

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. Retrieved from https://www.jstatsoft.org/v35/i12/