古典的テスト理論とラッシュモデルとの比較による英語テスト分析

スマイリ ジム*，増井三千代**

An Analysis of an English Test Comparing Classical Test Theory and the Rasch Model

SMILEY Jim, MASUI Michiyo

はじめに

テストは学習者の学習成果や能力を測定するツールとして，重要な役割を担っている．伝統的，または「古典的テスト理論」(Classical Test Theory：CTT) では，テストの合計得点は「単純素点」と言われ (Holster and Lake [2014])，正答した項目数を加算することによって算出される．各受験者の素点と受験者全体の素点は統計分析の基礎を形成するが，多くのデータを扱う場合，「個人を分類あるいは集合体にクラスタリングする必要がある」(Choppin [1983])．クラスタリングとは，「ある集団の中で，パターンが似た人を集めて，似たもの同士のグループを作る手法である」(三浦 [2004])．集団へ焦点を合わせることで，その集団の特性に依存した統計が可能になる．例えば，母集団あるいは標本の得点を利用して平均値，分散，偏差値などを求めれば，受験者集団の傾向や個人の相対的な位置づけを容易に確かめることができる．

だが，これらの統計結果を正確に解釈するには，測定道具，すなわちテス

* 東北文化学園大学総合政策学部専任講師
** 東北文化学園大学総合政策学部准教授

ト自体の検証も行わなくてはならない．もし欠陥[1)]のある項目がテストに含まれているとすれば，受験者は不備があるテストを受験して意味のない得点を得ることになる．テスト開発者はテストの信頼性を確かめ，必ずその機能を果たせるようにすべきである．粗悪なテストは見当違いの内容を測定するばかりでなく，テストとは直接関係のないところで，受験者を混乱させていることもある．例えるなら，熱湯の量を計るのに紙製の計量カップを使用するようなものである．このような測定道具，すなわち紙製の計量カップは，その意図するタスクに対しては不適切な手段である．

CTT では，こうしたテストの質に関わる問題を明らかにする方法として，項目分析（item analysis）を行う．項目容易度（item facility：*IF*）と項目弁別力（item discrimination：*ID*）という対のツールは，個々のテスト項目と全体的な得点との関係を統計的に示してくれる．しかしながら，多くの研究者が指摘している通り（大友［1996］，小山・木村［2011］），CTT に基づく分析には様々な弱点がある．例えば，ある受験者が２つの異なる集団で同一テストを受験した場合，この受験者の2つの得点に違いはなくとも，各集団の能力に差があれば平均値や順位などの統計データは異なる数値を示す．個々の受験者の能力を正しく測定するという意味においては，受験者集団の特性に依存する統計分析では不十分であるといえる．「相対的な位置関係は分かっても能力そのもの（絶対的評価）は明らかにならない」からである（小山［2010］p.15）．

また，ある受験者が２つの難易度が異なるテストを受験して，難しいテストで60点，易しいテストで70点を取得したとする．このような場合，単純に得点を比較できないのは明白であり，どちらの得点がより信頼できるのかも分からない．つまり，CTT では受験者の能力がテスト項目の難易度によって左右されるため，テストが異なる場合，受験者の能力を比較することは極めて難しい．とはいえ，教育現場において同じテストを繰り返し実施することは現実的ではない．

そこで，このような CTT の限界を克服するために開発されたのが，項目応答理論（Item Response Theory：IRT）である．IRT では項目の困難度と受験者の能力は独立して扱われ，共通の尺度上で受験者の絶対的な能力値を測定

1） 欠陥については，構成概念妥当性や表面的妥当性など信頼性に関する問題も含まれるが，本研究ではそれらの議論は対象外とする（Hughes［1989］）

することができる．そのため，「異質な受験者が，異なる項目を，異なる日時に，異なる場所で受験したにも関わらず，被験者は統一された処遇を受けることができる」（豊田［2002］，pp.16）．IRT に基づいた TOEFL が，世界中の大学の留学可否判断に使用されているのはそのためである．この IRT 分析を可能にする数理モデルの一つがラッシュモデル（Rasch Model[2)]）であるが，複雑な計算が求められるため，専用のソフトウェアを使用する必要がある．

本稿は，ラッシュモデルがもたらす情報はその研究に求められる時間を考慮しても CTT よりもはるかに価値があるのか，という問いに答えようとするものである．著者らが実施した小規模なテストを CTT 分析およびラッシュモデル分析にかけた事例研究を示す．各分析から得られるテスト項目情報を検討し，CTT とラッシュモデルの使用について実際的な判断を示す．なお，本研究は数学および統計に関する高度な知識がなくても比較的簡単に扱える分析ツールに焦点を当てているため，複雑な数理モデルの詳細については解説しない．

2　古典的項目分析手法

CTT の観点から，テストが意図した言語能力を的確に測定しているかを検証するには，項目分析（item analysis）が非常に有効である．テストを構成するそれぞれの項目を分析した結果は，テストの難易度レベルの調整や有効に機能していない問題の発見や修正といったテストの質改善に役立てることができる（斉田［2011］，pp.38）．

本章では，古典的項目分析の項目容易度（item facility：*IF*）と項目弁別力（item discrimination：*ID*）の算出方法について解説する．さらにテスト全体の項目バランスを把握する方法として，折半法（split-half method）を例として示す．

(1)　項目容易度（item facility：*IF*）

IF はそのテスト項目がどのくらい易しいかを示す指標で，次の式で算出される．

---

2)　デンマークの数学者 Georg Rasch が考案した IRT の基本数理モデルで，1パラメタ・ロジスティック・モデル（1PLM）を指す．

$$IF = \frac{\text{正答数}}{\text{総受験者数}}$$

あるテスト項目に対して，その項目に正答した数を受験者の総数で割る．受験者全員がその項目を正答した場合は，*IF*=1.00となる．それに対して，受験者全員が誤答した場合は，*IF*=0.00となる．すなわち，値が大きい項目ほど，易しいテスト項目（正答率が高い項目）であると解釈できる．Brown（2011：75）によれば，適切な *IF* 値は .30 ～ .70だとされている．

この簡単な分析は，受験者の能力を的確に測定できない極端に「易しすぎる」あるいは「難しすぎる」テスト項目を浮き彫りにしてくれる．算出した *IF* 値はテストの難易度やレベル調整に役立てることができる．

(2)　項目弁別力（item discrimination：*ID*）

ID はそのテスト項目がどれだけ有効に高得点者と低得点者を弁別しているかを示す指標で，次の式で算出される．

$$ID = IF\text{（上位25%）} - IF\text{（下位25%）}$$

まず，得点数をもとに受験者を上位群，中位群，下位群に3分割する．（割合は通常25% から33% の間で設定する．ここでは25% を基準とする）．あるテスト項目に正答した成績上位群の *IF* 値から成績下位群の *IF* 値を差し引いた値が，その項目の弁別力となる．弁別力指数は－1から1までの値をとり，プラスの値が大きいほど，そのテスト項目は高い弁別力を持つ項目であるといえる．具体的には，.40以上ならとても良い項目，.30 ～ .39なら悪くはないが改善の余地がある項目，.20 ～ .29なら改善が必要な項目，.19以下は削除か作り直しが必要な項目とされる（Brown［2011］p.75）．*ID* がマイナス値を示すときは，低得点者の方が高得点者より正解していたことになり，そのテスト項目は不適切であることを意味する．このようなことが起きた場合は，その理由を調査する必要がある．

(3)　折半法（split half method）

折半法はテストの全体的な信頼性に関する推定値を提供する（Hughes［1989］）．CTT はテストの信頼性を推定する方法をいくつか提唱しているが，

折半法は比較的扱いやすく，次の方法で求められる．テスト項目をなんらかの基準（例：奇遇法，前半後半）に基づいて2つに分割し，両得点間の相関関係を計算する．

例えば，受験者1は50項目から成るテストで$\frac{15点}{項目1\sim25}$と$\frac{13点}{項目26\sim50}$を得点したとする．これら2つの得点にはあまり差がない．一方，受験者2は$\frac{1点}{項目1\sim25}$と$\frac{13点}{項目26\sim50}$を得点したとする．前半と後半の得点には大きな隔たりがあり，これはテストの全体の不均衡さを示唆している．なお，全受験者の後半の得点に対する前半の得点との相関は，エクセルの *CORREL* 関数，*CORREL*（range1sthalf：range2ndhalf）を用いて求められる．ただし，そこで算出される相関係数は半分の長さのテストを基準にしているため，テスト全体の信頼性係数を計算するには，スピアマン・ブラウンの公式「信頼性係数=2×相関係数／（1＋相関係数）」を用いて修正する必要がある（斉田［2011］p.49）．

折半法は項目難易度や受験者の能力分析よりもテストの全体的な不均衡さについて教示してくれるため，テスト項目の配置を決めるときに非常に役立つ．しかし，単にバランスを保つために，テストの最初のセクションに難しい項目を多数配置することは，一部の受験者に心理的負担を与えかねないので注意が必要である．

3　ラッシュモデルによる分析手法

ラッシュモデルを搭載したソフトウェアは様々あるが，本研究では Winsteps Version 3.81.0を用いる．異なるレベルの分析に対応できる多種多様な機能を備えているが，本章では5つの主要な分析ツールのみ扱う．それらはテスト開発者に最も直接的な情報を提供するだけでなく，統計に関する高度な知識が無くても利用しやすいからである．

(1)　Summary and Fit Statistics

表1はラッシュモデル分析を用いたサンプルデータの基本統計量とフィット（*Fit*）統計量である．フィットとはテスト分析データがラッシュモデルと一致する度合いのことを指す．下記は，Winsteps がフィットを計算する際に行う手順の概略である．手順を直線的に示してあるが，実質的には何回か

の反復を経て可能な限り最も低い残差得点（Maximum Score Residual：MSR）に達するまで，プログラムは何度もその手順を実行する．

1. 素点データを読み取る．
2. 各受験者の合計得点を集計する．
3. 各項目の項目難度（または項目測定値）を計算する．
4. 項目を困難度の尺度上に配置する．
5. 各受験者の項目別の期待値を算出する．（総合得点を尺度の特定の位置に配置するが，これは各項目の困難度に照らして判断する）．
6. 残差（予測値と観測値の差）を計算する．
7. 現在のモデルから得られた情報を用い，繰り返し更新する過程を通して，モデルを改善する．
8. それ以上小さい MSR を計算できなくなったときに反復は止まる．

CTT とラッシュモデルの決定的な違いは，ラッシュモデルは受験者の回答が，必ずしも彼らの真の能力を反映しないことを理解している点である．例えば，多肢選択式テストにおいて，ある受験者は当て推量で正解するかもしれない．あるいは，高得点を取得できる受験生が，ケアレスミスで不正解になるかもしれない．テストの得点はこのような外的要因によって常に影響を受ける．

ラッシュモデルでは，受験者能力とテスト項目という2つの変数が関係している．10題からなるテストを例として，各変数を順に見ていくことにする．受験者は次の3つのいずれかに属すると仮定する．

1）能力が高すぎる→10問中すべて正答した．
2）能力が低すぎる→10問中すべて誤答した．
3）能力は1）と2）の間である．→10問中いくつか正答した．

CTT およびラッシュモデルは，1）能力が高すぎる，および，2）能力が低すぎる，に該当する受験者を事実上除外する．CTT は受験者に10点（100%）および0点（0%）という得点をそれぞれ与える．一方，ラッシュモデルは「極端な受験者」として分類し，彼らのデータを分析に含めない．別な見方をすれば，1），2）のようなテスト項目は，受験者のレベルを適切に測定しないということである．能力が高すぎる受験者にはより難しい項目，他方にはより易しい項目が必要である．CTT とラッシュモデルを効果的に利用できるのは，受

験者の得点が1点から9点の範囲内に収まる場合だけである．

では，ある受験者が10点中9点を取得した場合はどうだろうか．誤答が一つあるが，最終得点は分析可能な範囲内の最高得点である．得点に着目すれば，当該受験者は能力が高いと一見判断できるかもしれない．だが，テスト開発者はその誤答した項目にこそ着目すべきである．受験者は問題の中身を理解できなかったのか，それとも不注意による誤答だったのか，その原因を調べる必要がある．同様に，10点中1点しか取得できなかった受験者も同じである．問題を理解して正答したかもしれないし，単なる憶測で正答したかもしれない．それでも，10点中9点を取得した受験者は，10点中1点を取得した受験者よりもレベルは高い．テストでは，当て推量による正答やケアレスミスによる誤答といった実力とは違う「想定外」の回答が出てくる．これらを正しく判断するには，項目容易度（*IF*）が分析できなければならない．

いかなる数理モデルも仮説を立てる必要がある．CTT における項目容易度（*IF*）の基本的な考えとラッシュモデルがどのように項目の難易度を算出するかには1つの決定的な相違がある．*IF* はその項目に正答した受験者数を総受験者数で割って算出した正答率のことを指す．そこに個人の総合得点は一切考慮されていない．そのため，CTT ではもう1つの分析手法である項目弁別力（*ID*）を算出する必要がある．受験者の合計点に基づき，受験者を上位，中位，下位群に分け，あるテスト項目に対して，上位群の *IF* 値から下位群の *IF* 値を引いたものがその項目の *ID* となる．*ID* 指数は欠陥のあるテスト項目を発見するのに非常に役立つ．しかし，*IF* と *ID* を用いても，個々の受験者が当て推量で正答したかどうかを判断するのは実際には難しい．

一方，ラッシュモデルは個別のテスト項目の難易度を計る際，個人の合計得点を含めて分析するアルゴリズムを採用している．

$$log_e\left(\frac{P_{ni1}}{P_{ni\phi}}\right) = B_n - D_i$$

$$P_{ni1} + P_{ni0} = 1$$

受験者が問題を正答する確率（*P*）は，項目難度（*D*）と受験者の能力値（*B*）が連動した結果である．能力値 *B* の受験者（*n*）がある項目で1（すなわち，正答）を得る確率をその受験者が0（すなわち，誤答）を得る確率で割ったものは，そ

の項目に対する困難度を取り除いた受験者の能力と等しい．このモデルに立てられた仮説では，正答総数が，その受験者の能力に関していくらかの評価を提供すること，および全項目に対して単なる偶然の結果ではないことを含んでいる．

表1　ラッシュモデル分析によるサンプルデータのSummary and Fit Statistics

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
MEAN	7.9	15.0	52.44	7.78	.96	-.2	1.03	.1
S.D.	2.5	.0	14.67	.87	.44	1.2	.85	.9
MAX.	13.0	15.0	85.76	10.96	1.71	1.8	2.95	2.2
MIN.	4.0	15.0	30.58	7.38	.33	-2.0	.25	-1.1

REAL RMSE 8.55 TRUE SD 11.92 SEPARATION 1.39 Person RELIABILITY .66
MODEL RMSE 7.83 TRUE SD 12.41 SEPARATION 1.59 Person RELIABILITY .72
S.E. OF Person MEAN = 3.92

Person RAW SCORE-TO-MEASURE CORRELATION = 1.00
CRONBACH ALPHA（KR-20）Person RAW SCORE "TEST" RELIABILITY = .69

SUMMARY OF 13 MEASURED Item

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
MEAN	7.9	15.0	50.00	7.37	.95	-.1	1.03	.3
S.D.	4.1	.0	18.34	1.19	.27	.8	.73	.9
MAX.	14.0	15.0	77.72	10.65	1.41	1.4	2.65	2.6
MIN.	2.0	15.0	19.11	6.20	.55	-1.3	.26	-.9

REAL RMSE 7.82 TRUE SD 16.58 SEPARATION 2.12 Item RELIABILITY .82
MODEL RMSE 7.47 TRUE SD 16.75 SEPARATION 2.24 Item RELIABILITY .83
S.E. OF Item MEAN = 5.29

(2)　Point-Measure Correlation（*PT*）

表2の *PT* 係数は，CTT の *ID* といくつかの点で類似しており，各テスト項目の困難度とテスト全体の困難度との相関を示している．1.0という値は，能力の低い受験者全員がその項目を間違い，能力の高い受験者全員が正解したことを意味する．表2の .00という数字は，q2（question2，問2）の困難度とテストの残りの部分には相関関係がないという情報をテスト開発者に提供す

る．負の値は，得点の低い受験生の方が得点の高い受験生より正解しているということであり，改訂すべき項目であることを示唆している．

表2　ラッシュモデル分析によるサンプルデータの Point-Measure Correlation

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT		PTMEASURE-A		EXACT MATCH		Item
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.	OBS%	EXP%	
2	14	15	19.11	10.65	1.22	.5	1.99	1.0	.00	.24	93.3	93.2	q2
11	7	15	53.72	6.20	1.41	1.4	2.65	2.6	.22	.54	66.7	74.5	q11
12	4	15	66.05	6.89	1.39	1.1	2.11	1.4	.24	.52	73.3	79.3	q12
6	13	15	27.65	8.12	1.12	.4	.79	.3	.28	.33	86.7	86.4	q6
5	12	15	33.41	7.14	.99	.1	.67	.1	.43	.40	86.7	80.7	q5

(3)　Variable Maps

図1は受験者の能力値と項目難度を縦軸のスケール上に同時に示している．これは二つのデータを同一の尺度で比較可能なものにしていることを表す．中央の線から左側に各受験者，右側に各テスト項目（問題），左端に尺度がある．スケール上部に行くほど受験者の能力値と項目難度は上がる．例えば，最上部にある q10の左側には受験者がいない．これは，難度が高すぎて正答した受験者がいない項目であることを示している．同様に，下部の q7，q2，q6は受験者の能力よりも低い項目で，これらの易しすぎた項目は受験者集団の能力推定には役立たないことを示唆している．また，1つの項目に多くの受験者が並んでいる q3は，受験者たちを弁別するための情報に乏しく，少ない受験者に対して複数の項目が並んでいる q13，q14は，同じレベルの問題が多すぎることを表している．ここでも，テスト項目は目的に応じて分析する必要があることがわかる．

図1　ラッシュモデルによるサンプルデータの Variable Map

INPUT: 15 Person 15 Item REPORTED: 15 Person 15 Item 2 CATS WINSTEPS 3.81.0

```
MEASURE          Person - MAP - Item
                   <more>       <rare>
    90                      +     q10
                        s14   T
    80                      T+
                                  q8

    70                  s12  +S    q13   q14
                             S     q12   q15
    60        s03   s06   s15  +
                        s10
                  s02   s07  M     q11
    50                     +M
                  s04   s13         q1    q4
    40                  s09  +
              s05   s08   s11  S     q3
                                    q5    q9
    30                  s01  +S
                                    q6
    20                      T+
                                    q2
                             T
    10                             q7
                   <less>       <frequent>
```

(4)　Item Category ／ Option ／ Distractor Frequencies

表3は，受験者がどの選択肢（1 ～ 4）を選んだか，選択肢毎の受験者数とその割合（%）を表している．また，各選択肢に対する受験者の平均的な能力値も示している．これらを複合したデータは，テスト項目を改善するのに役立つ情報を提供してくれる．

表3の q11 を例に見てみると正解は得点値（Score Value）欄の1で示され，選択肢（Data Code）1 ～ 4の中の選択肢3が正解である．データカウント（Data Count）欄は，（このサンプルデータの受験者数は15人）何名の受験者がそれぞれの選択肢を選んだかを示している．この場合，正答した受験者は7名（47%）で，その平均的能力（Average Ability）は55.96% であった．平均的能力が36.47% だった受験者1名（7%）だけが，選択肢1を選択した．一方，平均的能力が52.83% だった受験者5名（33%）は選択肢4を選択した．正解の3と不正

解の4を選んだ受験者の平均的能力はかなり近い．このことからq11がそのレベルの受験者を適切に分別できたとは言い難い．

表3　ラッシュモデル分析によるサンプルデータのItem Category/Option/Distractor Frequencies

ENTRY NUMBER	DATA CODE	SCORE VALUE	DATA COUNT	%	AVERAGE ABILITY	S.E. MEAN	OUTF MNSQ	PTMA CORR.	Item
11	A1	0	1	7	36.47		.2	-.29	q11
	2	0	2	13	47.16	16.58	1.6	-.14	
	4	0	5	33	52.83	8.67	6.1	.02	
	3	1	7	47	55.96	4.31	1.3	.22	
12	B3	0	2	13	44.57	8.10	.4	-.21	q12
	2	0	7	47	51.16	4.61	1.0	-.08	
	1	0	2	13	53.06	16.60	2.0	.02	
	4	1	4	27	58.31	10.73	2.5	.24	
2	C1	0	1	7	52.67		2.1	.00	q2
	2	1	14	93	52.42*	4.21	1.1	.00	

4　事例研究

本事例研究のデータには，著者らが執筆した英語教育用教科書『*Nursing Care*』(Smiley and Masui [2013])の付属テストを用いた．この教科書は，大学教育課程の一環として英語を学ぶ看護学生1～2年生向けに書かれたものである．英語のレベルとしては，およそ英検3級から準2級を想定している．文部科学省が高校終了時に達成すべき目標を英検準2級としていることから(Commission on the Development of Foreign Language Proficiency [2011])，大学の英語コースにふさわしい教科書であるといえる．

本書は12ユニットから構成されており，ユニット1～6用およびユニット7～12用にそれぞれ同じ形式のテストが用意されている．テストはリスニング20問(q1～q20)，リーディング30問(q21～q50)から成り，4つの選択肢から正解を一つ選ぶ形式になっている．

本研究では，ユニット前半用のテストを使用した．学生は関連するユニットの学習を全て終えてから各テストを受けるため，基本的には目標基準準拠テスト（criterion-referenced test：CRT）と位置づけられるが（Hughes［1989］），集団基準準拠型（norm-referenced test：NRT）の内容もある程度含まれている．受験者は著者らの授業を履修する143名で，看護を専攻している．学生の英語力には大きなばらつきがあるが，各クラスの平均的な能力はほぼ同等と思われる．しかし，中等教育での習熟が不十分な学生も含まれていたため，部分的に未回答の答案を提出した者もいたが，分析には十分有用なデータを得ることができた．

4.1　古典的項目分析結果

表4はCTTを用いて分析した基礎統計量を示している．各項目の配点は1点で，このテストで取得できる最高点は50点である．

表4　『Nursing Care』テストの基礎統計量

Mean	26.1
SD	6.8
Max	41
Min	12

テスト結果は50% ～ 60% の範囲で平均値を示したが，先に述べたように，英検準2級レベルの実力に達していない学生がかなりいる．このコース全体にわたるタスクは，看護に関する内容を学ぶのと同時に，関連する基礎的な英語力を身につけることである．これらを考慮すれば，平均値26.1は妥当な値であるといえる．

(1) 項目容易度（IF）

表5　『Nursing Care』テスト項目容易度（*IF*）

上位5位		下位5位	
項目 #	IF	項目 #	IF
q24	.87	q26	.19
q15	.86	q47	.17
q18	.80	q44	.14
q40	.80	q29	.10
q9	.79	q37	.08

表5は項目容易度（*IF*）の上位5項目と下位5位項目を示したものである．適正な値の範囲は .30 ～ .70であることから，.85を超えている q24と q15は易しすぎるテスト項目である可能性を示唆している．これらは，今後このテストを改善していく中で，削除，または難度を変更すれば適正な項目になるはずである．一方，.10未満になったのは q37だけである．これは極端に難しい項目であることを示しており，原因を調査する必要がある．q29から q26もすべて .2未満の値を示しており，これらもまた難しすぎた可能性がある．

このことから，修正すべき項目が確実に3つあり，さらに詳しく分析する必要のある項目がその他に5 ～ 6つあることが分かった．

（2）　項目弁別力（*ID*）

表6　『Nursing Care』項目容易度（*ID*）

.40 and above	24
.30 to .39	8
.20 to .29	6
below .20 （negative correlation）	12 （1）

先にも述べたが，Brown［2011］はどの項目が高得点と低得点をうまく弁別するかについての指標を提唱している（表6参照）．.40以上は信頼できる項目，.39 ～ .3は，比較的良いとみなされる項目，.29 ～ .2は，改善が必要な項目である．項目の難易度を調整するかは，この *ID* によって決定される．*ID* は高い群と低い群のそれぞれで，正確に得点している受験者数を基に算出される．しかし，数値自体は直接情報を提供しないため，その項目に対する回答を詳細に調べる必要がある．なお，.2を下回る得点は，高い群と低い群の分別には役立たない．このテストでは，q27だけが負の相関を示した．これは，低い群の受験者の方が高い群の受験者よりも高得点を取得していたことを意味する．この項目は削除または作り直す必要がある．

(3) 折半法 (split-half method)

表7 折半法による相関係数と受験者平均点

Correlation Coefficient	.68
Average Score q1-q25	59.7
Average Score q26-q50	42.1

折半法を用いて相関関係を調べるため，50項目から成るテストを前半と後半に分け平均点を求めた．表7は各グループの平均とその相関係数である．折半法によって計算された信頼性係数は .68 を示し，テストが不均衡である可能性を示唆した．具体的には，前半のテスト項目群の方が後半の項目群よりも統計的にかなり易しいことが分かった．

さらに，2グループ間の平均の差を検証するため，対応のある2標本 t 検定 (Paired Two-Sample T-test) を行ってみたところ，$t = 14.4458$, $df = 142$, $p\text{-}value < 2.2e\text{-}16$ となり，有意差が観察された．なお，「対応がある」とは「グループ同士が関連している (related) 同一の実験参加者である」という意味である (寺内，中谷 [2012] p.100)．

受験者群がリーディングのスキルよりリスニングのスキルにはるかに熟達していた可能性もあるが，その不均衡が項目自体によるものか，あるいはその他の要因によるものかを詳しく検証する必要がある．

4-2 ラッシュモデルによる分析結果

(1) Summary and Fit Statistics

Winsteps の基本統計量は，エクセルで算出できるものと同様の情報 (平均，最大値，最小値，標準偏差) だけではなく，個人の能力値とテストの項目難度も示すことができる (表8参照)．

表8　ラッシュモデルによる Summary and Fit Statistics

INPUT：143 Person　50 Item　REPORTED：143 Person　50 Item　2 CATS　WINSTEPS 3.81.0

SUMMARY OF 143 MEASURED Person

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
MEAN	26.1	49.8	50.94	3.30	.99	.0	1.02	.1
S.D.	6.8	.8	7.14	.16	.15	1.0	.24	1.1
MAX.	41.0	50.0	68.69	4.07	1.43	3.2	1.70	3.2
MIN.	12.0	43.0	35.63	3.17	.69	-2.7	.53	-2.4

REAL RMSE 3.39　TRUE SD 6.29　SEPARATION 1.85　Person RELIABILITY .77
MODEL RMSE 3.31　TRUE SD 6.33　SEPARATION 1.92　Person RELIABILITY .79
S.E. OF Person MEAN = .60

Person RAW SCORE-TO-MEASURE CORRELATION = 1.00
CRONBACH ALPHA（KR-20）Person RAW SCORE "TEST" RELIABILITY = .79

SUMMARY OF 50 MEASURED Item

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
MEAN	74.5	142.5	50.00	1.99	1.00	-.1	1.02	.0
S.D.	30.2	.7	11.11	.28	.09	1.3	.16	1.4
MAX.	125.0	143.0	76.79	3.08	1.30	4.1	1.38	3.8
MIN.	12.0	141.0	29.75	1.77	.83	-2.5	.72	-2.3

REAL RMSE 2.05　TRUE SD 10.92　SEPARATION 5.33　Item RELIABILITY .97
MODEL RMSE 2.01　TRUE SD 10.92　SEPARATION 5.43　Item RELIABILITY .97
S.E. OF Item MEAN = 1.59

分析データがどの程度ラッシュモデルに適合しているかを見る指標として，インフィット（INFIT）とアウトフィット（OUTFIT）がある．一般的に，インフィット平方平均（INFIT MNSQ）とアウトフィット平方平均（OUTFIT MNSQ）の値は .70 ～ 1.30 の範囲に収まっていれば良いとされている．この基準から外れている場合はミスフィット（不適合）と解釈する（靜 [2007]）．表8をみると，受験者能力に関する INFIT MNSQ と OUTFIT MNSQ はそれぞれ .99 と 1.20 で非常に良い．また，項目難度に関する INFIT MNSQ と OUTFIT MNSQ も 1.00 と 1.02 で，分析データがラッシュモデルに非常に適

合していることを示している．これらは，受験者能力に関する信頼性係数 .79 および項目難度に関する信頼性係数 .97 にも反映されている．

(2) Point-Measure Correlation

表9 ラッシュモデル分析による Point-Measure Correlation

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT		PTMEASURE-A		EXACT MATCH		Item
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.	OBS%	EXP%	
27	51	142	57.39	1.84	1.30	4.1	1.37	3.8	-.09	.31	52.8	67.6	q27
33	36	143	63.01	2.01	1.16	1.6	1.38	2.5	.02	.29	74.1	75.3	q33
26	28	143	66.52	2.19	1.11	.9	1.36	1.9	.05	.26	81.8	80.5	q26
44	20	143	70.84	2.48	1.11	.7	1.28	1.2	.05	.23	86.0	86.0	q44
37	12	142	76.79	3.08	1.07	.4	1.24	.8	.06	.19	91.5	91.5	q37

表9は得点測定相関に関して下位5位までの項目を示している．q27だけが負の値であった．これは，CCT の *ID* でも指摘された項目で，残りはすべて .1 未満であった．全体の表（スペースの都合でここには再現していない）では，わずか15項目だけが .4以上の値を示した．Holster[2014]が推奨しているカットオフ値である．このような PT 係数は，全体として異なる能力レベルの受験者を弁別する測定道具として，うまく機能していないことを示唆している．CTT の *ID* は12の項目を問題視したが，ラッシュモデルでは35項目を検証すべきだとしている．

(3)　Variable Maps

図2ラッシュモデルによる項目難度と受験者の能力推定値の分布

```
MEASURE                          Person - MAP - Itemv
                                 <more>          <rare>
   77                                    +    q37
   75                                    +    q29
   72                                   +T
   70                                 1  +    q44
   67                                 0  +    q26  q47
   65                               0 0 0  T+   q21  q39
   63                   0 0 0 0 0 0 0 0 0 1 1  +    q33
   60                         0 0 0 0 1  +S   q31  q35  q46
   58                   0 0 0 0 0 0 0 0 0 1 1  S+   q20  q27  q42
   55       0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1  +
   53         0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1  +    q23  q25  q32  q36  q41
   51   0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1  M+M  q14  q16  q49   q5
   48             0 0 0 0 0 0 0 0 0 0 1 1  +     q1  q12  q30   q4   q8
   46           0 0 0 0 0 0 0 0 1 1 1 1 1 1  +    q10  q17   q2   q3  q48  q7
   43         0 0 0 0 0 0 0 1 1 1 1 1 1 1 1  S+   q28  q45
   41                   0 0 0 0 0 0 1 1  +    q19  q22  q34  q38  q43
   39                         0 0 1 1  +S   q11  q50   q6
   36                         0 1 1 1 1  T+   q13  q18   q9
   34                                    +    q40
   31                                    +    q15
   29                                    +    q24
                                 <less>          <frequent>
```

図2の中央から左側にある1と0は受験者を表し（1と0の差に意味はない），右側の「q 数字」はテスト項目を示す．上部 q37 と q29 には対応する受験者がいないことから，この2つの問題は極端に難しく，受験者集団の能力を超えていることを意味する．一方，下部の q40，q15，q24 は受験者には易しすぎる項目であることを示している．つまり，これら5つの項目は，テストの質に何の影響も与えていない不必要な項目であることがわかる．

図の中央では，多くの受験者が多くの問題に並んでいるのがわかる．例えば，左側の測定値46を見ると，46の能力値を持つと推定された14名の受験

者がいる．その右側には，同じレベルの項目が6つもある．これらの項目は，このレベルの受験者を分別するのに役立っていないことを示している．同じ分別機能を果たす項目がたくさんあるときは，難易度を変更して，これらの項目を見直すべきである．

143名の受験者に50項目のテストを実施するとき，各項目でおよそ3名の受験者を測定できれば理想的である．図2を見ると，このテストには同じレベルの項目が多すぎる，または少なすぎることが分かる．これは対応する項目がない測定値55で一目瞭然である．言い換えれば，q23，q25，q32，q36，q41の5項目すべてが，測定値54 ～ 57の受験者を弁別するのに機能している状態である．

(4)　Item Category ／ Option ／ Distractor Frequencies

表10　ラッシュモデル分析による Item Category ／ Option ／ Distractor Frequencies

ENTRY NUMBER	DATA CODE	SCORE VALUE	DATA COUNT	%	AVERAGE ABILITY	S.E. MEAN	OUTF MNSQ	PTMA CORR.	Item
33	A1	0	1	1	38.08		.2	-.15	q33
	3	0	1	1	46.63		.6	-.05	
	2	0	105	73	51.00	.67	1.1	.01	
	4	1	36	25	51.24	1.33	1.5	.02	
27	B2	0	6	4	45.37	2.69	.6	-.16	q27
	3	0	82	58	51.76	.82	1.3	.13	
	4	0	3	2	55.20	2.30	1.5	.09	
	1	1	51	36	50.03*	.94	1.4	-.09	
	MISSING	***	1	1#	50.37			-.01	
	C1	0	1	7	52.67	2.1	.00	q2	
26	C1	0	3	2	43.55	1.05	.4	-.15	q26
	4	0	14	10	43.72	1.74	.5	-.33	
	2	0	98	69	51.99	.64	1.2	.22	
	3	1	28	20	51.66*	1.56	1.4	.05	
[Selected Items]									
47	F3	0	57	40	48.72	.87	.8	-.25	q47
	4	0	16	11	52.46	2.03	1.3	.08	
	1	0	45	31	52.70	1.08	1.3	.17	
	2	1	25	17	51.86*	1.36	1.3	.06	
15	K1	0	4	3	43.89	1.30	.6	-.17	q15
	2	0	8	6	45.52	3.31	1.1	-.18	
	3	0	7	5	52.74	1.41	1.5	.06	
	4	1	124	87	51.41*	.63	1.0	.17	
8	U1	0	9	6	46.69	3.25	1.1	-.15	q8
	4	0	38	27	47.45	.89	.8	-.29	
	3	0	12	8	53.37	2.15	1.7	.10	
	2	1	83	58	52.64*	.75	1.0	.28	

Winstepsは受験者のデータがラッシュモデルにどの程度適合しているかによってこの表を整理している（表10参照）．あまりうまく適合しなかった項目は，右から3番目の欄の「OUTF MNSQ（OUTFIT MEAN SQUARE）」で確認できる．

さらに，テスト開発者にとても貴重な情報を提供してくれる欄は他にも2つある．データカウント（Data Count）は何名の受験者がどの選択肢を選んだかを示し，平均的能力（Average Ability）は，受験者の全体での平均的なレベルを表す．これらの数値は，各項目の選択肢がどれだけうまく様々なレベルの受験者を弁別したのかを正確に示してくれる．値の横のアスタリスク（*）は，正答した受験者の平均的なレベルが，誤答した受験者のレベルよりも低いことを意味している．高得点者が正しい選択肢を選び，低得点者がその他の選択肢を選ぶのが理想であり，本テストではそれが9回起こったことが分かった．

q27において正解を選んだのは，受験生の36%でレベルは50.03%だった．誤答選択肢3を選んだのは，58%でレベルは51.76%だった．レベル差はわずか1.73%であり，これらの受験者たちはほぼ同レベルであると判断できる．誤答選択肢4を選んだ受験者のレベルは55.2%だったが，受験者は2%にすぎなかった．テスト開発者は，なぜ高いレベルの受験者がこの項目を間違えたのか検証する必要がある．しかし，受験生の数としてはわずか3名なので，ケアレスミスと判断した方が説得力はあるかもしれない．

だが，それ以上に問題視すべき項目はq15である．最も高いレベルの受験者（5%）が誤った選択肢を選び，次の下位レベルの受験者（87%）が正答した．ここでも，テスト開発者はその間違った要因を調査する必要がある．

本研究で使用したテストに関する最大の懸念は，各選択肢が受験者の能力を適切に弁別していないことである．理想的なテストとは，表11のような結果になるものを指す．ここでは，3つの誤答選択肢間における差，さらに正答と誤答選択肢との間に明らかな境界がある．一方，本テストでは，ほとんどの項目がわずか数％の差で分けられていただけであった．

表11　ラッシュモデル分析による理想的なテストデータ

ENTRY NUMBER	DATA CODE	SCORE VALUE	DATA COUNT	%	AVERAGE ABILITY	S.E. MEAN
xU	1	0	9	6	26.69	3.25
	4	0	38	27	47.45	.89
	3	0	12	8	33.37	2.15
	2	1	83	58	62.64	.75

5　終わりに

CTT もラッシュモデルも，ともに改訂すべき項目を的確に指摘した．q27 は，*ID* 値と *PT* 係数で負の相関を示した．CTT の *IF* と *ID* の値は，改善の余地がある多くの項目を明らかにした．しかし，いずれの *IF* 値もそれだけではどの項目も断定的に批判するのに不十分だった．*IF* は問題のある項目に関する手がかりは提供したが，1つずつ，各項目をさらに分析しなければならなかった．*ID* は受験生をわずかに区別する項目が12あったことを示した．一方，Winsteps の *PT* 係数は，30を超える疑わしい項目を指摘した．

以上のことから，CTT の *IF* と *ID* は，潜在的な問題をまず抽出するという点からすれば，テスト分析ツールとして十分機能し，非常に有益な分析方法であると言える．しかしながら，ラッシュモデルは CTT を超えるさらに詳細で強力な分析ツールを提供する．本研究で示されたように，基本的な統計知識を身につけ，ラッシュモデル分析が行える統計ソフトウェアの主要ツールの使い方さえ習得できれば，文系の教員でも精度が高いテスト分析を行うことが十分可能になる．安定した結果を示す質の高いテストの開発，実施をしていくには，テスト理論の有用性をさらに示していく必要があろう．本稿がテスト理論を理解する上での一助となることを期待したい．

参考文献

Bond, T. G. & Fox, C. M. [2007] *Applying the Rasch Model : Fundamental Measurement in the Human Sciences,* New Jersey : Lawrence Erlbaum Associates.

Brown, J. D. [2011] *Testing in Language Programs,* New York : McGraw-Hill.
Choppin, B. [1983] "The Rasch Model for Item Analysis", in *CSE Report,* Vol. 219, pp.1-32, Center for the Study of Evaluation, Graduate of Education, University of California in Los Angels.
Linacre, J. M. [2014] Winsteps®（Version 3.81.0）[Computer software]. Chicago : MESA Press.
Holster, T. A. & Lake, J. [2014] "How High Can They Jump : An introduction to Rasch Measurement", in *Studies in the Humanities,* Vol. 78, pp. 148-122, Faculty of Literature, Fukuoka Women's University.
Hughes, A. [1989] *Testing for Language Teachers,* Glasgow : Cambridge University Press.
Commission on the Development of Foreign Language Proficiency. [2011] "Five Proposals and Specific measures for Developing Proficiency in English for International Communication", Ministry of Education, Culture, Sports, Science and Technology. Online accessed : August 6, 2014. http : //www.mext.go.jp/component/english/__icsFiles/afieldfile/2012/07/09/1319707_1.pdf
Smiley, J. & Masui, M. [2013] *Nursing Care,* Nagoya : Perceptia Press.
大友賢二［1996］『項目応答理論入門』大修館書店 .
小山由紀江［2010］「テストの歴史的変遷とコンピュータ適応型テストの意義」『New directions ／名古屋工業大学共通教育・英語 編』第28巻，pp.13-25.
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 NO.254，pp.59-78.
斉田智里［2011］「第2章 英語学力測定論」『テスティングと評価――4技能の測定から大学入試まで』大修館書店，p.38，p.49.
静哲人［2007］『基礎から深く理解するラッシュモデリング―項目応答理論とは似て非なる測定のパラダイム』関西大学出版部.
寺内正典・中谷安男［2012］『英語教育学の実証的研究法入門』研究社，p.100.
豊田秀樹［2002］『項目反応理論　入門編』朝倉書店.
三浦省五，前田啓朗・山森光陽，磯田貴道・廣森友人［2004］『英語教師のための教育データ分析入門――授業が変わるテスト・評価・研究』大修館書店.