項目応答理論によるテスト評価
―漢字テストを用いて―

2007MI022 深谷和典　　2007MI154 永田仁志
指導教員: 尾崎俊治

1　はじめに

私たちの人生においてテストは切り離せないものだろう．これまで，私たちは能力を測るために様々なテストを受けてきた．テストにはテスト理論と呼ばれるものが使われており，これを用いて能力を測定するのだが，従来のテスト理論ではいくつか欠点があり，能力を測るのに限界があった．そこで近年では，それを解消するために新しいテスト理論が登場してきた．本論文では，この新しいテスト理論について考察をしていこうと思う．考察をするにあたって，何らかのテストを用いる必要があるが，今回は私たちの身近な『漢字』を取り上げて，実際にテストを作成し，データを集め，新しいテスト理論について考察をしていきたいと思う．

2　テスト理論

従来，テストの作成や評価には，古典的テスト理論が用いられてきた．しかし，この古典的テスト理論には，能力を測定するのに限界があった．そこで近年では，項目応答理論と呼ばれる新しいテスト理論が導入されてきている．

2.1　古典的テスト理論

2.1.1　古典的テスト理論とは

古典的テスト理論（CTT: Classical Test Theory）を用いたテストでは，被験者が正答した項目の数を数え上げて得点とする正答数に基づく得点をもとに，平均値や分散，偏差値などのデータ分析をしたり，信頼性・妥当性の検討を行う([1]参照).

2.1.2　古典的テスト理論の問題点

古典的テスト理論には，次のような問題点がある([2]参照).
1）同一の測定単位が保てない.
2）平行テストの作成は不可能に近い.
3）被験者個人に対する測定の精度が求められない.

2.2　項目応答理論

2.2.1　項目応答理論とは

項目応答理論（IRT: Item Response Theory）は，項目反応理論とも呼ばれ，テストの難易度に依存しない連続尺度で被験者の特性（認識能力，物理的能力，技術，知識，態度，人格特徴等）や，テスト項目の難易度・識別力を測定するための統計理論である．この理論は米国やヨーロッパの多くの国で普及しており，日本でも英語試験であるTOEFLなどに使われている([1]参照).

2.2.2　項目応答理論の利点

項目応答理論には，次のような利点がある([4]参照).
1）測定精度をきめ細かく確認できる.
2）複数のテスト間の結果の比較が容易である.
3）平均点をテスト実施前に制御できる.
4）被験者ごとに最適な問題を瞬時に選び，その場で出題できる.

2.2.3　項目応答理論の前提条件

項目応答理論には次のような前提条件がある([2]参照).
【1】局所独立の仮定: ある問題に正答出来る確率は，他の項目に正答出来る確率に影響を受けない.
【2】一次元性の仮定: 全ての問題は，唯一の能力分野を測定するものである.
また本論文では，項目に対する2値のいずれかの応答（正解か不正解）で評価するといった条件も付け加える.

2.2.4　項目応答理論のモデル

項目応答理論において，テスト項目に対する被験者の特性値（能力値）と正答確率の関係を表現するために，次のようなモデル（項目特性関数: ICC）が存在する．①ロジスティックモデル，②段階応答モデル，③部分採点モデル，などである．本論文ではこのうち，ロジスティックモデルを取り上げる.

IRTに用いられるロジスティックモデルは「1母数ロジスティックモデル」,「2母数ロジスティックモデル」,「3母数ロジスティックモデル」の3種類があり，本論文では，3母数ロジスティックモデルを取り上げ，用いる変数は以下の通りである.

$P_j(\theta)$：項目jに正答する確率
θ：能力パラメータ
　被験者の特性（能力）の大きさを表す母数.
a_j：識別力パラメータ
　項目jが能力の高い被験者とそうでない被験者を識別する力を表す母数.
b_j：難易度（困難度）パラメータ
　項目jの難しさを決める母数．一般的には各項目に50%の正答率を持つ被験者の能力値を基準として決められる.
c_j：当て推量パラメータ
　多肢選択式のテストの場合に，項目jに被験者が偶然に正答する確率を表す母数.

基本的な考え方としては，能力パラメータと，項目の難易度パラメータの差をとり，ロジスティック曲線に当てはめ，正答する確率を求めるというものである.

3母数ロジスティックモデルのICCは，

$$P_j(\theta) = c_j + \frac{1 - c_j}{1 + e^{-Da_j(\theta - b_j)}} \tag{1}$$

である．3母数では難易度b_jと識別力a_j，当て推量 c_j の3変数を用いる．難易度b_jはおよそ -2.0 から 2.0 の間で推定され，-2.0 に近い項目は簡単で，2.0 に近い項目は難しいとされる．識別力a_jはおよそ 0.3 から 2.0 の間で推定される．当て推量 c_j は，実力では正解できない被験者が偶然正解する確率を表しているので，多肢選択項目では選択肢の逆数が目安の一つである．しかし，当て推量の真値は必ずしも目安には一致しない．また，定数Dは 1.701 という値で，ロジスティック関数を累積正規分布関数に近似するためのもので，θ の全域で確率に 0.01 以上の差が生じないようになっている([3]参照)．

2.2.5　項目特性曲線

a_j，b_j，c_jのような項目パラメータは，ある項目jの性質を示している．項目パラメータが定まると，受験者がその項目に正答する確率$P_j(\theta)$は各受験者の能力θ の１変数のみを持つ関数になり，縦軸に正答率，横軸に能力値としたグラフが描ける．このグラフを項目特性曲線と呼ぶ([3]参照)．

3母数モデルでは，識別力a_j，難易度b_j，当て推量c_jの3要素でグラフの形が決まる．3母数モデルの項目特性曲線の例として「a_j=0.7，b_j=-1，c_j=0.3」とした場合と「a_j=1.0，b_j=0，c_j=0.2」とした場合，「a_j=1.5，b_j=1，c_j=0.1」とした場合の3つのグラフを図1に示す．

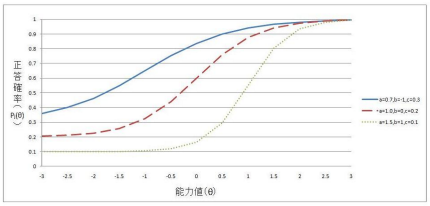

図 1: 3母数の項目特性曲線の例

仮に $\theta =$ 0.0 を平均的なレベルとしたら，このときそれぞれのb_jの項目に正答する確率は，横軸の 0.0 の点から垂線を立てて，曲線と交わった値であり，難易度の高い項目ほど正答しにくいことがわかる．また，$\theta > \theta'$ ならば $P_j(\theta) > P_j(\theta')$ ということがわかる．

識別力a_jが項目特性曲線の傾きを決定し，傾きが大きいほど，項目の難しさと人の特性の大きさに差があり，解答の正誤が分かれることを示している．識別力が高くなると，曲線が $\theta = b_j$付近で急激に立ちあがり，急激に立ちあがっている $a_j = 1.5$ の項目はある能力レベルで習得される項目で，緩慢に立ちあがっている $a_j = 1.0$ の項目はなかなか完全に習得されない項目である．

c_j は項目特性曲線の負の側の漸近線であり，$\theta = b_j$ における正答確率は 0.5 ではなく，$(1 + c_j)/2$ である([4]参照)．

2.2.6　テスト特性曲線

項目特性曲線は，各項目は互いに独立であるという前提を置いているので加法的である．すなわち，全ての項目特性曲線を足したものが求められる．これはテスト特性曲線と呼ばれ([3]参照)，横軸を被験者の能力値 θ，縦軸を被験者の真の得点(スコア) $T(\theta)$ とした曲線のグラフのことである．スコアは以下の式で求められる（N：項目数）．

$$T(\theta) = \sum_{j=1}^{N} P_j(\theta) \tag{2}$$

2.2.7　パラメータの推定

項目応答理論では，実際に集めたデータから，項目母数を推定する．推定方法には同時最尤推定法，周辺最尤推定法, ベイズ推定法，EMアルゴリズムなどいくつかあるが，どれも計算量が非常に多く，数学的には大変面倒である．そこでたいていは，専用のソフトウェアを利用する．有名なソフトとしては，RやBILOG-MG，Easy Estimationなどがある([5][6][7]参照)．

2.2.8　等化

複数のテスト間での比較をする為には異なるテストの得点が互いに交換可能になるような方法を考える必要がある．その方法はテストの等化と呼ばれ，古典的テスト理論では，こうした問題を解決することは非常に困難であるが，項目応答理論では比較的容易に可能である．その理由として，項目特性値が受験する受験者集団とは独立に求められることと能力特性値が受験するテストとは独立に求められることが挙げられる．

等化には，同一の能力水準に対して複数のテスト問題の難易度の間に共通の尺度を設定する「水平的等化」と異なった水準のテスト間に異なった尺度を設定する「垂直的等化」がある([2]参照)．

3　漢字能力測定

漢字テストを実際に作成し，学生に解答してもらい，データを集めた．このデータをもとに項目応答理論について考察をしていく．

3.1　漢字テストの問題形式

作成した問題の形式は，以下の通りである．

1）全て4択問題．
2）漢字の読み30問．漢字の書取り30問．四字熟語，部首，類義語，対義語がそれぞれ10問ずつの計100問．
3）問題の難易度は漢字検定3級から2級程度([8][9]参照)．
4）制限時間は20分．

3.2　テスト結果

今回行ったテストの被験者数は142人であった．

テスト結果は各項目の正誤を 0 か 1 (誤：0，正：1)で表し，表1のようにまとめた（一部省略）．表1をテストの項目反応データ，または反応パタンと呼ぶ．

表 1: 漢字テストの項目反応データ

s1	11111110010110010101···01110101110110111101
s2	11100101110101011101···11111101010101000101
s3	01110100100111110101···01110111110110100111
s4	01110100111110000101···01110111100110011101
:	:
s139	00111100111110010101···01010110110111001100
s140	00100101010101011011···11100111010110001001
s141	11110101111111010111···01110111110100111001
s142	11110101101111111101···01110111110110011111

3.2.1 古典的テスト理論を用いて

この結果をまずは，古典的テスト理論を用いてまとめる．

各項目の配点をすべて 1点とし，テスト得点の要約統計量をまとめたものを表2に示し，得点換算表を表3に示す．

表 2: テスト得点の要約統計量

平均点	54.718点	被験者数	142人
分散	171.112	中央値	55点
標準偏差	13.081	第1四分位	45点
歪度	0.081	第3四分位	64点
尖度	0.691	範囲	79
最高点	97点	四分位範囲	19
最低点	18点	最頻値	43点

表 3: テスト得点換算表

得点	順位	度数	累積度数	標準得点	偏差値
97	1	1	142	3.232	82.323
94	2	1	141	3.003	80.030
81	3	1	140	2.009	70.092
:	:	:	:	:	:
56	65	5	78	0.098	50.980
55	70	3	73	0.022	50.215
54	73	5	70	-0.055	49.451
:	:	:	:	:	:
35	137	1	6	-1.507	34.926
34	138	1	5	-1.584	34.161
28	139	1	4	-2.043	29.575
25	140	1	3	-2.272	27.281
18	141	2	2	-2.807	21.930

3.2.2 項目応答理論を用いて

次に項目応答理論を用いて考察していく．

今回の漢字テストは4択問題の形式なので当て推量を含む3母数ロジスティックモデルを用いる．パラメータの推定には統計ソフト「R」を用いて計算した([5]参照)．項目パラメータ, 被験者の能力パラメータの計算結果の一部を表4と表5に示す（表5は古典的テスト理論で，得点の高い順に並べてある.）.

表 4: 項目パラメータ

項目	識別力	難易度	当て推量
V1	1.180	-0.175	0.205
V2	1.459	-0.631	0.491
V3	0.538	-3.080	0.042
V4	0.532	-1.726	0.061
V5	0.953	1.042	0.420
:	:	:	:
V95	1.288	0.015	0.124
V96	0.825	1.539	0.245
V97	1.324	1.115	0.381
V98	0.538	-1.338	0.020
V99	1.649	0.288	0.222
V100	0.863	-2.292	0.043

表 5: 能力パラメータ

被験者	得点	能力パラメータ
s99	97	3.418
s85	94	2.822
s114	81	1.628
s86	79	1.795
s42	78	1.537
:	:	:
s47	59	0.092
s126	59	0.181
s130	59	0.314
s55	58	0.386
s106	58	0.226
:	:	:
s58	34	-1.305
s109	28	-2.734
s41	25	-2.761
s84	18	-3.642
s113	18	-2.913

まず，項目パラメータは，識別力がおよそ0から4の間の値になり，また難易度はおよそ−4から4の値になった．例えば，項目V2は識別力が高く，難易度がやや低いので，能力値がやや低い被験者を識別するのに向いていることがわかる．しかし，当て推量が高いので，能力が低い被験者も偶然正答する可能性も高いので注意が必要である．また，項目V3のように識別力がそれほど高くなく，難易度が低い項目は，全体的に正答率が高い問題であることが予想でき，また，能力値が低い被験者の識別には向いているとも取れる．

次に能力パラメータであるが，基本的には正当数の多い被験者の方が能力パラメータが高い値をとっている．しかし，被験者 s47 と s55 のように，CTTでの得点は s47 の方が高いが，能力パラメータは s55 の方が高い値をとっていることもある．これは項目応答理論が，偶然正答する可能性を考慮しているので，このようなことが生じたと考えられる．つまり，s47 は偶然の正答が多いだけで，実際は s55 の方が能力が高いと考えられる．

古典的テスト理論での得点は，今回のテストの良し悪しだけを表していたが，項目応答理論での能力パラメータは，被験者の能力の高さを表しているので，能力パラメータが高い被験者の方が能力が優れていると考えることができる．

この結果と式(1)から，被験者のスコアを求めると表6のようになる．

表 6: IRTでのスコア

	V1	V2	V3	···	V99	V100	スコア
a	1.180	1.459	0.538	···	1.649	0.863	
b	-0.175	-0.631	-3.080	···	0.288	-2.292	
c	0.205	0.491	0.042	···	0.222	0.043	
s99	0.99942	0.99998	0.99749	···	0.99988	0.99978	92.822
s85	0.99807	0.99990	0.99568	···	0.99936	0.99948	90.230
s114	0.97927	0.99814	0.98724	···	0.98226	0.99698	79.847
:	:	:	:	:	:	:	:
s126	0.73881	0.94012	0.95378	···	0.55307	0.97530	57.502
s130	0.78334	0.95548	0.95885	···	0.62518	0.97958	59.533
s55	0.80536	0.96223	0.96137	···	0.66412	0.98159	60.660
:	:	:	:	:	:	:	:
s41	0.20940	0.49356	0.59037	···	0.22215	0.36298	26.227
s84	0.20575	0.49129	0.40055	···	0.22201	0.15890	21.618
s113	0.20824	0.49276	0.55750	···	0.22210	0.31732	25.256

スコアを見ると，s126 と s130 はCTTでは，両者とも 59 点となっており差をつけることができなかったが，IRTでは57.502 と 59.533 となっており，差が出ていることがわかる．また， s126 と s55 のように，CTTでは， 59 点 と 58点 で s126 の方が s55 よりも点数が高いのだが，IRTでは，57.502 と 60.660 で s55 の方が s126 よりも値が大きくなっており，実際は s55 の方が能力が高いことがわかる．

次に，テストの等化についてみていく．これは複数のテストを比べるものだが，今回はテストを一度しか行えなかった．そこで，100問のテストを50問ずつに分け，テストAとテストBとして比べていく（奇数番号の項目を集めたものをテストA，偶数番号の項目を集めたものをテストBとする.）.

それぞれのテストの項目パラメータ，能力パラメータとスコアを表7から表10にまとめ，テスト特性曲線を図2と図3に描いた．

表 7: テストAの項目パラメータ

項目	識別力	難易度	当て推量
V1	1.837	0.339	0.407
V2	0.408	-3.738	0.024
V3	1.489	1.081	0.456
V4	7.382	0.856	0.296
V5	0.729	-0.620	0.557
:	:	:	:
V46	0.820	2.124	0.041
V47	0.203	-1.188	0.066
V48	0.994	-0.138	0.000
V49	8.576	1.024	0.424
V50	1.677	0.329	0.210

表 8: テストBの項目パラメータ

項目	識別力	難易度	当て推量
V1	0.983	-0.862	0.387
V2	0.518	-1.718	0.031
V3	0.359	-4.126	0.041
V4	1.176	0.770	0.295
V5	1.309	0.381	0.666
:	:	:	:
V46	1.002	-1.065	0.013
V47	0.228	1.699	0.040
V48	0.977	1.431	0.246
V49	0.542	-1.425	0.006
V50	1.334	-1.431	0.419

表 9: テストAの結果

被験者	能力パラメータ	スコア
s99	2.736	45.880
s85	2.526	45.180
s86	1.826	41.581
s42	1.519	39.489
⋮	⋮	⋮
s1	0.123	24.485
s130	0.112	24.407
s126	0.077	24.163
s81	0.048	23.965
⋮	⋮	⋮
s113	-2.254	14.046
s109	-2.369	13.748
s84	-2.483	13.468
s41	-2.614	13.163
平均	-0.014	25.222

表 10: テストBの結果

被験者	能力パラメータ	スコア
s99	2.653	47.024
s85	2.322	46.030
s114	1.792	42.752
s100	1.423	39.627
⋮	⋮	⋮
s24	0.104	30.670
s6	0.102	30.659
s19	0.096	30.626
s20	0.088	30.582
⋮	⋮	⋮
s113	-2.289	16.373
s109	-2.815	14.067
s136	-3.197	12.796
s84	-3.338	12.411
平均	-0.009	30.163

表 11: テストA→テストB

被験者	能力パラメータ	スコア（等化前）	スコア（等化後）
s99	2.736	45.880	47.116
s85	2.526	45.180	46.739
s86	1.826	41.581	44.627
s42	1.519	39.489	43.002
⋮	⋮	⋮	⋮
s1	0.123	24.485	30.265
s130	0.112	24.407	30.117
s126	0.077	24.163	29.647
s81	0.048	23.965	29.262
⋮	⋮	⋮	⋮
s113	-2.254	14.046	15.816
s109	-2.369	13.748	15.417
s84	-2.483	13.468	15.043
s41	-2.614	13.163	14.641
平均	-0.014	25.222	29.924

表 12: テストB→テストA

被験者	能力パラメータ	スコア（等化前）	スコア（等化後）
s99	2.653	47.024	44.847
s85	2.322	46.030	42.389
s114	1.792	42.752	38.067
s100	1.423	39.627	35.431
⋮	⋮	⋮	⋮
s24	0.104	30.670	27.449
s6	0.102	30.659	27.437
s19	0.096	30.626	27.402
s20	0.088	30.582	27.355
⋮	⋮	⋮	⋮
s113	-2.289	16.373	13.896
s109	-2.815	14.067	12.287
s136	-3.197	12.796	11.457
s84	-3.338	12.411	11.205
平均	-0.009	30.163	26.668

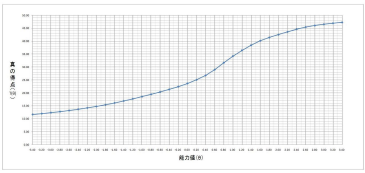

図 2: テストAのテスト特性曲線

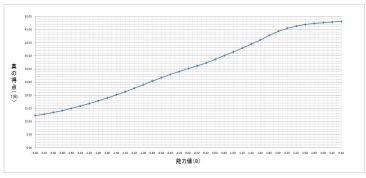

図 3: テストBのテスト特性曲線

今回は難易度比較での等化をする．方法は，それぞれのテストの難易度の平均をとり，その差をそれぞれの項目から引くというという方法である．

つまり，テストAをテストBに等化する場合は，テストAの難易度の平均は 0.39696 で，テストBの難易度の平均は -0.17482 なので，その差の 0.57178 をテストAのそれぞれの難易度から引き，これを新しい難易度としてスコアを求めていくということである．同様に，テストBからテストAに等化する場合は，難易度の平均の差である -0.57178 をテストBのそれぞれの難易度から引いて，スコアを求めればよい．

等化前と等化後の比較をわかりやすくするために，テストAからテストBに等化したテスト特性曲を描いたものを図4，テストBからテストAに等化したテスト特性曲を描いたものを図5に示し，表11と表12に等化前と等化後のスコアをまとめた．

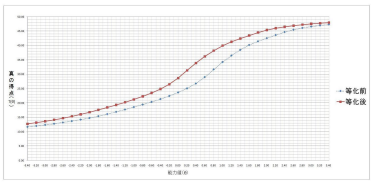

図 4: テストAを等化したテスト特性曲線

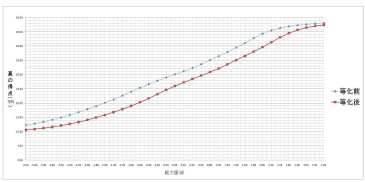

図 5: テストBを等化したテスト特性曲線

2つの表から，それぞれのテストを比べることができる．例えば，「 s86 のテストAのスコア」と「 s114 のテストBのスコア」のどちらが良いかを見てみる．s86 のテストAのスコアは 41.581 であり， s114 のテストBのスコアは 42.752 で，単純にスコアを比べるとs114 の方が良いのだが，s86 のテストAのスコアをテストBに等化すると 44.627 となり，s114 のテストBのスコアよりも良くなる．反対に，s114 のテストBのスコアをテストAに等化すると 38.067 になり，s114のテストAでのスコアは s86 のスコアよりも悪くなっている．つまり，被験者86の方が優れているということがわかる．このように項目応答理論では，複数のテストを比べることが可能なのである．これは，入社試験などで今年と去年の能力の比較をする際などに非常に効果的ではないかと考えた．

4 おわりに

今回の研究では，データ数が満足のいくものではなかった．しかしそれでも，項目応答理論の有用性を知るには十分な研究ができたといえるだろう．特に等化については，就職試験や入学試験など，あらゆる面での活躍が大いに期待できるものだと感じた．今後，項目応答理論が普及していけば，能力測定の精度が現在よりももっと正確なものになり，テストが今よりも意味があるものになるのではないかと考えた．

参考文献

[1] CiNii Article - 項目応答理論の概要(項目応答理論の基礎と応用-学力テストの分析を通して-,研究委員会企画チュートリアル)：
http://ci.nii.ac.jp/naid/110004570827

[2] 大友 賢二：『言語テスト・データの新しい分析法　項目応答理論入門』．大修館書店，1996.

[3] 項目応答理論 - Wikipedia：
http://ja.wikipedia.org/wiki/%E9%A0%85%E7%9B%AE%E5%BF%9C%E7%AD%94%E7%90%86%E8%AB%96

[4] 豊田 秀樹：『項目反応理論[入門編]-テストと測定の科学-』．朝倉書店，2002.

[5] Rで項目反応理論 - RjpWiki：
http://www.okada.jp.org/RWiki/?R%A4%C7%B9%E0%CC%DC%C8%BF%B1%FE%CD%FD%CF%C0

[6] SSI - Scientific Software International, Inc.：
http://www.ssicentral.com/irt/index.html

[7] 項目反応理論（IRT）とEasyEstimationのページ：
http://irtanalysis.main.jp/

[8] 資格試験対策研究会：『漢字検定3級〔頻出度順〕問題集』．高橋書店，2010.

[9] 資格試験対策研究会：『漢字検定2級〔頻出度順〕問題集』．高橋書店，2010.