

[原著]

ラッシュモデルを使ったテスト等化 ープレースメントテストに見る入学時英語力の推移ー

渡 辺 雄 一

Rasch model test equating: transition of the English ability
of incoming students over four years

Yuichi WATANABE

熊本保健科学大学 共通教育センター

テスト等化は、異なる問題から構成される複数のテストの結果を比較可能とするものである。今回はラッシュモデルという統計手法を使い、毎年新入生を対象に実施される英語プレースメントテストの平成21年度版から24年度版までの等化をおこなった。等化されたテスト結果をもとに年度別、専攻別の比較をおこなった。年度別では、平成22年度の新入生の英語力が最も高く、平成24年度新入生が続いた。専攻別では、理学療法学・医学検査・看護の方が作業療法学・言語聴覚学よりも英語力が高かった。ただし、これらの年度のプレースメントテストは問題構成および問題数の両面において必ずしも同等とはいえず、結果の解釈には注意を要する。

キーワード：テスト等化，ラッシュモデル，プレースメントテスト，英語力，大学初年次

I. 緒言 (はじめに)

熊本保健科学大学では、学部1年次生に対し入学直後にプレースメントテスト（以下「配置テスト」という）を実施し、能力別編成により必修の英語授業をおこなっている。テストは制限時間60分で、表1にあるような問題構成となっている。解答は全て多肢選択式で、正解であれば1点、不正解であれば0点となる。テストは毎年改良を加えてきた結果、一部共通な問題と異なる問題から構成されている。このような状況において入学時の英語力の推移を見たい場合には、どのような方法が可能であろうか。異なる問題を含む以上単純に平均点を比較することはできない。そこで、ひとつの方法としては全年度に共通な問題のみを取り上げ、この平均点を比較する事が考えられる。表1からわかるように英文1と英文2は4年間に渡り使用されており、また質問も同一のものが出題された。そこで、この16問に対する正答率を比較するという事は可能である。しか

し、テスト等化という方法を使うと共通問題に対する回答をもとに独自問題を含むテスト全体が比較可能となる。この論文では、ラッシュモデルを使ったテスト等化の方法を説明し、その等化の結果比較可能となった4年度に渡る配置テスト結果を分析し、入学時の英語力の推移を報告する。

配置テストの目的は、新1年次生を英語力に応じて、3（医学検査学科、看護学科）ないし4（平成23年度に言語聴覚学専攻が設置された以降のリハ学科）のグループに分けることである。通常、テストの質は信頼性と妥当性の観点から論ぜられる。配置テストの場合、信頼性とは精度の問題であり、受験者をいかに均質なグループに分けることができるかは測定誤差の大きさにかかっている。この測定誤差を反映したものが、表1下段におけるK-R20という指標である。これは、テストの内部一貫性を表し、1.0に近いほど測定誤差が少なく、より細かなクラス分けが可能となる。一方、この精度をより直接的に表しているのがSeparationである。これは、テ

表1：プレースメントテストの問題構成、信頼性、受験者数

	年度	平成24	平成23	平成22	平成21
語彙問題	共通	20問	20問		
	独自	6問	32問*		
文法語法問題	共通	20問	20問		20問
	独自				10問
長文読解	英文1	8問	8問	8問	8問
	英文2	8問	8問	8問	8問
	英文3	9問			
	英文4		8問	8問	
	英文5			8問	8問
合計		71問	96問*	32問	54問
Separation		2.60	2.27	1.72	1.70
K-R20		0.87	0.84	0.75	0.74
受験者数		371名	371名	370名	341名

* 語彙問題のうち前半の26問（全て独自問題）は正答率が高すぎたため、クラス編成及び分析には用いられなかった。そこで、今回の分析の対象としたのも残り70問である。

スト精度から、いくつの均質なグループに分けることができるかを示す指標である。例えば、Separation が3.0であれば、受験者を3つの均質なグループに分けるだけの精度があるということになる。過去4年度のSeparationを見ると、毎年改善してきているが未だに3には至っていない。これは受験者の英語力に大きな差がないことが主な原因である。同じ大学の入学者は入学試験によりすでに選別がなされているため、その英語力も均質性が高い。このことはTOEICなどの大規模試験と比較するとわかりやすい。TOEICでは、初級者から最上級者までが受験するため、全受験者で見るとより細かな能力グループに分けることが容易である。それに対し、ひとつの大学の学生の場合、均質性が高いため、その中で細かな能力に分けることは困難となる。

一方、妥当性は均質性の中身の問題である。英語力で均質といっても、それは読解力においてなのか、聴解力、会話力においてなのか、語彙力、文法力、または総合的な英語力においてなのか。例えば、配置テストにより分けられたグループは、読解力においては均質であっても、リスニング力では様々な能力が混在しているかもしれない。本学においては、英語読解力における均質なグループを作る目的で配置テストが作成されてきた。この点がどの程度達成されているかは容易に検証できない。配置後の各ク

ラスの担当講師によるアンケート調査などの実施が検討される。

II. 方法

1. 英語力の測定単位としてロジットを使用

各年度の配置テストの結果は、問題（以下「項目」という）数が異なるため各項目に対する素点（1または0）の合計点では英語力の比較ができない。また、そもそも素点の合計点は通常間隔尺度として扱われるがこれにも問題がある。たとえば、合計点10点と30点の間の能力差と70点と90点の間の能力差は同じ（等間隔）といえるだろうか。前者よりも後者のほうがより大きな能力差を表している可能性は否定出来ないし、またその逆も可能である。そこで、素点の合計点に代り能力を等間隔尺度に表したのが、ラッシュモデルによるlogitである。ロジットは正答確率を対数オッズ変換したものである。ロジットによる英語力の計算にはコンピュータソフトWinstepsを使った^{1, 2)}。

2. 等化

複数年度に渡る配置テスト結果を比較するためには、すべてのテストが同じ参照枠組（frame of reference）によっていることが必要である。すなわち、たとえばロジット1.0で表された英語力は、どの年度においても同じ能力を表していなければならない。実際にはある年度のテストを基準として、他の年度をこの年度の参照枠組に合わせることで、共通の参照枠組が可能となる。今回は、最も完成度が高いと思われる平成24年度の配置テストを基準とし、他の年度のテストをこれに合わせることにした。方法としては、複数年度に共通な項目を使ったcommon item equatingをおこなった^{3, 4)}。等化のために使用した共通項目は英文1と英文2に付随する合計16問である。

3. 平成23年度と24年度の等化

平成23年度と24年度における共通項目の難易度をロジットで求め、両年度間の関係を見た。下の散布図（図1参照）にあるようにこの16項目の難易度は両年度において殆ど変わらなかった。両年度間の相関係数は0.96、傾きは0.94、直線のx軸との交差点は0.03ロジットであった。それぞれ、すべての項目

の難易度が両年度で完全に一致していた場合は、1.0, 1.0, 0.0となる。これからもわかるように共通項目の難易度は両年度において殆ど変わらなかった。

そこで、傾きと x 軸との交差に関してごくわずかな修正を加えただけで平成23年度のテスト全体が平成24年度のテストに等化された。

ラッシュモデルにおける能力測定では、いくつかの項目に正解したかではなく、難易度の異なる項目にそれぞれどのくらいの確率で正答できるかという確率による能力の測定をおこなう。能力が1.0ロジットであれば、難易度が1.0ロジットの項目に50%の確率で正答できることを意味する。共通項目の難易度がほぼ同じであるということは、両年度の全項目がほぼ同じスケールにあるということの意味する。すなわち、平成24年度での難易度1.0ロジットと平成23年度の難易度1.0は同じ意味を持つということである。これを、能力で見ると同じスケールにある項目群により測定された能力は同じ意味を持つということである。やはり、平成24年度での能力1.0ロジットと平成23年度の1.0は同じ能力を表すということである。

4. 平成22年度と24年度の等化

平成22年度における共通項目の難易度をロジットで求め、平成24年度との関係を見た。両年度間の相関係数は0.89、傾きは1.07、直線の x 軸との交差点は0.39ロジットであった（図2参照）。

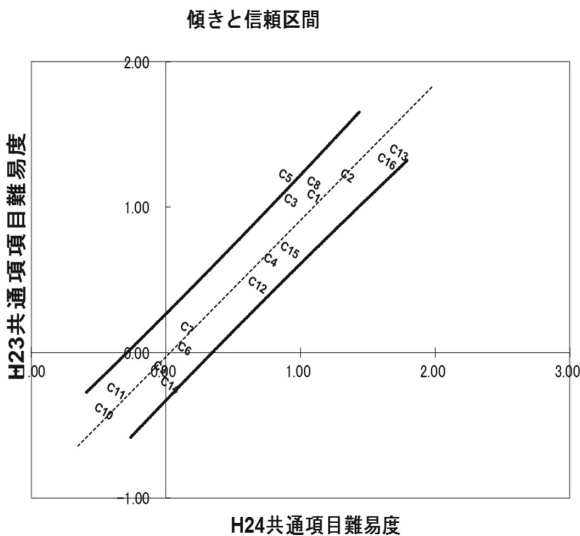


図1：共通項目のH24年度とH23年度間の関係

両年度で有意に異なる難易度を持つ項目が5つあったためこれらを除外し、再び散布図を描いた。すると、両年度間の相関係数は0.98、傾きは1.05、直線の x 軸との交差点は0.47ロジットであった。そこで、傾きと x 軸との交差に関してそれぞれ1.0と0.0となるよう平成22年度テストを調整した。具体的には、平成22年度テストの全項目についてそれぞれの難易度（単位はロジット）に対し（1/1.05）を乗じ（0.47）を加える線形変換をおこなった。これによって、平成22年度の項目難易度は平成24年度のものと同じスケールに基づくものとなった。その結果、各受験者の英語力（単位はロジット）も両年度で同じスケールに基づくことになり、比較可能となった。

5. 平成21年度と24年度の等化

平成21年度における共通項目の難易度をロジットで求め、平成24年度との関係を見た。両年度間の相関係数は0.97、傾きは0.82、直線の x 軸との交差点

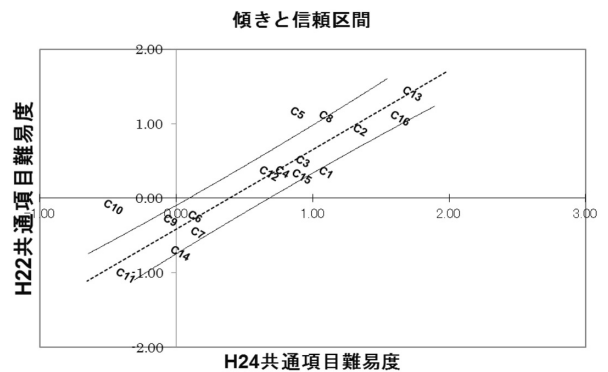


図2：共通項目のH24年度とH22年度間の関係

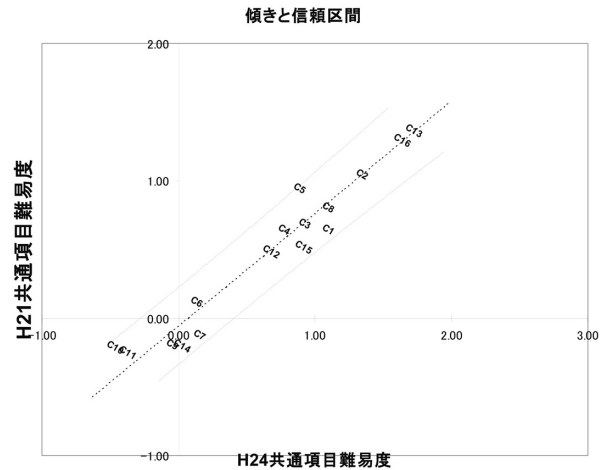


図3：通項目のH24年度とH21年度間の関係

は0.07ロジットであった（図3参照）。

両年度で難易度が大きく異なる項目は無かったため、この16の共通項目を基に等化をおこなった。平成22年度テストの各項目の難易度（単位はロジット）に対し（1/0.82）を乗じ（0.07）を加える線形変換をおこなった。これによって、平成21年度の項目難易度および英語力は平成24年度のものと同じスケールに基づくものとなった。

6. 各年度の比較

上記の等化の手順により、平成21年度から24年度までの四つの配置テスト結果は相互に比較可能となった。ここで、英語力を表すロジットをそのまま使うことはわかりづらいためこれを0から100までのスケールに線形変換する。これにより、平均値や標準偏差はこの新しいスケールに基づき計算されるが、このスケールも等間隔尺度あり、素点の合計点のような順序尺度ではない。このスケールに基づき、年度間、専攻間において英語力に差があるかどうかを検証した。統計分析は2要因（年度と専攻）の分散

分析を使った。

Ⅲ. 結 果

年度別・専攻別の英語力に関する記述統計は表2に掲げるとおりである。これをグラフで表したのが図4である。

2要因の分散分析の結果、主効果「年度」と「専攻」はともに危険率5%で有意であった：それぞれ $F(3,1435) = 13.484$, $F(4,1435) = 8.254$ 。交互作用は有意ではなかった： $F(10,1435) = 0.417$ 。主効果について危険率5%で Tukey の方法による多重比較をおこなった。

この結果、平成22年度入学生の英語力が最も高く、次に平成24年度、最後に平成21年度と23年度であった（表3）。一方専攻間では、理学療法、医学検査、看護が言語聴覚、作業療法よりも英語力が高かった（表4）。最も大きかった理学療法と作業療法間の差の効果量は Cohen's d で0.50であった。これは Cohen の指針によると medium に相当する大きさである⁹⁾。

表2：年度別・専攻別記述統計量

年度	専攻	平均値	標準偏差	N
H21	医学検査	53.65	6.21	124
	看護	52.15	5.78	125
	理学療法	54.83	8.71	45
	作業療法	50.97	5.89	47
	総和	52.89	6.49	341
H22	医学検査	56.25	7.34	127
	看護	56.06	6.68	148
	理学療法	57.86	6.71	49
	作業療法	53.93	6.94	46
	総和	56.10	6.99	370
H23	医学検査	53.69	6.94	114
	看護	53.29	6.59	107
	理学療法	53.77	5.42	50
	作業療法	51.37	4.88	49
	言語聴覚	52.03	6.13	51
総和	53.05	6.32	371	
H24	医学検査	55.26	7.07	118
	看護	54.91	6.29	115
	理学療法	55.56	7.76	48
	作業療法	52.31	6.99	40
	言語聴覚	52.63	5.18	50
総和	54.52	6.76	371	
総和	医学検査	54.74	6.97	483
	看護	54.21	6.52	495
	理学療法	55.51	7.31	192
	作業療法	52.12	6.24	182
	言語聴覚	52.33	5.66	101
総和	54.16	6.77	1453	

Ⅳ. 考 察

1. 等化と年度間の比較

等化したテスト間の比較により、平成22年度入学生の英語力が最も高いという結果となった。しかし、この結果の信頼性には疑問がある。第一にどの専攻

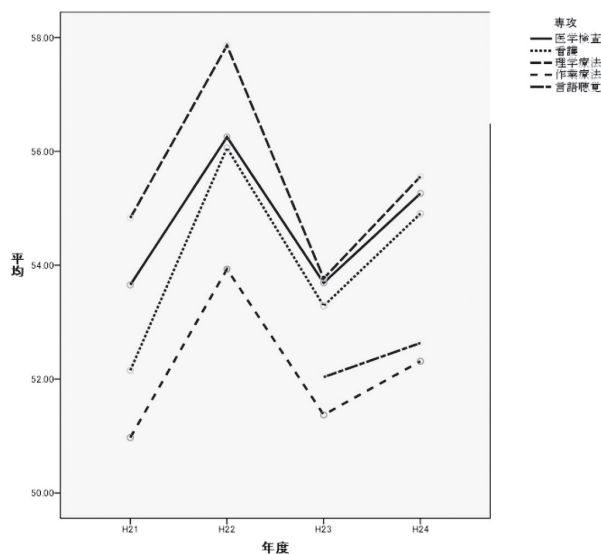


図4：年度・専攻別の平均値

表3：平均値の年度間の多重比較

年度	N	サブグループ		
		1	2	3
H21	341	52.89		
H23	371	53.05		
H24	371		54.52	
H22	370			56.10

表4：平均値の専攻間の多重比較

専攻	N	サブグループ	
		1	2
作業療法	182	52.12	
言語聴覚	101	52.33	
看護	495		54.21
医学検査	483		54.74
理学療法	192		55.51

(設置前の言語聴覚を除く)においてもこの年度だけ飛び抜けて平均が高かったことの原因が不明である。第二にこの年度だけは、他年度と異なり英文の数が多く(5)そのかわり語彙や文法語法問題がないといういわば特殊な形式であった。そのため等化により他年度と同じスケールと比較可能であるとして分析をおこなったがこれには無理があった可能性がある。本来等化は対象となるテストがパラレルフォームであることが前提となる^{3, 4)}。パラレルフォームとは毎年複数回実施される TOEIC のように同じテストで、問題構成、問題数、難易度、制限時間などが同じであるが問題自体は異なる複数のテストのことである。本学における配置テストは60分で読解能力を中心とした英語力を測るという目的は同じであるが、問題数や問題構成が毎年異なり、その点で厳密な等化をおこない難い面がある。平成22年度は他年度と大きく問題構成が異なったが、平成23年度と24年度の間においても語彙問題の数の上で大きな差があり(52問対26問)、このことが後半の読解問題解答に影響を与えた可能性は否定出来ない。また、平成21年度は文法語法問題と読解問題はあったが語彙問題がなかったという問題構成の点で平成23年度、24年度と異なった。なぜ等化の結果平成22年度が飛び抜けて英語力が高い結果となったかについて今後さらなる検討が必要である。

2. 専攻間の比較

年度間の比較と異なり、専攻間の比較においては年度ごとの問題構成や問題数の差異に直接影響を受けない、または受けたとしても同様に受けるため、比較の結果は信頼出来るものである。

V. 結語 (おわりに)

今回の報告は、ラッシュモデルを使ったテスト等化の手順の説明が主であり、その結果としての年度間の比較それ自体は解釈に注意を要する。テスト等化により、共通項目を含んでいれば異なる問題から成る複数のテストの結果を比較することができる。このことにより、年度間の異なる学習者集団の能力の比較だけでなく、同じ学習者集団の異なる時期、たとえば入学時から卒業時までの能力の推移をパラレルフォームを使って見ることができる。ただ、今回対象となった平成22年度テストのように問題構成の違いが等化に予想しなかった影響を与えた可能性があり、この点に関する研究がさらに必要である。

文 献

- 1) Linacre, J.M.(2012a). *Winsteps*® (Version 3.75.0) [Computer Software]. Beaverton, Oregon: Winsteps.com
- 2) Linacre, J. M. (2012b). *Winsteps*® Rasch measurement computer program User's Guide. Beaverton, Oregon: Winsteps.com
- 3) Skaggs G, Wolfe E. W. (2010). Equating designs and procedures used in Rasch scaling. *Journal of Applied Measurement*, 11(2), 182-95.
- 4) Wolfe E. W. (2000). Equating and item banking with the Rasch model. *Journal of Applied Measurement*, 1(4), 409-34.
- 5) Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (second ed.). Lawrence Erlbaum Associates.

Rasch model test equating: transition of the English ability of incoming students over four years

Yuichi WATANABE

Abstract

Test equating makes it possible to compare multiple tests made of different sets of items. In this study, Rasch model was used to equate four English placement tests administered between 2009 and 2012 to incoming freshmen. The data derived from the equated tests were used to compare the English ability of the students across the four years and also across the majors. The results indicated that the 2010 students had the highest English ability followed by the 2012 students. In terms of their disciplines, the physical therapy majors, the medical technology majors, and the nursing majors were higher in English ability than the occupational therapy majors and the speech therapy majors. However, the difference across the years needs to be interpreted with caution because the equatability of the four placement tests came into question due to differences in their composition and number of items .