報　告

項目反応理論とテスト結果の診断

研究開発部助手　鈴　木　規　夫
（情報処理研究部門）

1　はじめに

世の中，テストと名のつくものはいたるところで氾濫している。我々が幼いときから経験した学校の授業の一環として行われるテストや，学習塾で行うテスト。また，高校や大学へ進学を希望する者の選抜（入学）試験あるいは会社や公共機関で実施する採用試験の他，性格検査や知能検査等もみなテストの範疇に入る。いずれにしてもこれらのテストは，学力，知能，性格，興味等の心理的特性を測るために使われている。従って，測定の道具として使われるからには，できるだけ正確に個々の特性が測られる必要があろうが，実際にはなかなか難しい問題である。

これらの問題は，古くから理論的研究は数多くなされているが，現在にいたってもなお十分な結論が出されていない。その中で近年特に注目をあびているテスト理論の一つとして項目反応理論（Lord, 1980）がある。筆者は幸い文部省の在外研究員として，米国イリノイ大学へ1986年６月から10カ月間滞在する機会を得て，教育工学研究所のDr. Tatsuoka夫妻の指導のもと，この項目反応理論に関連した研究を行うことができた。そこで，本稿ではこの項目反応理論の基本的な考え方と最近の応用例について述べたいと思う。

2　項目反応理論

今，テストがｎ個の項目から構成されているとき，生徒がこのテストに反応するパターンは2^n通りの組み合わせが考えられる。例えば，ｉという生徒がはじめの３題の項目に正答（１）し，残りの（ｎ－３）題に誤答（０）した場合，反応パターンは，

$u_i=〔1, 1, 1, \cdots\cdots, 0, 0, 0〕$

となる。このとき，一般には正答した項目の重み（この場合１）の総和で表されたテスト得点によってこのテストで測ろうとしている能力とみなすことが多い。この場合，この生徒の能力は３点ということになる。しかし，実際には項目１は非常に易しくほぼ全員が正答できる項目であり，項目２は逆に非常に難しくほぼ全員が誤答するような項目であったとすると，同じ正答の１であってもその意味合いは非常に違ってくるはずである。そこで，項目反応理論では，単純にテスト得点である３点をその生徒の能力とは考えず，その生徒のもつ潜在的な能力を仮定することから始まる。潜在能力は，字のごとく目では見ることのできないその人がもつそのテストに対応した能力のことである。テストがいろいろな能力を測っているのであれば，その潜在能力は多次元になり，一つの能力を測っているのであれば一次元になる。テストが何次元の能力を測っているかは，因子分析などの方法を用いて調べられる。項目反応理論では，現在主として一次元の能力を仮定し，そのモデルにもとづいて理論が展開されている。

一般にテストを実施したとき，ある項目に対し，結果として正答（１）あるいは誤答（０）の２つの値しか表れてこないけれども，能力の低い者は正答する確率は低く，能力が高くなるにつれその確率は高くなりやがて１に近づくと考えることは自然である。モデルはこの考えにたって，ある項目に対し，ある潜在能力θをもった者がどの程度の確率$P(\theta)$で正答するかを関数で近似することによって表される。近似される関数としては，階段関数，正規分布関数，ロジスチック関数等があるが，取り扱いの容易さからロジスチック関数が主流を占めている。この場合，項目ｊの反応確率$P_j(\theta)$は

$$P_j(\theta)=\frac{1}{1+\exp(-Da_j(\theta-b_j))}$$

$(j=1,2,\cdots\cdots n)$

と表される。このモデルをロジスチック・モデルと呼ぶ。$P_j(\theta)$はある潜在能力θもった者の項目ｊに正答する確率を与えるものである。Ｄは1.7で，正規分布関数に近づくように選ばれている。θが小さい（$\theta\to-\infty$）と$P_j(\theta)$は０に近づき，θが大きい（$\theta\to\infty$）と$P_j(\theta)$は１に近づく。従って，$P_j(\theta)$は０から１の間の値をとる。

図１は，昭和60年度の数学Ⅰの18項目についてそれぞれロジスチック関数をあてはめた場合の曲線を示したもので，この曲線を項目特性曲線または項目反応曲線と呼ぶ。図は，横軸に潜在能力θをとり，縦軸にそれに対応する各項目の正答する確率$P_j(\theta)$をとっている。

また，ロジスチック・モデルには，２つのパラメータa_jとb_jを含んでいるので，これを２パラメータ・ロジスチック・モデルという。a_jは値が大きいと曲線の傾斜は急になり，潜在能力が低い

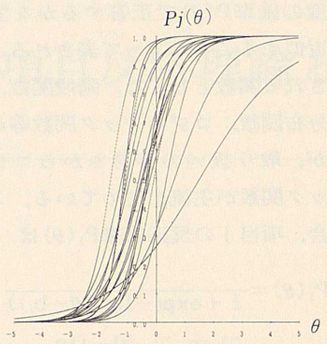

図１　項目特性曲線

者は正答する確率が低いが，ある一定の潜在能力をもった者より急に正答する確率が上がることを意味し，逆にa_jの値が小さいと曲線の傾斜はなだらかになり，潜在能力の違いによる正答確率の変化が小さいことを意味する。このことから，a_jは項目の識別力と呼ばれている。

b_jは正答確率がちょうど0.5になる潜在能力の値を示している。易しい項目ではその値は低く，難しい項目では高いことから，b_jは項目の困難度と呼ばれている。なお，多肢選択の場合，あて推量の問題がしばしば生じるが，そのような場合あて推量の確率c_jを組み入れた３パラメータ・ロジスチック・モデルもあるがここでは取り上げないことにする。

項目反応理論では，さらに同じ潜在能力θをもった者がある項目ｊとｋを正答する確率は互いに統計的に独立であるという局所独立の仮定を設けている。

さて，実際の問題としてこのモデルを使う場合，潜在能力θや項目パラメータa_j，b_jが分かっていなくてはならない。すなわち，どのようにしてこれらの値を推定するかといった問題がある。この問題を解くために，まず先の局所独立の仮定を利用してある潜在能力θをもった者の各項目に正答する確率の積を表す尤度を求め，次に，それを全ての生徒に対して求め，その尤度を最大にするようにパラメータa_j，b_jおよびθの推定量を求める方法がある。この方法は最尤推定法と呼ばれている。表１はこの方法によって推定された項目パラメータを，昭和60年度の

表１　項目パラメータ
（昭和60年度数学Ⅰ）

項目番号	問題番号	解答記号	a_j	b_j
1	Ⅰ	アイ	1.217	−1.028
2		ウエオ	1.109	−1.212
3		カキク	1.030	−0.605
4		ケコサ	1.007	−0.917
5		シス	0.911	−0.211
6	Ⅱ	アイウ	1.833	−1.091
7		エオカ	1.978	−1.141
8		キク	1.521	−1.600
9		ケコサシ	1.655	−0.356
10		スセソ	1.282	−0.075
11		タチツテト	1.253	0.365
12		アイ	0.889	−0.464
13		ウエ	0.866	−0.507
14		オカ	0.976	−0.329
15		キク	1.355	−0.282
16		ケコ	1.309	−0.326
17		サシ	0.392	1.537
18		スセ	0.672	0.648

表２　適合度の検定（項目番号１）

潜在能力θ 下限	潜在能力θ 上限	人数	正答率	モデル値
− 5.77	0.95	484	0.318	0.307
− 0.95	0.54	483	0.731	0.693
− 0.54	0.32	484	0.829	0.800
− 0.32	0.15	484	0.847	0.858
− 0.15	0.01	484	0.843	0.893
− 0.01	0.10	483	0.907	0.915
0.10	0.22	484	0.926	0.932
0.22	0.43	484	0.917	0.949
0.43	1.11	483	0.992	0.974
1.11	4.87	484	1.000	0.996

$\chi^2 = 3.814 < \chi^2_{0.05}(10)$

数学Ⅰの各項目について示したものである。表２は，問題番号Ⅰのアイの項目（項目番号１）についてロジスチック・モデルと実際のデータとの適合度を調べた結果を示したものである。この場合，かなり一致していることが分かる。

次に下式のようなｎ個の項目特性曲線の総和$T(\theta)$を考えてみる。

$$T(\theta) = \sum_{j=1}^{n} P_j(\theta)$$

この曲線はそのテストの（合計）得点の期待値をθの関数として表したものでテスト特性曲線と呼ばれている。この曲線は，項目の識別力や難易度の違いによって形状が変化し，曲線の傾斜が急であれば，潜在能力の差に比べ得点の差は大きな差となって表れ，逆に傾斜がなだらかであれば，潜在能力に差があっても，得点上には小さな差しか現れない。図２は上記の数学Ⅰのテスト特性曲線について描いたものである。図中，下側に示した分布はθの分布を表しており，左側の分布はθが正規分布すると仮定したときのテスト特性曲線による得点分布を示したものである。

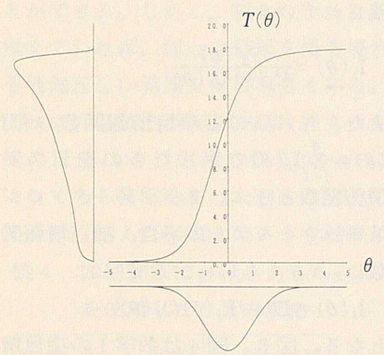

図２　テスト特性曲線

選抜試験のようなテストを実施した場合，生徒の能力を弁別するための情報ができるだけ多く得られることが望ましい。ある項目が易しすぎてどのような生徒にも正答が得られるのであれば，その項目は選抜するというテストの目的から言えばまったく意味のないものとなる。逆に全ての生徒が誤答するのであれば，同じように意味のない項目となってしまう。これに対し，ある能力をもった生徒は正答し，そうでない生徒は誤答するような項目からは，生徒の能力を識別するための情報を得ることができる。そこで，項目がどの程度の識別力をもっているかを測る道具として，次の項目情報関数が定

義されている。

$$I_j(\theta)=\frac{P_j'(\theta)}{P_j(\theta)Q_j(\theta)}$$

また，これらの各項目情報関数の和$I(\theta)=\sum_{j=1}^{n} I_j(\theta)$で表したものをテスト情報関数と呼ぶ。2パラメータ・ロジスチック・モデルの場合，項目情報関数は，

$$I_j(\theta)=D^2a_j^2P_j(\theta)Q_j(\theta)$$

となる。図3，図4は数学Ⅰの項目情報関数とテスト情報関数をそれぞれ示したものである。図3をみると，潜在能力が－2から0の範囲に曲線がピークを示している項目が多いことがわかる。また，図4からテストは潜在能力が－2から0の範囲の者について情報量が多いことを物語っている。すなわち，数学Ⅰからなるテストを考えたとき，この範囲の潜在能力をもった者に関して識別力が高いテストであるといえる。

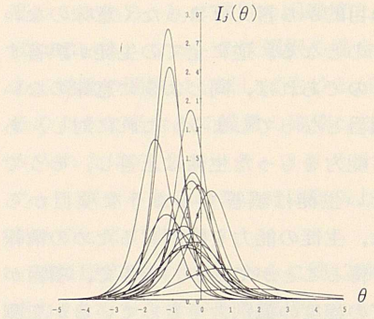

図3　項目情報関数

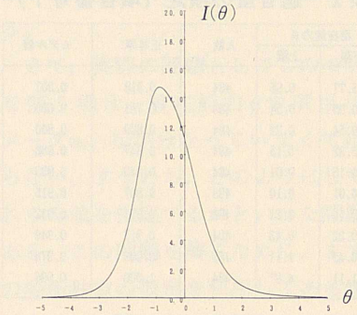

図4　テスト情報関数

うことができる。

一般に，テスト作成者の経験から，例えば易しい項目には小さい重みをつけ，難しい項目には大きい重みをつけることはしばしば行われている。そのとき，重みのとり方の1つとして，最大のテスト情報関数を与える重みのとり方がある。これは，重みw_jを潜在能力θの関数として，

$$w_j=\frac{P_j'(\theta)}{P_j(\theta)Q_j(\theta)}$$

によって表す方法である。ここで，$P_j'(\theta)$はθに関して微分したものである。2パラメータ・ロジスチック・モデルの場合，θに関係なく，項目の識別力のみに依存し，最適な重みは

$$w_j=Da_j$$

と表される。表3は，先の数学Ⅰの各項目の最適な重みと実際の配点とを対比して示したものである。この場合，最適な重みは，合計200点になるよう換算してある。易しい項目は，比較的項目の識別力は小さく，最適な重みも小さくなる傾向にある。

表3　最適な重みと実際の配点

項目番号	最適の重み	配　点
1	6.87	6
2	6.26	6
3	5.82	6
4	5.69	6
5	5.14	12
6	10.35	12
7	11.17	
8	8.59	6
9	9.34	6
10	7.24	6
11	7.07	10
12	5.02	
13	4.89	12
14	5.51	
15	7.65	8
16	7.39	8
17	2.21	8
18	3.79	8

3　項目反応理論にもとづいたテスト結果の診断

今，正負の符号のついた整数の加算や減算の問題，例えば，5＋3，5－3，5－（－3）等の演算について考えてみよう。我々はこの演算の仕方あるいはルールについて小・中学校で学んだわけであるから，もし，正しいルールを用いていれば正しい結果を得ることができる。しかし，誤ったルールを覚えていれば，誤った結果を得る場合も偶然正しい結果を得る場合もある。従って，加算や減算の問題に対する生徒の反応はその生徒の使うルールによってある特徴をもったものとなるであろう。例えば，表4のような結果を得ることができる。

表4　四則演算におけるあるルールを使用したときの反応

No	演　算	ルール0	ルール1	ルール2
1	5－3＝	＋2（1）	＋2（1）	＋2（1）
2	5－（－3）＝	＋8（1）	＋2（0）	＋2（0）
3	－5－（－7）＝	＋2（1）	－2（0）	＋2（1）
4	－5－（－3）＝	－2（1）	－2（1）	＋2（0）
5	－5－3＝	－8（1）	－2（0）	＋2（0）
	得　点	（5）	（2）	（2）

ルール0：正しいルールを使用。
ルール1：マイナスの符号が1つでもあるときは，絶対値の大きい方から小さい方を引き，大きい方の符号をつける。
ルール2：ルール1の法則でいつもプラスの符号をつける。

表4でわかるように，ルール1を用いた場合は2個の正答を，ルール2を用いた場合も2個の正答を得ることができる。ここでは，これらの反応パターンから，生徒が一体どのようなルールを用いて解答しているかを項目反応理論との関係から診断する方法について考えてみる。

一般に，易しい項目に正答し，難しくなるにつれ正答する項目が減ってくるような反応は，素直な反応と考えられるが，逆に，易しい項目に誤答するとか，難しい項目に正答するような反応は何か誤ったルールあるいは誤認知

によって生じたものと考えられる。もし，そのような反応の違いを一つの指標で表し，診断することができれば便利である。この考え方から導出された一つの指標として注意指標ξがある。

この指標は，易しい項目に対しては正答する確率が高く，難しい項目になるほど正答する確率が低くなるような曲線を規準としたとき，ある潜在能力θをもった者の実際の反応とその曲線との隔たりの程度を数値で示したものである。もし，易しい項目で誤答する数が多く，難しい項目で正答する数が多いような反応パターンであれば，注意指標の値は大きくなる傾向を示す。明らかにこのような反応パターンは異常である。注意指標はこれに対する注意信号を与える。

さて，ある１つの反応パターンに対して１つの注意指標が得られるわけであるが，同じように異なる反応パターンに対してもそれに対応する指標が得られる。先の四則演算の例でみると，合計点が２点になる反応パターンは10通りあるので，この場合の注意指標は10個である。同じ合計点であっても異なった注意指標が得られることになるので，それぞれの反応パターンについて区別することができる。

ところで，この考え方をルールに適用した場合について考えてみる。今，１つのルールがあり，そのルールによって固有の反応パターンが作り出されるとする。そうすれば，このルールからそれに対応した１つの注意指標を得ることができる。従って，ｎ個のルールがそれぞれ固有の反応パターンをもっていれば，ｎ個の注意指標が得られることになる。

従って，ｎ個のルールの注意指標が予め得られているとすれば，生徒の反応パターンから得られた注意指標を比較することにより，その生徒がどのようなルールを用いて解答したを診断することが可能となる。その診断の基礎となる注意指標と潜在能力との関係を２次元の座標系に表したものをルール・スペース（Tatsuoka, 1983, 1986, Tatsuoka & Tatsuoka, 1985）と呼ぶ。

潜在能力θをもった者が一定して特定のルールを使っていれば，１つの注意指標ξが得られ，ルール・スペース上のそのルールの点（ξ，θ）に位置づけられる。予め全てのルールをそのルール・スペース上に布置し，生徒の反応から，どの生徒がどのルールを用いて解答したかを調べる。

しかし，実際には，１つのルールだけではなく，いろいろなルールを組み合わせて解答したり，誤認知によって解答したりするので，ルール・スペース上のある特定のルールの点に合致するとは限らない。そのため，一般には統計的な分類の問題として取り扱われ，判別関数やベイズ決定理論等によってある反応がどのルールに属するかを決定する。

上に述べた考え方は，単に加算や減算といった単純な四則演算の誤認知の診断だけに限定されるわけではない。例えば，あるテストを実施したとき，医学系進学者は，ある特定の反応パターンをとり，芸術系進学者はそれとは異なった反応パターンをとるような場合，そのテストにより適性にあった進路に関する診断を行ったり，２組の生徒にそれぞれ異なった方法で学習指を行い，２組に同じテストを実施し，その結果から学習指導方法を評価する指導法の研究にも適用できる。これらの例のようにテスト使用者側の関心や目的に応じかなり広い範囲での応用が可能であろう。

〔参考文献〕

Lord, F.M.: Applications of item response theory to practical testing problems. Educational Erlbaum Associates, 1980.

Tatsuoka, K.K.: Rule space: An approach for dealing with misconceptions based on item response theory. Journal of Educational Measurement, 1983, 20, 4, 345-354.

Tatsuoka, K.K.: Diagnosing cognitive errors: Statistical pattern classification based on item response theory. Behaviormetrika, 1986, 19, 73-86.

Tatsuoka, K.K. & Tatsuoka, M.M.: Bug distribution and pattern classification (Research Report 85-3-ONR). Urbana IL: University of Illinois, Computer-Based Education Research Laboratory, 1985.