語彙サイズテストによるプレイスメントの試み

Trial of Placement Using Vocabulary Size Test

横　内　裕一郎*

Yuichiro YOKOUCHI

要　旨

本稿は弘前大学教養教育英語科目におけるプレイスメントテスト実施に向けた調査の第一段階として、望月語彙サイズテスト（2003）を用いて測定された受容語彙サイズのクラスレベル平均が現行のクラス編成と整合性があるか、現行のクラス編成方法と潜在ランク理論を用いたクラス編成方法で、どのように編成が変わるかを調査したものである。本研究の結果から、現行のクラス編成でも初級・中級・上級のそれぞれで平均語彙サイズが有意に異なることが明らかになった一方、潜在ランク理論に基づく潜在ランクの推定結果をクラス編成に応用した場合、中級クラスの編成がこれまでと大きく変わる結果が得られた。この結果から、現行のセンター試験の得点を利用したプレイスメントでは中級クラスに該当する学生を適切に弁別できていない可能性が明らかになった。この結果を踏まえつつ、プレイスメントテスト導入に向けてどのような要素を考慮すべきかをまとめた。

キーワード：プレイスメントテスト，語彙サイズテスト，潜在ランク理論

1. はじめに

多くの大学で新入生の英語運用能力を測定し、適切なレベルの授業に入学者を割り振ることを目的としたプレイスメントテストを実施している。プレイスメントテストとは、受験者の能力を測定し、適切なクラスに割り振ることを目的として行うテストのこと（Alderson, Clapham, & Wall, 1995）で、TOEICやTOEFLなどのテスト作成機関が作成したテストを利用するか、各大学・学部が独自に作成した試験を使うことが多い。プレイスメントテストを実施する場合、どのテストを使用して何の能力を測定するのかをまず検討し、テストタスク、対象になる受験者の能力、予算、結果を受験者に提示するまでの時間など様々な要素を検討し、ようやく適切なプレイスメントテストの実施が可能となる。

現在、弘前大学教養教育英語科目では熟達度別クラスを導入しているものの、プレイスメントテストは実施していない。クラス編成の際には大学入試センター試験の英語科目の得点が使用されているが、2020年に大学入試センター試験が廃止されることから、適宜プレイスメントテストを実施して適切なクラス編成を行う準備を進める必要がある。大学入試センター試験が廃止となった後には「高等学校基礎学力テスト（仮称）」や「大学入学希望者学力評価テスト（仮称）」が導入される見通しとなっている

＊弘前大学教育推進機構教養教育開発実践センター
Center for Liberal Arts Development and Practices, Institute for Promotion of Higher Education, Hirosaki University

（文部科学省高等教育局, 2016）。また、大学入試にTOEFLや英検、TEAPなどの外部試験を導入することが推奨されるようになった（文部科学省初等中等教育局国際教育課外国語教育推進室, 2014）ため、新入生の英語運用能力を統一の尺度で比較することが難しくなる。そこで本研究では、将来的に弘前大学において教養教育英語科目のプレイスメントテストを実施することになった場合に備え、プレイスメントテストとしてふさわしいテストを検討することを目的に、その第1段階として弘前大学の学士課程1年生を対象に望月語彙サイズテスト（望月, 1998）を実施し、小泉・飯村（2010）、小泉（2011）と法月（2013）を参考に、4技能の下位概念である語彙の測定を通じて適切なプレイスメントが可能であるかを調査した。また、現行のクラス編成方法と潜在ランク理論に基づくクラス編成で、各レベルの編成がどのようになるのかを検討した。

2. 先行研究

2. 1　プレイスメントテスト実施の実態

各大学のプレイスメントテスト実施状況を調査した研究に、清水（2000）や杉森（2003）などがある。清水（2000）は全国616の国公私立大学を対象に質問紙調査を行い、そのうち200校から得た回答を集計した結果、2000年当時では国公私立大学全体で48％の大学でプレイスメントテストを実施しており、国立大学だけに絞れば34.8％にとどまるものの一定数の大学でプレイスメントテストを実施していたことがわかる。また、杉森（2003）も清水（2000）を受けて同様の調査を行い、全国208校から得た回答を分析している。4年制大学では64％の大学が何らかの統一テストを実施しており、そのうちの55.7%がプレイスメントを目的とした統一テストを行っていることが報告されている。全国を対象に各大学で実施されているプレイスメントテストの実態を一括にまとめた調査はこの他に見当たらないが、各大学で実施されているプレイスメントテストの結果についての報告や使用したテストの妥当性検証が数多く報告されており（例．Kimura, 2009; 小泉, 2011; Yamanaka & Kondo, 2016）、プレイスメントテストの実施に対する教育者の関心は高い。

プレイスメントテストを実施する場合、既存の外部試験を利用することが第一選択肢となるが、それ以外の選択肢として、各大学や学部で独自に作成した試験を使用したり、入学試験の成績を使用したりすることも可能である。テストを作成したり運用したりする場合には、その問題の妥当性と信頼性が確保されていることが大前提だが、プレイスメントテスト実施の際には妥当性や信頼性だけでなく、実用性、特に結果をできるだけ早く得ることができるかどうかを重視する必要がある。プレイスメントテストの場合、年度の始めから1週間から10日程度でクラス編成の結果を通知しなければならないため、結果を素早く伝えられるようにしなければならない（吉田, 2009）。テスト作成機関が作成したテストは、妥当性と信頼性が検証済みであり、費用はかかるもの、結果がすぐ得られることに利点がある。一方、各大学で独自の試験を作成する場合、大学が真に測りたい能力を限定して試験を作成できるものの、短い期間で採点を完了させなければならず、担当者にかかる負担は大きい。また、プレイスメントテスト実施後にテストの妥当性と信頼性の検証がなされるべきだが、大学・学部独自の試験を行う場合には専門家の助言や補助が必要となるだろう。

2. 2　語彙サイズテスト

テストを作成する、既存の試験を使用する、いずれの場合においてもテストの構成概念に何の技能が含まれるか、どのような技能の下位項目が含まれるかを考慮することがテストの運用に向けての第一段階である。語彙は他の4技能の下位技能としてみなされることが多く、各技能を測定するテストの場合、構成概念の1つとして捉えられることが多い。語彙サイズテストは数多く存在し、受容語彙サイズを測定するもの、発表語彙サイズを測定するもの、ESL学習者を対象としたもの、日本人英語学習者を

対象としたものなど様々ある。語彙サイズと各技能の関係を調査した研究は数多くあり、例えば、水本（2006）は独自の語彙サイズテストとTOEICの相関が高かったことから、語彙サイズを習熟度の予測に使うことができると報告している。各技能に目を向けても、それぞれの技能と語彙サイズの関係に強い相関があるという結果が数多く報告されている（例. Laufer, 1992; Mizumoto & Shimamoto, 2008; Tamura, 2011）。プレイスメントとして語彙サイズテストを使用する場合、例えばリーディングを対象としたクラスのプレイスメントの場合Nation (1990) によるVocabulary Levels Testを使用し、ライティングやスピーキングを対象とした場合にはProductive Vocabulary Test (Laufer & Nation, 1999) を使用するなど、編成したいクラスで取り扱う技能によって別の語彙サイズテストを使うことも可能である。日本人英語学習者を対象とした語彙サイズテストとして最も著名なものとして、日本人英語学習者のための語彙サイズテスト（通称：望月語彙サイズテスト）（望月, 1998）がある。このテストをプレイスメントとして使用した結果を報告した研究として小泉（2011）がある。小泉（2011）は私立大学におけるプレイスメントとして望月語彙サイズテストとCASECを使用し、どちらがよりプレイスメントテストとして有効かを調査している。本節で紹介した先行研究はほんの一例に過ぎないが、語彙サイズテストはプレイスメントテストとして適切に機能する可能性がある。そのため、授業計画や英語科目の教育目標から大きく逸脱しないことが確認した上で、何らかの語彙サイズテストを弘前大学の教養教育英語科目のプレイスメントテストとして活用することは選択肢の 1 つとして十分にありえるだろう。

2. 3　潜在ランク理論

プレイスメントテストの結果を解釈し、クラス編成を行う場合、素点をそのまま比較するだけではプレイスメントテスト受験者に対して説得力のある説明ができないことがある。例えば、上位 25％、下位 25％をそれぞれ上級と初級にクラス編成し、残りを中級クラスに割り振るとする。年度ごとに上級と中級、中級と初級の閾値が異なり、例年であれば上級クラスに配分される受験者が 1 点の差で中級クラスに割り振られるというようなことが考えられる。入学試験でも同じように 1 点の差で合否が決定するため 1 点の格差を一概に否定する事はできないが、同じ得点を取った受験者同士でも正答した問題の難易度によって受験者の能力は大きく異なることをテストの作成者や実施者、結果の使用者は把握していなければならない。

Shojima (2007)によって提唱された潜在ランク理論（以前はニューラルテスト理論と呼ばれていた）は古典的テスト理論や項目応答理論に変わる最新のテスト理論の 1 つである。潜在ランク理論は 1 点刻みでテストの結果を示すのではなく、5 段階～ 20 段階程度の比較的少ない順序尺度でテストの結果を解釈することを可能にする手法である。荘島（2010）は、テストの得点は「解像度」の低い、すなわち精度の低い測定道具であると主張している。例えば、体重計は 1 グラムの単位でどのような環境でも他者との明確な比較が可能なのに対して、100 点満点のテストにおける 1 点の違いで他者との比較を行うことができないことから、試験の得点は体重計などに比べて十分な精度の測定道具ではないと言える（荘島, 2008; 2010; 木村, 2009）。この考えに基づき、1 点の違いを根拠にクラス編成を行うのではなく、統計的な処理を行い、受験者のレベル分けをすることが適切なクラス編成の根拠となる可能性がある。

潜在ランク理論は、項目参照プロファイル（Item Reference Profile; IRP）を調査することによってそれぞれの潜在ランクに該当する受験者の解答パターンを判別し、どのような問題で間違いを起こしているのかを分析することができる。その他、テスト参照プロファイル（Test referenced profile; TRP）、潜在ランク（Latent rank）、潜在ランク分布（Latent rank distribution; LRD）、ランク・メンバーシッププロファイル（Rank membership profile; RMP）、ランク・メンバーシップ分布（Rank membership distribution; RMD）、境界カテゴリ参照プロファイル（Boundary category reference profile; BCRP）、項目カテゴリ参照プロファイル（Item category reference profile; ICRP）、といった指標を参考に、受験者の予測得点や特定の潜在ランクにある受験者の特定の項目への正答確率、受験者のレベル、特定の潜在ランクに分布する

受験者数、受験者がそれぞれの潜在ランクに当てはまる確率などが推定される（木村, 2013, 2016; 小泉・飯村, 2010; 小山・木村, 2011; 荘島, 2010, 2015; Shojima, 2007, 2008, 横内, 2014）。現段階で潜在ランク理論を積極的に取り入れてプレイスメントを実施したり、テストサービスを提供している例は筆者の知る限りないが、ことプレイスメントというテストの目的に限っては、潜在ランク理論によるレベル推定は積極的に取り入れられるべき手法の 1 つである。

3. 研究

3. 1　目的

本研究は、弘前大学教養教育英語科目のクラス編成をセンター試験の得点を利用する以外の方法で行う方法を追求することを目的にとして、小泉・飯村（2010）と法月（2013）を参考に、望月語彙サイズテスト（1998）を用いてプレイスメントのパイロットを行った。現行のクラス編成が適切に機能しているかどうかを検証し、より適切なクラス編成方法を検討することを目的に、以下のリサーチクエスチョンを設定し、調査を行った。

RQ1: 現行の熟達度別クラスにおける平均語彙サイズに差は生じるか

RQ2: 潜在ランク理論を用いたクラス編成と現行の熟達度別クラスに差は生じるか

3. 2　参加者

本学の教養英語の授業を受講する学生を対象に望月語彙サイズテストを行った。初級 57 名、中級（下）120 名、上級 58 名が本実験に参加した。そのうち、テスト開始から最後まで出席していた日本人大学生（1 年生）のうち、90% 以上の解答がマークリーダーで適切に読み取れたデータのみを分析対象にした。最終的に初級 53 名、中級（下）99 名、上級 58 名のデータが分析対象となった。本調査への参加者には調査目的を伝えず、実験当時にどの程度語彙の問題に回答できるかに挑戦する課題という形式で実験を行った。これは、実験参加者が事前に授業の成績に関係ない課題であることに気づいた場合、適切なパフォーマンスを得られない可能性が極めて高くなり、適切な解釈ができなくなるためである。なお、本調査のデータ収集に協力してくださった教員には、本調査の目的と概要、実施方法について口頭で説明を行い、本調査が弘前大学英語ワーキンググループの承認を受けた上で実施した調査であることを伝えた上で同意を得た。

3. 3　手法

3.3.1　マテリアル

望月（1998）による望月語彙サイズテストのうち、相澤・望月（2010）に収録された筆記版語彙サイズテストを利用し、受験者の受容語彙サイズを測定した。望月語彙サイズテストを選択した理由として、選択肢が日本語で記載されていることから受験者の英語力のレベルに関わらず適切に試験を実施することが可能だと判断したためである。B4 用紙両面の問題用紙に望月語彙サイズテストの全問 182 問が掲載されるよう調整し、受験者は A4 のマークシートに解答する形式で問題に解答した。

3.3.2　データ収集の手順

2016 年 10 月～ 11 月にかけて、教養教育英語科目の Writing 及び Speaking の各級の授業時に 30 分程度時間をとり、望月語彙サイズテストを実施した。試験は筆者の作成した音声ファイルを元に時間管理がなされ、1 問につき 5 秒の解答時間が与えられた。受験者には解答時にはマークにチェックを入れるに留めるよう指示を行い、試験終了後 5 分程度時間をとってマークを塗りつぶすよう指示を与えた。その後、問題用紙と解答用紙の両方を回収し、Area 61 マークリーダーを用いて採点を行った。ただし、

マークが薄かったり、枠線を大きくはみ出していたりなどして解答を正しく読み取れないデータを 20 件以上含むデータは不良データとして分析から除外した。

3.3.3 分析手法

RQ1 では、素点の合計点と語彙サイズの記述統計を示し、素点を従属変数、在籍クラスのレベルを要因とする一元配置分散分析を行った。RQ2 では、潜在ランク理論のうち、2 値の SOM（Self-organizing map: 自己組織化マッピング）モデルを適用し、分析を実施した。分散分析には R パッケージの psych と anova-kun を用い、潜在ランク理論による分析には Exametrika (Shojima, n.d.) を用いて分析を行った。

4. 結果・考察

4. 1 RQ1: 現行の熟達度別クラスにおける平均語彙サイズに差は生じるか

現在のクラス編成が適切に行われているかを調査するため、語彙サイズテストの正答数（素点）が各クラスレベル間で差が生じるか一元配置分散分析による分析を行った。表 1 がクラス別の語彙サイズテストの平均点と標準偏差、95% 信頼区間の値である。平均点を見る限り、これまでに望月語彙サイズテストを使用してきた研究に比べ低い傾向が見られるが、これは 1 問あたりの解答時間を 5 秒に制限したことで難易度が高まったことが原因であると考えられる。一元配置分散分析の結果、$F(2, 189) = 40.60$, $p < .00$, $\eta^2 = 0.30$ で効果量は大であった。多重比較の結果、初級と中級では $t(189) = 9.01$, $p < .00$, $r = 0.55$、初級と上級では $t(189) = 5.99$, $p < .00$, $r = 0.40$、中級と上級では、$t(189) = 4.49$, $p < .00$, $r = 0.31$ という結果だった。この結果から、クラスレベルごとの平均語彙サイズは初級・中級・上級それぞれで明確に異なることがわかった。

表 1
クラス別の語彙サイズテストの結果

クラス		*N*	*Mean*	*SD*	95% CI
初級	素点	53	54.55	11.29	[43.26, 65.83]
	語彙サイズ	53	2097.97	434.07	[1662.90, 2532.03]
中級	素点	99	69.20	23.12	[46.09, 92.32]
	語彙サイズ	99	2661.62	889.06	[1772.56, 3550.67]
上級	素点	40	87.17	15.00	[72.17, 102.18]
	語彙サイズ	40	3352.88	577.01	[2775.87, 3929.90]

Note. 95% CI means 95% confidential interval.

4. 2 RQ2: 潜在ランク理論を用いたクラス編成と現行の熟達度別クラスに差は生じるか

続いて、受験者の解答パターンから潜在ランク理論に基づくレベル分けを行った上で、その結果が現行のクラス編成方法によるレベル分けとどの程度合致するかを調査した。潜在ランク理論による分析を行うためには、潜在ランクがいくつあるのかを事前に推定する必要がある。現行の方法では、初級・中級・上級の 3 段階に本学学生のレベル分けがなされているため、最小限の潜在ランク数を 3 と設定した。ランク数を増やした場合のほうがモデルにデータが合致する可能性があるため、潜在ランクが 4 の場合と 5 の場合で情報量基準がどのように変化するかを確認した上で分析を実施した。その結果、AIC（赤池情報量基準）はランク数が 4 つの場合で最も数値が小さくなり、BIC（ベイズ情報量基準）と CAIC（一貫性のある赤池情報量基準）では潜在ランクが 3 つの場合で最も数値が小さくなった。一般的にこれらの情報量基準は小さいほどモデルへの適合があるとみなすが、今回は AIC と BIC 及び CAIC

で情報量基準が最小となったランク数が異なったため、ランク数が 3 の場合と 4 の場合いずれのケースでも現行のクラス編成と結果が合致するかを調査した。表 2 は潜在ランク数が 3 の場合と 4 の場合におけるテスト適合度を示したものであり、図 1 は潜在ランク数が 3、4、5 のときの情報量基準の変化を表したものである。

表 2
潜在ランク数が 3 の場合と 4 の場合のテスト適合度

	ランク数 3 の場合		ランク数 4 の場合	
	テスト適合度	RMP に基づく テスト適合度	テスト適合度	RMP に基づく テスト適合度
カイ 2 乗値	1588.238	1501.991	1025.343	955.797
自由度	4004	4004	3822	3822
P 値	1.000	1.000	1.000	1.000
NFI	0.493	0.500	0.657	0.667
RFI	0.493	0.500	0.657	0.667
IFI	1.000	1.000	1.000	1.000
TLI	1.000	1.000	1.000	1.000
CFI	1.000	1.000	1.000	1.000
RMSEA	0.000	0.000	0.000	0.000
AIC	-6419.762	-6506.009	-6618.657	-6688.203
CAIC	-23466.773	-23553.021	-22890.805	-22960.350
BIC	-19462.773	-19549.021	-19068.805	-19138.350

Note. NFI = normed fit index; RFI = relative fit index; CFI = comparative fit index; RMSEA = root mean square error of approximation; AIC = Akaike information criterion; CAIC = consistent AIC; BIC = Bayes information criterion. AIC, CAIC, BIC は小さいほどテスト適合度が高いとみなす.

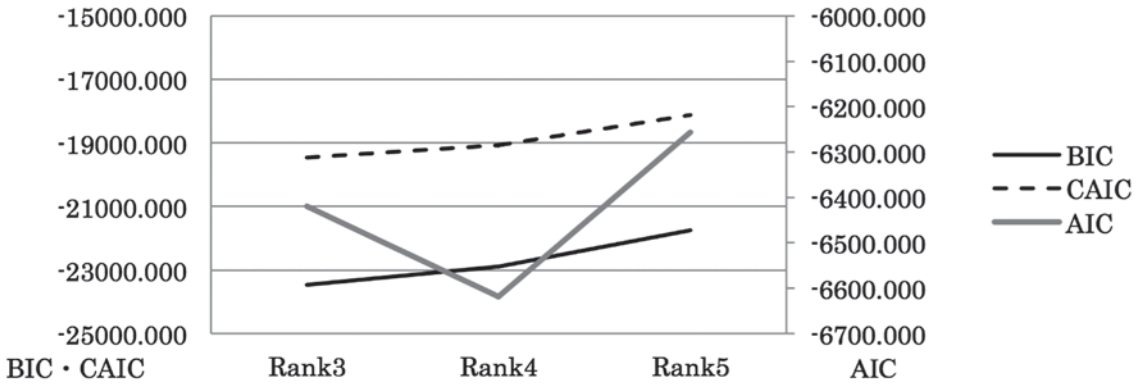

図 1. ランク数ごとの情報量基準.

まず、潜在ランクを 3 とした場合の受験者の分布についてまとめる。表 3 の潜在ランク分布をみると、Rank1（最も熟達度が低い群）に 68 名、Rank2（中位群）に 45 名、Rank3（上位群）に 79 名が当てはまることになる。今回のデータでは中級クラスからの参加者が最も多く 99 名だったが、この分析では Rank2 に当てはまる参加者が最も少なくなった（図 2 参照）。

表 3
潜在ランクを 3 とした場合の各ランク TRP・LRD・RMD（事前分布を指定しない場合）

	Rank 1	Rank 2	Rank 3
テスト参照プロファイル（TRP）	51.901	66.633	86.105
潜在ランク分布（LRD）	68	45	79
ランク・メンバーシップ分布（RMD）	67.253	49.856	74.891
相対 TRP	0.285	0.366	0.473
相対 LRD	0.354	0.234	0.411
相対 RMD	0.350	0.260	0.390

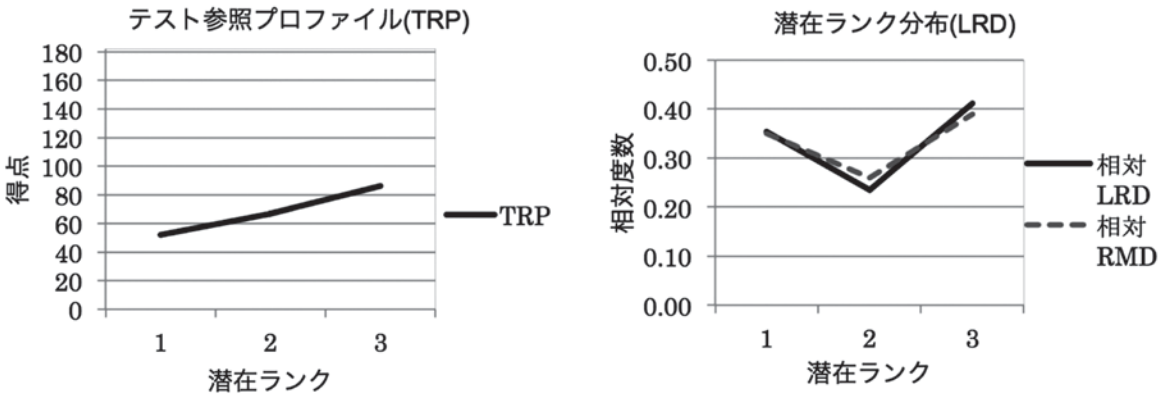

図 2. 潜在ランクを 3 とした場合の TRP（左）・LRD（右）.

今回の結果と現在受験者が所属するクラスとのクロス集計が表 4 になる。表 4 の網掛け部が現在のクラス編成と潜在ランクがマッチしている人数を示したものである。現在初級に所属する学生は全員 Rank1 に該当しており、上級に在籍する学生 40 名のうち 82.5%に及ぶ 33 名が Rank3 に該当する結果となった。つまり、初級と上級に関してのレベル分けは現状の方法で充分なされている可能性が高いことがわかる。一方で中級に関しては Rank1 に 15 名、Rank3 に至っては 46 名にも及ぶ半数近くが上級と同等扱いとなり、現在中級を受講している学生のうち半数は別のレベルに該当していた可能性があることが今回の分析結果から明らかになった。

表 4
クラスレベル別のランク（潜在ランクを 3 とした場合）

	Rank1	Rank2	Rank3	元の人数
初級	53	0	0	53
中級	15	38	46	99
上級	0	7	33	40
合計人数	68	45	79	

Note. $\chi^2(4) = 149.637, p < .00$.

表 5
潜在ランクを 4 とした場合の各ランク TRP・LRD・RMD（事前分布を指定しない場合）

	Rank 1	Rank 2	Rank 3	Rank 4
テスト参照プロファイル（TRP）	52.921	55.673	74.131	87.835
潜在ランク分布（LRD）	53	40	35	64
ランク・メンバーシップ分布（RMD）	53.328	38.644	38.556	61.472
相対 TRP	0.291	0.306	0.407	0.483
相対 LRD	0.276	0.208	0.182	0.333
相対 RMD	0.278	0.201	0.201	0.320

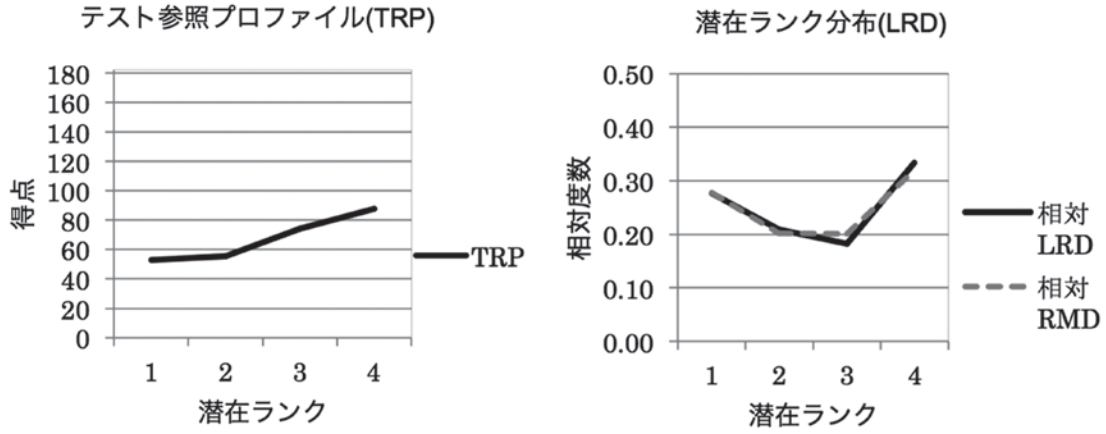

図 3. 潜在ランクを 4 とした場合の TRP（左）・LRD（右）

潜在ランクを 4 とした場合でも同様の結果が得られ、初級の受験者は全員が Rank1 に当てはまる結果になった。一方、上級クラスの学生のうち、32 名が Rank4 に当てはまるため、明確に上級の学生は高い能力を持っていることがわかる（表 6 参照）。

表 6
クラスレベル別のランク（潜在ランクを 4 とした場合）

	Rank1	Rank2	Rank3	Rank4	元の人数
初級	53	0	0	0	53
中級	0	39	28	32	99
上級	0	1	7	32	40

Note. $\chi^2(6) = 231.858, p < .00.$

今回の実験協力者のうち中級の学生は入学時にセンター試験の得点を元にクラス編成された結果、初級と「1 点の違い」で配分された学生と、センター試験を受験せず、推薦入学者だからという理由で中級（下）のクラスに配分された学生で構成される。データ収集の期間が入学後 7 ～ 8 ヶ月経過してからであったことが今回の結果に影響を及ぼしている可能性は否めない。しかし、中級を受講する学生の半

数が他のレベルに割り振られるべきであったかもしれないという事実は重く、適切なプレイスメントテストの重要性を認識した上で教養教育英語科目の計画や編成を行わなければいけないだろう。

続いて、今回の実験を通じて得られた実用性に関する報告をする。吉田（2009）でも言及されていたように、プレイスメントテストの実施の際には信頼性と妥当性だけではなく、テストの実用性に関しても考慮がなされなければならない。今回の実験では、データの回収が完了してから、潜在ランク理論に基づくクラス編成を実施するまでに約 1 週間を要した。今回の実験参加者数が 1 学年全体の 10 分の 1 程度の人数であったにも関わらず分析に大きな時間を要した理由として、マークシートの読み取りエラーが非常に多く、目視でデータの確認を行ったり、読み取りデータの更新が複数回必要だったりしたことが挙げられる。精度の高いマークリーダーと専用のマークシートがあれば、より効率的にテストを実施することも可能だが、現状のままで確実に 1 週間以内にプレイスメントを完遂することは難しいだろう。しかし、語彙サイズテストの結果からクラスレベルごとの平均語彙サイズが異なることが明らかになったことから、既存の語彙サイズテストをプレイスメントテストに応用することも十分選択肢に入るだろう。

References

相澤一美・望月正道（2010）.『英語語彙指導の実践アイディア集：活動例からテスト作成まで』東京：大修館書店.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation.* Cambridge: Cambridge University Press.

木村哲夫.（2009）.「ニューラルテスト理論による英語プレイスメントテストの作成と評価」『関東甲信越英語教育学会研究紀要』23, 23–34.

木村哲夫（2013）.『潜在ランク理論を用いたコンピュータ適応型テストのためのアルゴリズムの提案と実装』早稲田大学審査学位論文.

木村哲夫（2016）.「*潜在ランク理論*」『日本言語テスト学会20周年記念特別号』217–222.

熊谷龍一・荘島宏二郎（2015）.『教育心理学のための統計学：テストで心を測る , 心理学のための統計学』東京: 誠信書房.

小泉利恵・飯村英樹（2010）.「ニューラルテスト理論の特徴：古典的テスト理論・ラッシュモデリングとの比較から」『JLTA (Japan Language Testing Association) Journal』13, 91–109.

小泉利恵（2011）「プレイスメント・テストの有効性： 2 種類のテストの比較と学生の反応から」『常磐国際紀要』15, 1–15.

小山由紀江・木村哲夫（2011）.「Neural Test Theory を使ったCan-do Statements の分析」『統計数理研究所共同研究リポート』254, 59–77.

清水裕子（2000）.「 4 年制大学におけるプレイスメント・テストの実施状況に関する研究」『JACET 全国大会要綱』39, 138–139.

荘島宏二郎（2010）.「ニューラルテスト理論―学力を段階評価するための潜在ランク理論―」植野真臣・荘島宏二郎（編著）『学習評価の新潮流』東京: 朝倉書店.

荘島宏二郎（2014）. Exametrika (Version 5.3) [Computer software]. Retrieved from http://www.rd. dnc.ac.jp /~shojima/exmk/jindex.htm

荘島宏二郎（n.d.）.「潜在ランク理論（ニューラルテスト理論）」Retrieved from http://www.rd.dnc.ac.jp /~shojima/ntt/jindex.htm

Shojima, K. (2007). Neural test theory. *DNC Research Note,* 07–02. Retrieved from http://www.rd.dnc.ac.jp/~shojima/ntt/jpaper.htm

Shojima, K. (2008). Neural test theory: A latent rank theory for analyzing test data. *DNC Research Note*, 08–01.

Retrieved from http://www.rd.dnc.ac.jp/~shojima/ntt/jpaper.htm

杉森幹彦（2003）.「英語統一テスト・習熟度別クラス編成・到達目標の設定および測定に関する実態調査の報告」『政策科学』10, 3–26.

Tamura, F. (2011). The relationship between vocabulary size and writing, *ARELE (Annual Review of English Language Education in Japan)*, *22*, 281–296.

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. Boston: Heinle & Heinle.

法月健（2013）.「受容語彙力を測定するプレイスメントテストにおけるラッシュモデルと潜在ランク理論に基づく規準設定の試行」『公益財団法人日本英語検定協会英語教育研究センター委託研究言語テストの規準設定報告書』2, 81–103.

水本篤（2006）.「語彙サイズテストは何を測っているのか？―語彙サイズテストの開発における問題点―」『統計数理研究所共同研究レポート』190, 71–80.

Mizumoto, A., & Shimamoto, T. (2008). A comparison of aural and written vocabulary size of Japanese EFL university learners, *Language Education & Technology*, *45*, 35–52.

望月正道（1998）.「日本人英語学習者のための語彙テスト」『財団法人語学教育研究所紀要』12, 27–53.

文部科学省初等中等教育局国際教育課外国語教育推進室（2014）.『英語教育の在り方に関する有識者会議 英語力の評価及び入試における外部試験活用に関する小委員会　審議のまとめ』Retrieved from http://www.mext.go.jp/b_menu/shingi/chousa/shotou/102/102_2/houkoku/1350999.htm

文部科学省高等教育局（2016）.『高大接続改革の進捗状況について』Retrieved from http://www.mext.go.jp/b_menu/houdou/28/08/1376777.htm

横内裕一郎（2014）.「能力記述文による自己評価と実際のスピーキング能力の関係―英検Can-doリストとCEFR-Jを使って―」『EIKEN BULLETIN』26, 218–230.

吉田弘子（2009）.「英語プレイスメントテスト分析：言語テストの観点から」『大阪経大論集』60, 93–103.

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In P. J. L. Arnauld, & H. Bejoint, (Eds.), *Vocabulary and applied linguistics* (pp. 126–132). London: Palgrave Macmillan.