■名古屋学芸大学　教養・学際編・研究紀要　第10号　2014年2月

英語力測定テストとしてのIRTテストの活用

The Use of the IRT Test as an English Proficiency Test

鈴木　　薫、大島　龍彦、永井　靖人
Kaoru SUZUKI, Tatsuhiko OSHIMA, Yasuhito NAGAI

Abstract

Although a wide variety of English proficiency tests are available to teachers and educational institutes, it is difficult to choose the most appropriate one. The majority of tests require an investment of a certain amount of time and money, making it hard for teachers to administer them in class. Tests should be valid and reliable enabling a beneficial backwash effect on teaching, and should also be practical. The English vocabulary test based on Item Response Theory, IRT, was introduced in this research to highlight these points. The results of the tests, which were given to 194 college students, are statistically analyzed and compared with those of the nationwide research, with the data also being compared to identify some differences among three departments the participants belong to. It transpired that the English vocabulary IRT test could be a candidate for precise measurement of English proficiency. Moreover, it was found that there was significant correlation between this test and another IRT English proficiency test used widely by many educational institutes in Japan.

1　はじめに

大学生たちが入学前に6年以上学習してきた英語という言語は、彼らにとって外国語であり、母国語である日本語とは習得の仕方も異なるため、その能力は日本語以上に多様である。さらに、近年では日本の企業の海外進出により、英語圏で成長期を過ごす者もいるため、第2言語環境の中で英語を習得する者も少なくない。また、学校教育で学習を積み上げてきた者と積み残しをしてきた者との間には、差が生じるのは自然なことである。英語を専門

として学ぶ学部がない大学においては、英語力を基準として入学者をふるいに掛けることをしていない場合もある。大学生の英語力は、均一ではないことは自明の理である。

教育をする側から見ると、担当する学生の英語力について全く何の情報もない状態で授業を開始することが求められ、授業中にテストをして初めて知る場合もある。もし各々が実施するテストが信頼性や妥当性についてしっかりと裏付けされたものではなく、授業で学習したことについての復習テストのような限られた範囲に対するものだとしたら、客観的に英語力を測定しているとは言い難い。このような状況において、短時間で正確に総合的な英語力を測定することができる物差しが求められる。

2　研究の背景

グローバルコミュニケーションで求められるスキルとして、英語力は不可欠である。英語力を示す指標として様々な検定試験が存在し、それぞれの目的に応じて利用されている。現代の日本でよく知られているものとして、小学生・中学生・高校性にとって馴染みの深い STEP(実用英語技能検定試験)、アメリカの大学に留学する際の指標となる TOEFL(Test of English as a Foreign Language)、就職活動をする者や社会人が最もよく利用する TOEIC(Test of English for International Communication) などを挙げることができる。しかし、これらのテストにはある程度の所要時間が必要となり、容易に実施できない。

今回の調査に利用した IRT 診断テストは、より短時間（約30〜40分）で語彙について調査するものであり、授業時間を圧迫することなく利用できるものである。リスニングテストが含まれていないので、音源の準備や確認をする手間を省くことができ、センター試験などでよく起こるリスニングテストのトラブルを、一切心配する必要がない。

さらに、英語の習熟度が高くなればなるほど、顕著な差があらわれるのは語彙力であるので、語彙力を適切に測ることができれば、信頼性や妥当性の高いテストであると仮定できる。

本調査は、大学生を対象として英語 IRT テストを実施することで、大学生の英語力の現状を明らかにする。さらに、大学生の英語力を診断するために有効な尺度となるかを検証するため、これまで利用してきた英語能力判定テストの結果と比較する。

3　項目応答理論

項目応答理論または項目反応理論（Item Response Theory、以降 IRT と略す）とは、すでに欧米では多用されているテストの作成、運用を行なうための新しい数理モデルである。従来の素点方式によるテストは、「古典的テスト理論（Classical Test Theory：CTT）」と呼ばれ、この方式では、受験する集団の特性によりテスト得点が変動してしまい、その値から問題固有の困難度を表すことはできない。受験者の能力についても同様で、問題群の困難度により得点が変動してしまい、その値から能力を十分に推定できないという難点がある。

一方、項目応答理論は、テスト項目に対する受験者の反応と潜在特性（能力や性格）を表す関数を用いることで、テスト項目の特徴分析や受験者の能力を推定する。一般的にはロジスティック関数が用いられるが、他にはベイジアン・ネットワークや S-P 表を用いる方法もある。

IRT を用いることで、①複数のテスト間の結果の比較が可能、容易になる、②平均点をテスト実施前に設定できる、③受験者ごとに最適な問題を瞬時に選び、その場で出題する、といったことが可能になる（豊田、2012）。IRT を用いたテストの代表的なものに、TOEFL、TOEIC、「語彙・読解力検定」、国家資格である「IT パスポート試験」がある。

要するに IRT を利用することの最大の利点は、受験者数・受験者集団の特性に影響を受けることなく、受験者の能力を診断できることである。調査の対象となる学生の数や学力の多様化に柔軟に対応することができ、異なる学部での横断的利用にも適している。

4　研究調査

4-1　対象者

名古屋学芸大学の学生192名（A 学部57名・B 学部76名・C 学部59名）

4-2　手順

①英語 IRT テストを調査対象者に実施する。

② IRT 診断テストを管理運営する株式会社エヌ・ティ・エスが、独自に全国の様々な大学を対象として実施した調査データと、調査協力グループ全体のデータを比較する。

③学部別のデータを比較する。

5　結果

全国データとの比較と学部間の比較について結果を報告する。

5-1　全国データとの比較

学年レベルと英検レベルに関連する全国データとの比較を、表1・2に提示する。

表1　学年レベルの度数分布

	高3以上	高2	高1	中3以下
A 学部	34	12	11	0
	（59.6%）	（21.1%）	（19.3%）	（0%）
B 学部	36	20	18	2
	（47.4%）	（26.3%）	（23.7%）	（2.6%）
C 学部	14	14	29	2
	（23.7%）	（23.7%）	（49.2%）	（3.4%）
3学部	84	46	58	4
	（43.8%）	（24.0%）	（30.2%）	（2.1%）
全国%	33%	16%	42%	9%

学年レベルごとの度数分布を全国データと調査協力グループとの間で比較した（表1参照）。まず、全国と調査協力グループ全体（A，B，C 学部の合算）との間には、有意な違いが認められた（$\chi^2(3) = 17.775$，$p < 0.01$）。さらに残差分析をおこなったところ、高3以上、高2レベルに相当する者の割合は、調査協力グループが全国を上回り（$p < 0.05$；$p < 0.10$）、高1、中3レベル以下に相当する者の割合は全国よりも低いことが示された（$p < 0.05$；$p < 0.01$）。

次に全国と各学部との間で度数分布を比較したところ、全国と A 学部とは有意に異なっていた（$\chi^2(3) = 14.502$，$p < 0.01$）。残差分析をおこなったところ、高3以上レベルに相当する者の割合は全国を上回り（$p < 0.01$）、高1、中3レベル以下に相当する者の割合は全国よりも低いことが示された（$p < 0.01$；$p < 0.05$）。全国と B 学部とは有意に異なっていた（$\chi^2(3) = 10.681$, $p < 0.05$）。残差分析をおこなったところ、高3レベル以上に相当する者の割合は全国を上回り（$p < 0.10$）、高1、中3レベル以下に相当する者の割合は、全国よりも低いことが示された（$p < 0.05$；$p < 0.01$）。一方、全国と C 学部とでは、度数の分布に有意な違いは認められなかった（$\chi^2(3) = 3.782$，*n.s.*）。

以上から、調査協力グループの成績を高校の学年レベルで見ると、調査グループ全体では高3レベル以上、高2レベルに達している者の割合が全国平均よりも高く、高1、中3レベルの者は全国平均よりも低かった。学部ごとに全国平均と比較しても、違いがほとんどない学部もあるが（C学部）、他の2学部は高2以上のレベルに達している者が全国平均よりも多かった。よって、調査協力グループは全国調査の大学生全体よりも良好なレベルの英語力を有しているといえる。

表2　英検レベルの度数分布

	2級	準2級	3級	4級以下
A学部	1	27	28	1
	（1.8%）	（47.4%）	（49.1%）	（1.8%）
B学部	0	27	45	4
	（0%）	（35.5%）	（59.2%）	（5.3%）
C学部	0	13	40	6
	（0%）	（22.0%）	（67.8%）	（10.2%）
3学部	1	67	113	11
	（0.5%）	（34.9%）	（58.9%）	（5.7%）
全国%	3.0%	27.0%	48.1%	21.8%

英検レベルごとの度数分布を全国データと調査協力グループとの間で比較した（表2参照）。検定に先立ち、2級レベルに達している者が1名しかいなかったので、2級レベルの度数を除き、準2級、3級、4級以下を分析の対象としたことを断っておく。全国と調査グループ全体（A、B、C学部の合算）との間には、度数の分布に有意な違いが認められた（$\chi^2(2)=31.326,\ p<0.01$）。さらに残差分析をおこなったところ、3級レベルに達している者の割合は全国平均よりも高く（$p<0.10$）、4級レベル以下は全国平均よりも低かった（$p<0.01$）。準2級レベルでは、調査協力グループと全国平均との間に有意差はなかった。

次に全国と各学部とで度数分布を比較した。A学部では英検4級以下に相当するレベルの者が1名しかいなかったので、4級以下のカテゴリーを分析から除外した。その結果、全国平均とA学部との人数の偏りは有意でなかった（$\chi^2(1)=2.325,\ n.s.$）。全国とB学部との違いは有意であった（$\chi^2(2)=11.304,\ p<0.01$）。残差分析をおこなったところ、準2級、3級に相当する者

の割合は全国と有意差がなく、4級レベル以下に相当する者の割合は、全国よりも低いことが示された（$p < 0.01$）。全国とC学部との違いは有意であった（$\chi^2(2) = 6.945,\ p < 0.05$）。残差分析をおこなったところ、準2級に相当する者の割合は全国と有意差がなく、3級に相当する者の割合は有意に多く（$p < 0.05$）、4級レベル以下に相当する者の割合は、全国よりも低いことが示された（$p < 0.05$）。

以上より、英検のレベルで調査協力グループの英語力を全国と比較すると、2級レベルに達している者は少なく、準2級は同程度であった。つまり、調査協力グループの英語力は大学生に求められる到達度をやや下回っているといえる。

調査協力グループの英語力について、本研究で用いたIRTテストの結果からは、大学横断的、相対的には優位に立っているが、到達度で評価するとやや低いと考えられる。とりあえず学部卒業後の就職といった実利面で考えると、公務員や教員、大手企業を志望しても、客観テストの段階で劣勢に立たされることが予想される。よって、今後はリメディアルを意識したカリキュラム再編、授業の再設計が必要になってくると考えられる。

5-2　学部間の比較

表2に英語IRTテストの結果を学部別に示す。一元配置の分散分析をおこなった結果、学部の主効果は有意であった（$F(2, 192) = 13.93,\ p < 0.01$）。Tukey法を用いた多重比較によると、A学部とB学部の平均がC学部より有意に大きかった（$MSe = 8.68,\ p < 0.01$; $MSe = 8.10,\ p < 0.01$）。しかし、A学部とB学部との平均の差は有意でなかった。

表2　学部別の英語IRTテスト得点

	A学部	B学部	C学部
Mean	628.8	618.4	585.5
SD	44.8	45.9	51.1
n	57	76	62

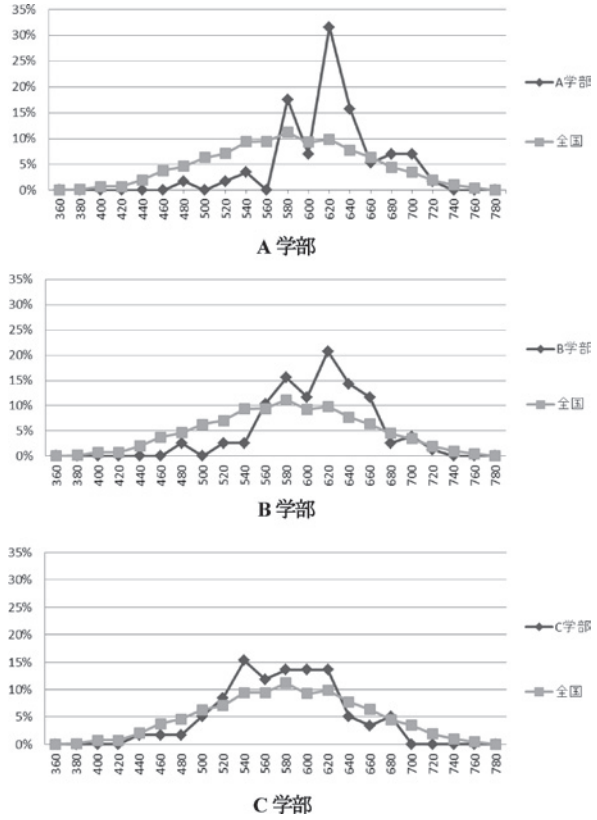

図1 スコア分布

次に、学部ごとのスコア分布を図1に示す。

A 学部は600点以上620点未満のあたりが谷間となりその前後が二つのピークになっている。B 学部は A 学部ほどではないが、同様に600点以上620点未満のところで低くなる箇所が現れている。この谷間が A 学部も B 学部も習熟度を2分する境界を示していると予測できる。C 学部ではこのような境界は観察できないが、3つの学部の中でもっとも多様な集団であると解釈できる。

6　英語能力判定テストとの比較

英語能力判定テストは、中学生や高校生にも馴染みのある実用英語技能検定の過去問をベースに作成された IRT テストの一種である。出題項目は、語

彙・熟語・文法、文章構成、読解、リスニングの4つのパートに分かれ、実施時間は説明も含めて最低でも約70分で、1回の授業時間を必要とするものである。

本調査の対象となっていない別の大学生グループを対象に英語 IRT テストと英語能力判定テストを実施し、ピアソン相関係数を算出した（$r = 0.6207$, $df = 41$, $p < 0.01$）。

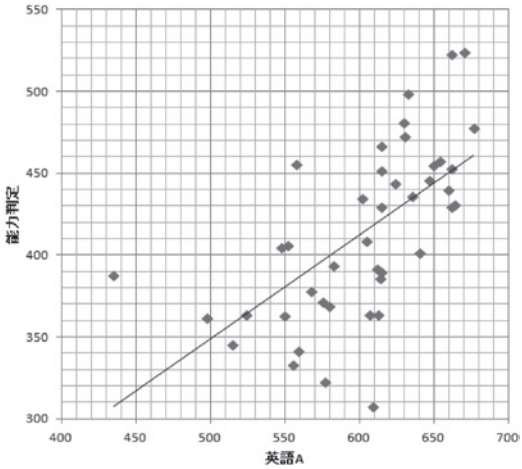

図2 英語 IRT テストと英語能力判定テストの相関

図2に示されているように、英語 IRT テストと英語能力判定テストの全体スコアとの間には、高い相関があると判断できる。しかし、英語能力判定テストのセクションごとのスコアとの間では、文章構成（$r = 0.5257$, $df = 41$, $p < .01$）とリスニング（$r = 0.3736$, $df = 41$, $p < 0.02$）との間に有意な相関がみられたが、語彙（$r = .2275$, $df = 41$, *n.s.*）と読解（$r = .247$, $df = 41$, *n.s.*）とは有意な相関は検出されなかった。

全体スコアとの間に高い相関が検出されたにもかかわらず、セクションごとにばらつきが生じたことは今後の課題となるが、総合的英語力を測定する指標としては有効であると解釈できる。小野（2007）は、NIME（メディア教育開発センター）プレースメントとして英語語彙力診断テストを、英語能力判定テストと同時に実施しているが、今回の調査結果よりもさらに高い相関（$r = .776$, $n = 550$）があると報告している。この NIME プレースメントが

改良されて、今回使用した英語IRTテストとしてリリースされている。

2つの研究調査で同じ結果を得ていることから、英語IRTテストは英語能力判定テストとの相関性において安定していると判断できる。

7　考察

本調査で採用したIRT診断テストは、数十万名にのぼる中学生・高校生に対して大規模調査を実施し、その情報を元に厳選した問題を使用している。よって、限りなく大きい母集団との相対評価となっていることで、スコアの信頼性は高い。さらに、学年レベルや英検レベルという段階別評価の結果も提示することで、学生の能力を感覚的に把握することができる。例えば、TOEFL500点とTOEIC500点はどれくらい違うのかが、具体的に把握できる者は少ないが、日本で英語教育を受けた者にとって馴染みのある英検のレベルであれば、何となくわかる人は多いであろう。

さらに、総合的な英語力の指標として従来から利用されている英語能力判定テストとの相関性が高いという点においても、本調査で採用したIRT診断テストの有効性や信頼性は安定していると解釈できる。

本調査では、少数ではあったがその存在が注目された中3以下レベル・英検4級レベルの学習者に対して、何らかの方法でリメディアル教育の実施が必要となることが明らかとなった。現代においては、大学生の学力低下が問題視され、英単語の意味を辞書で調べてもその日本語の漢字が読めなかったり、百万以上の数字などは英語どころか日本語でもすぐに読めなかったりする学習者がいる。英語において非常に初歩的な文法力が欠如していたり、英語で曜日が即座にわからなかったりする。英語力は中学校の3年間で習う英語がベースとなり、それを応用しながらさらに能力を伸ばすことが必須条件である。文法力と語彙力の欠如がこれらの学習者の英語力向上を妨げている。オーラル・コミュニケーションを重視した教育を行っても、簡単な会話の運用はできるかもしれないが、内在する文法や語彙の知識が乏しければ、高度な内容のリスニングやスピーキングはできない。将来的な英語力の向上を促すような文法や語彙に関する基礎的能力を備えることが必要不可欠である。

次に、特に図1で顕著に現れている二つの山について考察する。この二つの山は、調査協力グループの英語力が二分されていることを示している。つまり、今回の調査対象者たちには習熟度別の指導が望ましい。しかし、日本人の国民性として皆が一緒であることを好む意識が潜在的にあることや、平

等という権利を画一という言葉と混同してしまう風潮から、習熟度によるグループ別の指導に抵抗を感じることでの弊害が現れていることも事実である。仲が良い友達と一緒に授業を受けることができないことで不満を抱く今時の学生たちの気質も問題となる。できる学生にとっては、クラス分けは好評であるけれども、意欲はあるが学力が少し足りなくて習熟度の低いクラスに入ってしまった学生は不満を持ち、底辺にいる学生たちは完全に学習意欲をなくしてしまう場合も多々ある。低いレベルのグループに入って劣等感を抱えた学生への配慮が教員側の精神的な負担にもなっている。教員によっては、授業内容をあからさまに区別することを避けるために共通のテキストで全く同じ内容の授業を行い、習熟度別の指導が十分に機能していない教育現場もみられる。習熟度別グループ分けによる指導とは逆に、習熟度の違う者同志を接触させることで刺激を与えるほうがよいという考え方もある。このような問題を上手く解決する教育法として、集団学習と個別学習を組み合わせたCALL(Computer Assisted Language Learning)の活用がある。誰もがスマートフォンを携帯する時代において、教室で実施するCALLと携帯端末を利用した教育システムの構築が、多様な英語力の学習者を指導するために利用することができる。よって、IRTテストによる診断によって、個々の学習者の英語力を把握するとともに、CALLを利用した個別指導により各々のレベルに合せた教育を実現することが究極の指導法の1つとして挙げられる。

8　おわりに

今回の調査で実施した英語IRTテストは、短時間に実施することが可能で、なおかつ各学部の特色を把握するのに十分に機能することを示す結果が得られた。今後は、習熟度別の分析や同時に実施している日本語IRTテストとの比較研究も進める予定である。

従来利用してきた英語能力判定テストの全体スコアとの比較において、ある程度の相関が検出されているが、英語能力判定テストの各セクションとの相関については、文章構成やリスニングとの間には相関がみられたが、語彙や読解との間には相関がなかった。ゆえに、さらなる研究調査として、英語能力判定テスト以外のテストスコアとの比較研究を行い、教育現場に有益な情報を提供していきたい。

参考文献

Brown, J.D. (1988). *Understanding Research in Second Language Learning*. Cambridge University Press.

Chomsky, N. (1986). *Knowledge of Language.* Praeger Publishers.

Crystal, D. (1997). *English as a Global Language*. Cambridge University Press.

Dulay, H., M. Burt & S. Krashen (1982). *Language Two*. Oxford University Press.

Henning, G. (1987). *A Guide to Language Testing*. Newbury House Publishers.

Hughes, A. (1989). *Testing for Language Teachers*. Cambridge University Press.

O' Malley, J. M. & A. U. Chamot. (1990). *Learning Strategies in Second Language Acquisition.* Cambridge University Press.

大友賢二・中村洋一（2002）.『テストで言語能力は測れるか ～言語テストデータ分析入門～』桐原書店.

小野博（2007）「基礎学力測定を目的としたプレースメントテストの開発と日本人大学生の日本語・英語力構造」『平成16年度～平成18年度科学研究費報告書（基盤研究（B）（2））16300281「e-learning における学力対応型学習プログラムの開発に関する実証的研究」・平成17年度～平成18年度科学研究費報告書（萌芽研究）17652068「複数の音声認識回路を利用した外国語学習システムの構築に関する実証的研究」日本人大学生を対象とした日本語･英語教育　―リメディアル教育から実力養成教育への展開』3-10.

豊田秀樹（2012）.『項目反応理論［入門編］』朝倉書店.

Yalden, J. *Principles of Course Design for Language Teaching*. Cambridge University Press.

＊本研究は平成24年度名古屋学芸大学学長裁量経費（研究課題「プレースメントテストとしての英語 IRT テストの活用」）を利用した研究である。