多肢選択式テストによる測定の精緻化の試み

東京大学教育情報科学研究室　平　井　洋　子

An Attempt to Improve the Precision of Measurements by a Multiple-Choice Test

Yoko HIRAI

In attempt to improve the measurement precision by a multiple-choice test, two IRT multi-category models were applied to an English test and compared to a binary model. The multi-category models were graded response model and nominal response model. Each item of the test consisted of 5 alternatives, which were classified according to their correctness into 3 categories : "correct," "nearly correct," and "far from correct." It was found that the two multi-category models showed the almost equal effect on improving the measurement precision compared to 2-parameter logistic model. The improvement was most remarkable for subjects with low ability.

目　　次

Ⅰ．問　　題

大学や高校の入学試験をはじめ，多くの能力テストでは採点手続きの容易さから多肢選択式のテスト項目が採用されることが多い。正答とされる選択肢の数は複数の場合もあるが，1つだけの場合が一般的である。このとき，被験者が選んだ選択肢が正しければ得点として1点を与え，正しくなければ0点を与える。正答得点が項目によって異なる場合もあるが，この1点―0点という得点化に対して何らかの重みを加えた得点と考えられるので，基本形としては1点―0点の場合を考えれば良い。このような採点法は一般に2値採点と呼ばれる。

この採点法の問題点は，被験者が誤った選択肢を選んだ場合（誤答のとき）を一律に扱うことである。正しい選択肢がどれかわからない被験者でも，選択肢のうち少なくとも一部は誤っていることは理解しているかも知れない。そのような被験者と，どの選択肢も同じように見えてしまう被験者とでは，持っている能力のレベルが異なると考えられる。選択肢の内容の面から考えても，1つの項目の中に「誤りの度合い」が異なる選択肢が混在していることが多い。このようなとき，どのような誤答選択肢を選んだかということと被験者の能力レベルとのあいだには，何らかの関係があると想定できる。そして，その場合は，2値採点ではなく，完全ではないが部分的な知識に対しても部分点を与えるような多段階の採点法（以下，多値採点と呼ぶ）のほうが好ましい。1つの項目から取り込む情報も多く，被験者の能力をより正確に

評価できるからである。

被験者がどの誤答選択肢を選んだかということから情報を引き出そうとする試みは，実は古くから行われている。Nedelsky(1954)やCoombs, Milholland and Womer (1956) らは，能力の低い層の被験者の中にも，正答がどれかわからなくても，誤答選択肢のいくつかを見分けて却下することができる被験者とできない被験者がいることを指摘した。その後，1960年代から1970年代にかけて，各選択肢にそれぞれ異なる重みをつけて多値採点し，テストの測定精度を向上させようとする試みが古典的テスト理論の立場から数多くなされた (Davis and Fifer, 1959, Sabers and White, 1969, Hendrikson, 1971, Reilly and Jackson, 1973)。測定精度の指標には，主に信頼性係数や妥当性係数が用いられた。これらの研究はその重みづけの方法によって大きく2つに分類できる。1つは経験的重みづけを行った研究で，もう1つは専門家による評定を用いた研究である。

経験的重みづけ法は，粗点データから各選択肢にかける重みを統計的に計算する方法である。重みの計算方法としては，ガットマンの重みづけや相互平均法，双列相関係数などが多く用いられた。これらの重みづけの結果，従来の正答数による採点に比べ，信頼性係数が高くなったという報告も多いが，結果は一貫していない。妥当性係数に関しては，むしろ低くなったという結果が多い。経験的重みづけ法の長所は，重みを導き出す手続きが客観的で明解であるという点である。短所としては，得られた重みが計算に用いられた被験者集団に依存し，一般性に欠けるという点が指摘できる。また，真に測りたい内容以外の要因を得点に取り込み，テストの焦点をずらしてしまう恐れもある。

専門家による評定を重みとする方法では，複数の専門家が各選択肢の正しさの度合いを評定し，その結果に基づいて重みを決める (Hambleton, Roberts and Traub, 1970, Patnaik and Traub, 1973)。この方法は，重みがデータに依存せず，真に測りたい内容に選択的に重みを与えることができる反面，重みが主観的になりやすく，重みの大きさが持つ意味が曖昧になりやすいという短所がある。

全体としてみると，古典的テスト理論の立場からの研究では，結局どのような重みづけが適切なのかが明らかになっていない。さらに，部分的知識を採点に反映できたかどうかの指標として，信頼性係数や妥当性係数を用いることにも限界がある。信頼性係数や妥当性係数は全被験者層をまとめた指標であるが，部分的な知識を持つ被験者は能力が中・低レベルに集まっているからである。選択肢に重みをつけることで，実際に能力の低い被験者層での測定精度が向上したとしても，そのことが信頼性係数や妥当性係数にどう反映されるかは必ずしも明確でない。重みづけの目的からすれば，いろいろな能力レベルにおいて，測定の精度がそれぞれどの程度向上したかがわかる指標のほうが適切である。また，信頼性係数や妥当性係数が計算に用いられた被験者集団に依存する，という点も指標として用いるには好ましくない性質である。

ところで，1970年代ごろから項目反応理論に基づいた採点法が発達し，古典的テスト理論とはまったく異なる立場から，被験者の能力推定が行えるようになった。項目反応理論では，被験者がテスト項目にどう反応するかを，被験者の潜在的な特性（たとえば能力）で説明する。項目反応モデルにはいろいろな種類があるが，そのなかの多値採点モデルを用いると，被験者が誤答選択肢に反応した場合にも，反応した選択肢に応じて情報を引き出すことができる。さらにこれを2値採点モデルと比較することによって，被験者の能力推定の精度がどの程度向上したかを能力レベル別にみることができる。

本研究は，多肢選択式テストを多値採点することで被験者の能力測定の精度を上げることを試みるものである。この目的にてらすと，古典的テスト理論の枠組みよりは項目反応理論の枠組みのほうが，多値採点の場合のモデルが存在している点と，被験者の能力レベル別に能力推定の精度をみることができる点の2点において好ましい。そこで，本研究では，項目反応理論における多値採点モデルを採用し，2値モデルをあてはめた場合に比べてどの程度能力推定が改善されるかを検討していく。

II. 方　法

A. データ

1. テスト

本研究で用いたテストは基礎的な英語の能力を測るテストである。内容は主に語彙力，文法力，読解力を測る項目からなり，形式は小問が25問，長文が3題で15問，計40問構成ですべて5肢選択式である。また制限時間は30分である。被験者は18才から36才まで合わせて2788名，うち男性が59.4%を占める。学歴構成でみると，4年制大学生が84.3%，短期大学生が11.4%，その他4.4%となっている。

のちにモデルのあてはまりの交差妥当性を確認するため，2788名のデータから層化抽出法により無作為に398名を抜き出し，「あてはまり確認用データ」とした。残り2390

名のデータはモデルのパラメタ推定用とした(表1)。このような人数に分割したのは，5カテゴリ25項目で段階反応モデル（Samejima, 1969）を適用した場合，プログラムパッケージMULTILOGで安定したパラメタの推定を行うには2000名程度のデータが必要という報告があり（Reise and Yu, 1990），一方，モデルのあてはまりの確認には数百人規模のデータがあれば目的を果たすことができると思われたからである。

表1　正答数得点の要約統計量

	「推定用」	「確認用」
人　数	2390人	398人
最高点	4点	5点
最低点	40点	39点
平　均	24.86点	24.92点
標準偏差	7.68	7.61

2．項目選択

本研究で採用する項目反応モデルは，潜在変数の1次元性を仮定している。すなわち，被験者の各項目への反応は，その被験者の潜在的な能力だけによって規定される。たとえば，被験者の得点が問題を解くスピードにも左右されているとしたら，能力のほかにスピード因子という潜在変数も仮定しなければならない。これは1次元性を満たさない。

ところで本研究で用いたテストは，制限時間30分で40項目に答えるテストである。このテストでは，時間が不足して最後の項目まで到達できなかった被験者もいると思われる。このままでは1次元性の仮定が満たされない。そこで，大半の被験者がその項目までは到達したと思われる項目までを分析に使用し，それ以降の項目は切り捨てることにした。具体的には被験者の95%が到達したことを目安にした結果，第34問以降を切り捨てた。

次に，残った33項目について主因子分析を行った。共通性の推定値にはSMCを用いた。33項目のうち，第1因子負荷が0.4未満の項目を除外した結果，8項目が除かれ，25項目が残った。この25項目についてあらためて主因子分析を行ったところ，第1固有値が第2固有値の9.8倍の大きさだった。また，どの項目も第1因子負荷が一様に高かった(表2)。このことから，この25項目には共通して1つの能力を測る項目が集められたといえる。

表2　主因子法による因子負荷（第2因子まで）

ITEM	第1因子	第2因子	ITEM	第1因子	第2因子
1	0.5342	−0.1606	14	0.6426	−0.1079
2	0.6030	−0.1548	15	0.5728	−0.2389
3	0.5775	−0.1328	16	0.5398	0.1711
4	0.4795	−0.1702	17	0.4642	−0.1082
5	0.4826	0.2013	18	0.6076	−0.2222
6	0.6037	0.0085	19	0.4716	0.0928
7	0.5151	−0.2036	20	0.4045	0.1481
8	0.5247	−0.0372	21	0.5833	0.1160
9	0.4520	−0.1243	22	0.4788	0.3554
10	0.5920	−0.0693	23	0.7315	0.3721
11	0.5591	0.1337	24	0.4106	0.0520
12	0.6669	−0.0103	25	0.4974	0.2521
13	0.5941	−0.1223			

3．多値採点

項目選択の結果残った25項目について，多値採点を行った。誤答選択肢に部分点を与える方法は，専門家による評定を用いた（Hambleton et. al., 1970)。

多値採点では，各項目につき5つある選択肢をその「正しさの度合い」から3つのカテゴリに分類した。このカテゴリ分けは，カテゴリ3が正答，カテゴリ2が「惜しい」誤答，カテゴリ1が「見当外れな」誤答，という3値採点に相当する。正しさの度合いの評定は，テスト項目の作成者，アメリカからの帰国子女，日本の英語教育で育った(英語の成績の良い)大学院生，の各1名が行った。カテゴリ3は正答の選択肢1つだけとし，カテゴリ2とカテゴリ1は，この3者による正しさのランクづけを単純合計したランク合計点の差の最大のところで分けた。したがって1つのカテゴリに入る選択肢の数は1つから3つまで様々である。

B．項目反応モデル

1．段階反応モデル

前項で多値採点されたデータは，正答カテゴリ3点，次に正答に近いカテゴリ2点，正答から遠いカテゴリ1点，という順序尺度になっている。したがって項目反応モデルにはカテゴリ間に順序性を仮定する段階反応モデル（Samejima, 1969）を採用し，2値採点データに対応する2パラメタ・ロジスティックモデルと比較する。

段階反応モデルでは，あるカテゴリ以上の得点をとる確率が2値モデルで表される。このとき，あるカテゴリ

得点をとる確率は，2つの隣り合う2値モデルの差で表される。

今，ある項目において，特性値 θ を持つ被験者がカテゴリ得点 k をとる確率を $P_k(\theta)$，カテゴリ境界 k より上の得点をとる確率 $P_k^*(\theta)$ とすると，段階反応モデルでは

$$P_k(\theta)=P_{k-1}^*(\theta)-P_k^*(\theta),\quad 1\leqq k\leqq m \tag{1}$$

ただし

$$P_0^*(\theta)=1,\quad P_m^*(\theta)=0$$

と定義される。本研究では，カテゴリ境界 $P_k^*(\theta)$ に2パラメタ・ロジスティック関数を用いる。

また，段階反応モデルにおける項目 j の情報関数 $I_j(\theta)$ は以下のように定義される（Baker, 1992）。

$$I_j(\theta)=\sum_{k=1}^{m}\left\{\frac{[P'_k(\theta)]^2}{P_k(\theta)}-P''_k(\theta)\right\} \tag{2}$$

ただし，$P'_k(\theta)$ および $P''_k(\theta)$ はそれぞれ $P_k(\theta)$ の1次微分，2次微分を示す。

2．名義反応モデル

Bock (1972) の名義反応モデルは，項目への反応が名義尺度になっている場合を扱う項目反応モデルである。反応カテゴリ間に順序性を仮定しないところから，名義反応モデルは段階反応モデルの高次のモデルと捉えることができる。項目反応が名義尺度の場合，カテゴリ得点の「合計点」は意味をなさない。古典的テスト理論ではこのような場合を扱うことができないが，項目反応理論では対応するモデルが存在する。

ある項目で被験者がカテゴリ k を選ぶ確率を $P_k(\theta)$ とすれば，名義反応モデルでは

$$P_k(\theta)=\frac{\exp(a_k\theta+c_k)}{\sum_{v=1}^{m}\exp(a_v\theta+c_v)},\quad v=1,\cdots,k,\cdots,m \tag{3}$$

ただし，$\sum_{k=1}^{m}P_k(\theta)=1$，$\sum_{k=1}^{m}(a_k\theta+c_k)=0$

と定義される。

また，名義反応モデルにおける項目 j の情報関数は(2)式で定義される。

III．結　果

A．段階反応モデルのあてはめ

1．項目パラメタの推定

2パラメタ・ロジスティックモデルと段階反応モデルについて，コンピュータプログラム MULTILOG Ver. 6.0 (Thissen, 1991) を用いてパラメタの推定を行った。用いたデータは「パラメタ推定用データ」2390名である。パラメタの推定にさいしては，特に制約条件は入れなかった。

2．段階反応モデルのあてはまりの評価

段階反応モデルが実際のデータをどのくらい良く記述しているかを検討する。データにはパラメタの推定に用いたものとは別の「あてはまり確認用データ」398名を用いる。

まず前項で推定された項目パラメタに基づき，モデルから期待されるカテゴリへの反応確率を求める。ついで「あてはまり確認用データ」から実際の反応率を計算する。この2つの反応率の食い違いが大きいほどモデルは妥当性に欠けることになる。

図1，図2は2つの項目について2通りの反応率を図示したものである。図1は，能力の非常に高い層と非常に低い層とを除いて，理論値と実際の反応率とが良く合っている。能力の非常に高い層と非常に低い層には該当する被験者が少なく，実際の反応率が不安定になったため，あてはまりも悪いのであろう。一方図2は，カテゴリ2のあてはまりが悪く，実際の反応率が理論値に反して能力の低い層でも下がらない。段階反応モデルでは，中央のカテゴリは必ず山形のグラフになるが，それがあてはまっていないわけである。そこで項目パラメタを計算した「パラメタ推定用データ」に対し数量化3類を行ってカテゴリの順序性を確認した。その結果，この項目はカテゴリ1とカテゴリ2の順序が逆になっていることがわかった。なお，順序性が逆転していたのはこの項目だけであった。

もともとカテゴリの順序は，選択肢の「正しさの度合い」に基づいて決定されている。しかし数量化3類の結果からは被験者は選択肢の正しさの度合いを逆に認知していたことがうかがえる。このような場合，データを優先させてカテゴリの順序を入れ替えれば内容的な正しさの順序が崩れてしまう。むしろカテゴリの順序性を仮定しないより一般的なモデルを導入したほうが，内容もデータも活かすことができると思われる。そこで，次にカテゴリ間の順序性を仮定しない名義反応モデルを用い

た分析を行う。

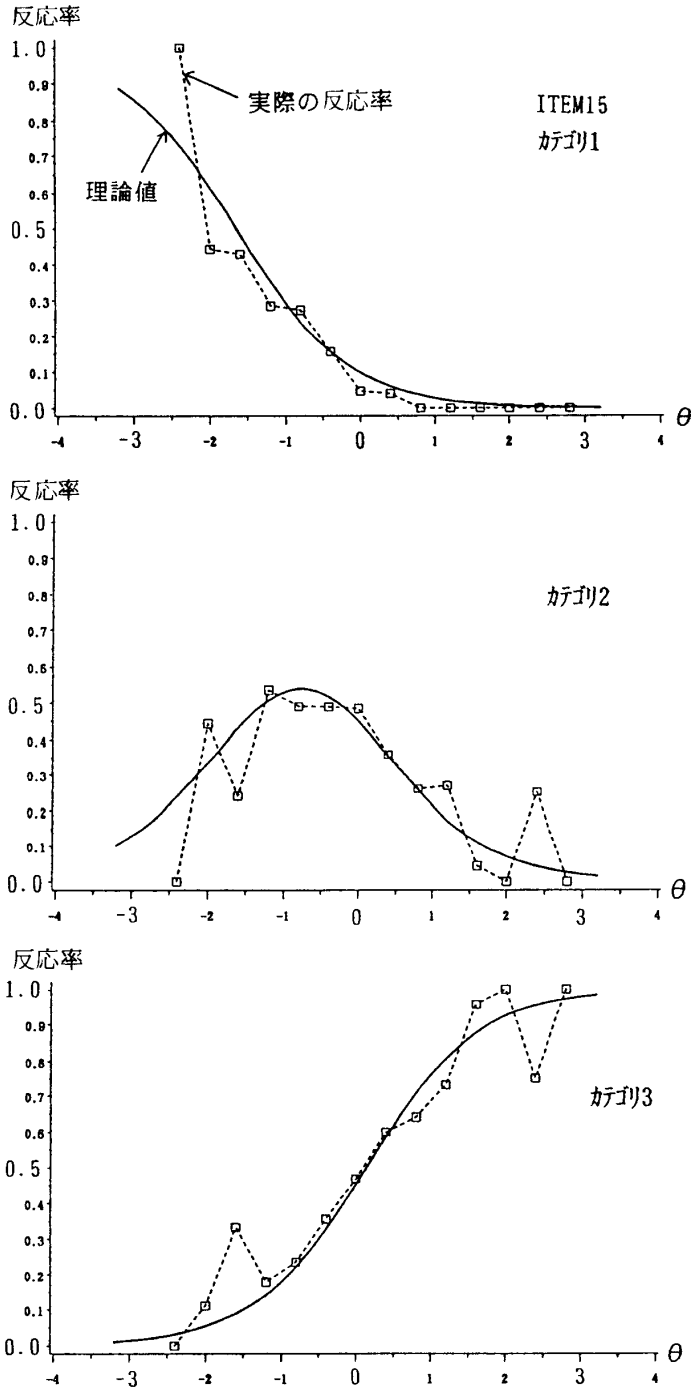

図1　段階反応モデルのあてはまり（ITEM15）

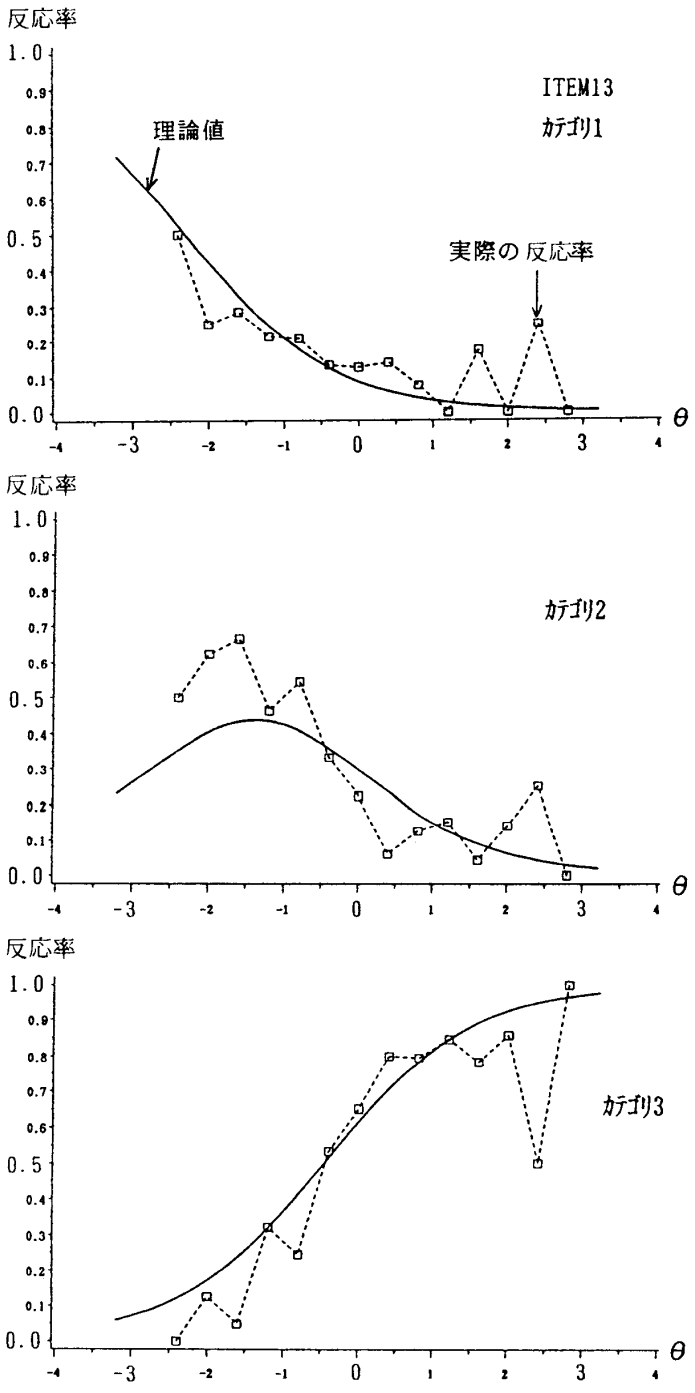

図2　段階反応モデルのあてはまり（ITEM13）

B．名義反応モデルのあてはめ

1． 項目パラメタの推定

段階反応モデルの場合と同じく，「パラメタ推定用データ」を用いて MULTILOG でパラメタの推定を行った。制約条件は特に入れなかった。推定されたパラメタに基づいてカテゴリ特性曲線を描いてみると，カテゴリ 3（正答）については，各項目とも 2 パラメタ・ロジスティックモデルや段階反応モデルとほぼ同じ曲線が得られた。このことから，正答カテゴリの働きはどのモデルの下でも安定して記述できたといえる。

2． 名義反応モデルのあてはまりの評価

名義反応モデルが実際のデータをどのくらい良く記述しているかを，「あてはまり確認用データ」398名を用いて検討する。検討の方法は段階反応モデルのときと同様で，モデルから期待されるカテゴリへの反応確率と，実際の反応率とを比較する。

図 3 は，段階反応モデルでのあてはまりが比較的良かった項目である。この項目は名義反応モデルでもあてはまりが良い。名義反応モデルでは，データが順序性を持つ場合，それを反映するように項目パラメタが推定されていることがわかる。段階反応モデルであてはまりの良い項目は，より高次のモデルである名義反応モデルでもあてはまりが良いといえるだろう。

一方図 4 は，順序性が逆転していた項目である。この項目は段階反応モデルであてはまりが悪かったが，名義反応モデルでは特にカテゴリ 2 のあてはまりが良くなっている。

25項目全体でみると，名義反応モデルのあてはまりは，段階反応モデルに比較して大きく改善されたとはいえなかった。その原因として考えられるのは，ほとんどの項目でカテゴリの順序性が保たれていたことである。順序性が保たれていなかった項目は25項目中 1 項目しかなかった。つまり，どちらのモデルをあてはめても，実質的には変わらなかったことになる。

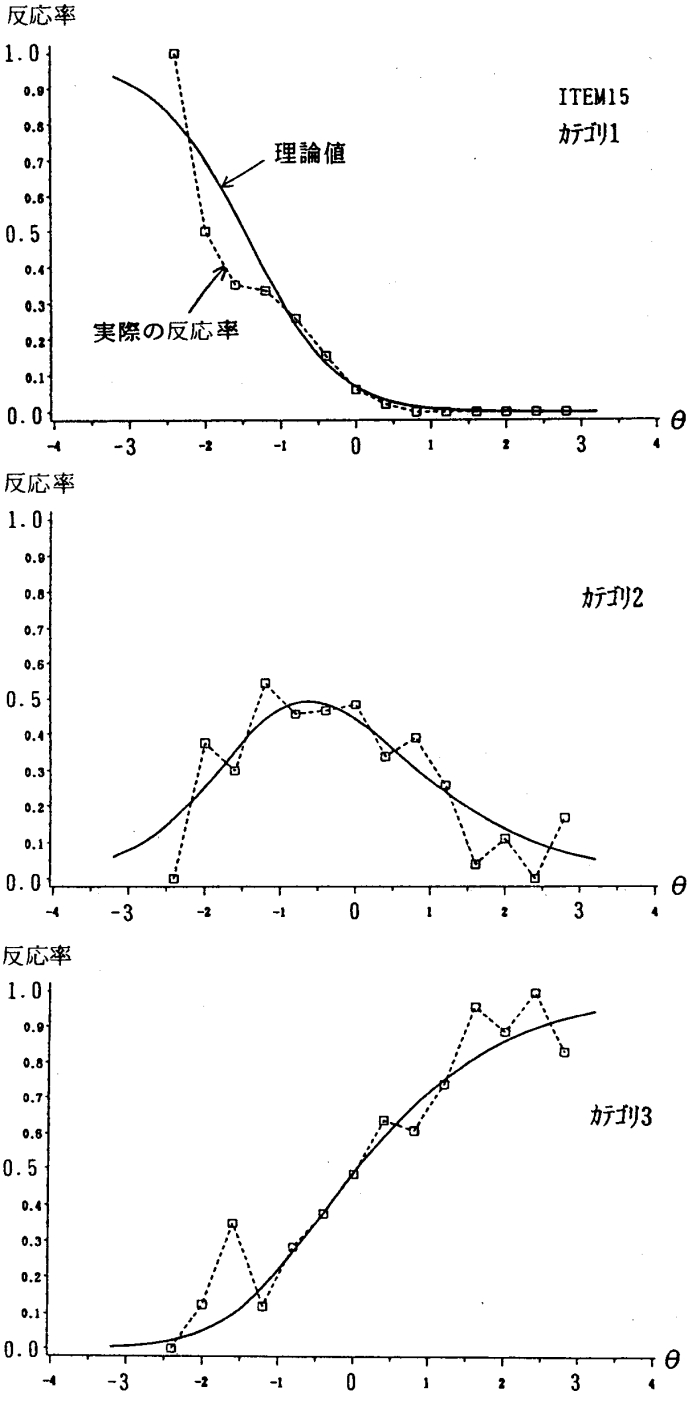

図 3　名義反応モデルのあてはまり（ITEM15）

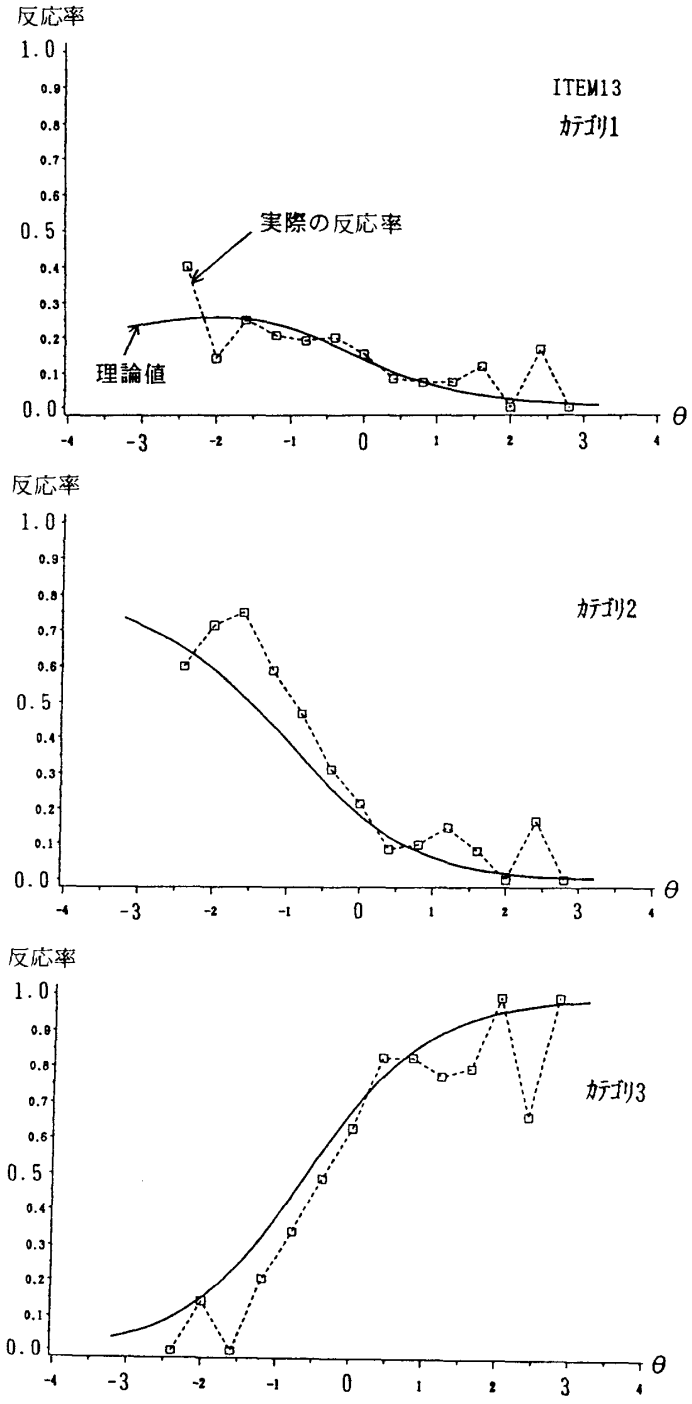

図4　名義反応モデルのあてはまり（ITEM13）

C．情報量による測定精度の比較

1．テスト全体

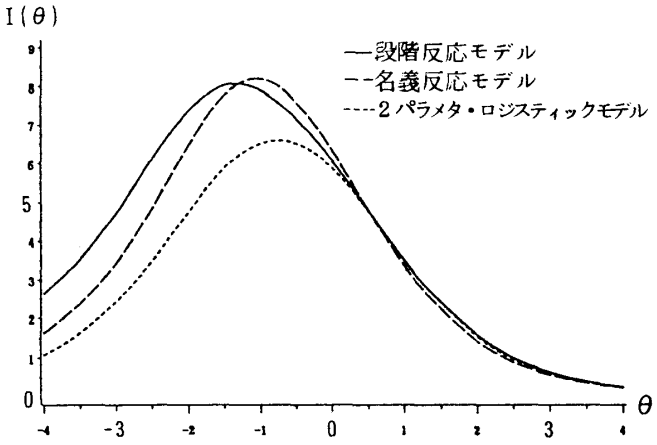

図5　3つのモデルによるテスト情報量

25項目全体で，多値採点による測定精度向上の度合いを検討する。図5は，テスト情報量を3つの項目反応モデルで比較したものである。図では，名義反応モデルの情報量のピークは2パラメタ・ロジスティックモデルより左，段階反応モデルよりやや右にある。名義反応モデルは，2パラメタ・ロジスティックモデルより，能力の低い被験者層を精度良く測定していることがわかる。また，それよりも能力の低い層を良く測定しているのが，段階反応モデルということになる。能力の高い被験者層では，3つのモデルはどれも同じような測定精度をしている。

名義反応モデルが能力の低い方の層をより良く測定するようになったことは，採点を2値から3値にするさいに，誤答カテゴリを「惜しい誤答」と「見当外れな誤答」の2つに分割したことに対応している。その3値採点の段階反応モデルとほぼ同じだった理由には，ほとんどの項目でカテゴリの順序性が保たれており，2つのモデルが実質的に同じ働きをしていたことが考えられる。

2．項目ごと

a．3つのモデルが互いに異なる情報曲線を持つ場合

図6は，2つの多値採点モデルと2パラメタ・ロジスティックモデルが互いに異なる情報曲線を持った例である。2パラメタ・ロジスティックモデルより能力の低い層を精度良く測定する形で段階反応モデルの平板な情報曲線がある。段階反応モデルでは，この項目は3値採点にした結果，幅広い能力層を同じような高い精度で測るようになったということができる。一方図6では，名義反応モデルはさらに能力の低い層を重点的に測っていることがわかる。名義反応モデルに従えば，この項目は，

正しい選択肢を選べたかどうか（2パラメタ・ロジスティックモデル）より，「見当外れな誤答」に引っかからなかったかどうかで，被験者の能力を識別していることになる。このような知見は，名義反応モデルをあてはめて初めて得られたといえよう。

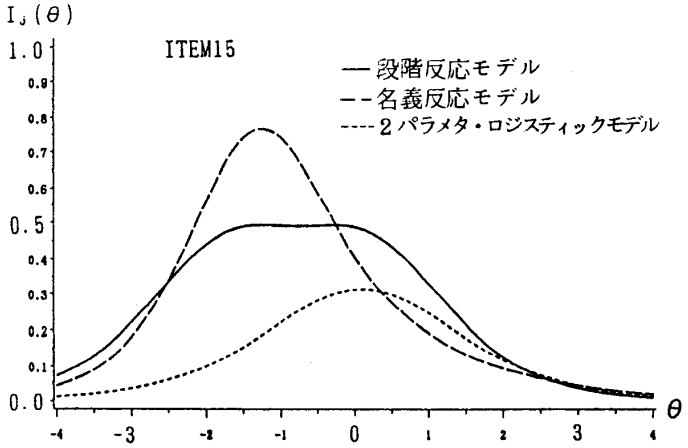

図6　3つのモデルの項目情報量（ITEM15）

b．2パラメタ・ロジスティックモデルと名義反応モデルが同じ情報曲線を持つ場合

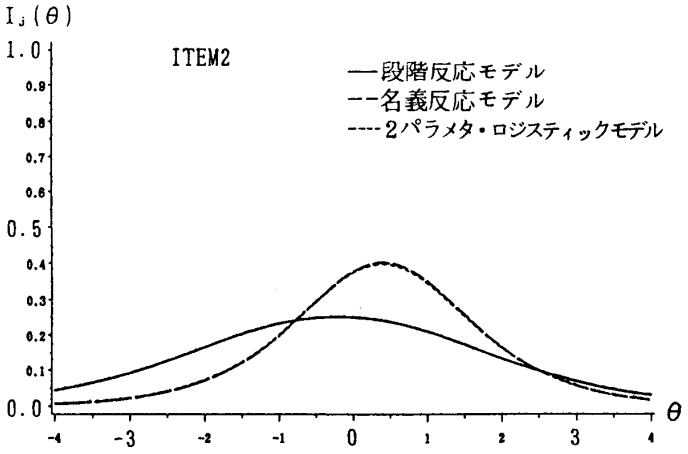

図7　3つのモデルの項目情報量（ITEM2）

図7は，2パラメタ・ロジスティックモデルと名義反応モデルがほぼ同じ情報曲線を持ち，段階反応モデルだけが異なる項目である。3つの情報曲線がこのようなパターンをとる理由は，名義反応モデルで誤答カテゴリ1，2の働きが区別されず，どちらも2パラメタ・ロジスティックモデルの「誤答」カテゴリと同じ働きをしているように記述されたことによる。つまりこの項目は，名義反応モデルによると実質的に2値採点した場合と等しいわけである。一方段階反応モデルでは，カテゴリ間に順序性の制約があるため，2つの誤答カテゴリの働きは異なるように記述される。

c．3つのモデルの情報曲線が等しい場合

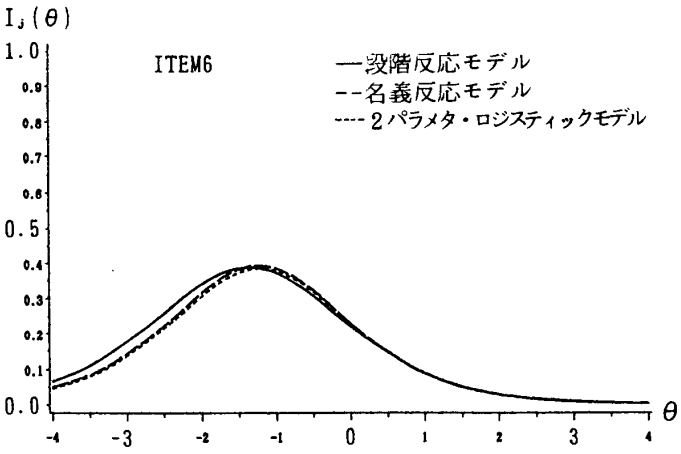

図8　3つのモデルの項目情報量（ITEM6）

図8は，3つのモデルがほぼ等しい情報曲線を持つ例である。このような項目は全体的に正答率の高い項目が多かった。誤答カテゴリを2つに分割しようとしても，「惜しい誤答」のカテゴリを選択する被験者が少なく，「惜しい誤答」カテゴリが事実上ないに等しくなったのである。もとのデータがこのような2値性を持つ以上，多値採点モデルをあてはめてもその効果は表れなくて当然といえよう。

IV．まとめと考察

能力推定精度の向上という目的から考えれば，2値採点に対する2つの多値採点モデルの効果はほとんど同じであった。まず，テスト全体の情報量では，能力の低い層で測定精度が2値採点の場合よりも向上し，その向上の度合いが2つの多値モデルでほぼ同じであった。また，モデルのあてはまりの見地からしても，2つの多値モデルの間に大きな違いは見られなかった。このことは，実用上はよりシンプルなモデルである段階反応モデルで間に合うことを示唆する。

一般に選択肢の内容から段階採点を行う場合は，カテゴリ間の順序性が必ずしも保証されないので，より制約の少ない名義反応モデルのほうが適切である。本研究でも，個々の項目レベルでは，名義反応モデルのほうが各カテゴリの働きをより自由に記述できた。ただその段階反応モデルとの記述の違いが，主として相対的頻度の少ない低能力者層で生じていたため，項目情報量やテスト情報量に大きな改善が表れなかった可能性がある。

とはいえ今後，多値採点の方法を，たとえば数量化3

類によるカテゴリへの重みを考慮に加えるなどして改善すれば，カテゴリ間の順序性はかなりの程度確保できるであろう。その場合は，よりシンプルでパラメタ推定の安定した段階反応モデルのほうが好ましい。

いずれにせよ，2値採点より3値採点のほうが，被験者の能力をより良く推定できることが確かめられた。このことは，同じ測定精度を保った場合，少なくとも能力の低い被験者層では，より少ない項目数で能力測定ができることを意味する。被験者の負担を軽減するためにも，多肢選択式テストで広く多値採点が導入されることが期待される。

（指導教官　渡部洋教授）

付　記

本論文は，著者の東京大学大学院教育学研究科修士学位論文（1993）に基づいて書かれた。

引用文献

Baker, F. B. (1992) *Item response theory : parameter estimation techniques*. New York, NY. Marcel Dekker.

Bock, R. D. (1972) Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.

Coombs, C. H., Milholland, J. E. and Womer, F. B. (1956) The assessment of partial knowledge. *Educational and Psychological Measurement, 16*, 13-37.

Davis, F. B. and Fifer, G. (1959) The effect on test reliability and validity of scoring aptitude and achievement tests with weights for every choice. *Educational and Psychological Measurement, 19*, 159-170.

Hambleton, R. K., Roberts, D. M. and Traub, R. E. (1970) A comparison of the reliability and validity of two methods for assessing partial knowledge on a multiple-choice test. *Journal of Educational Measurement, 7*, 75-82.

Hendrickson, G. H. (1971) The effect of differential option weighting on multiple-choice tests. *Journal of Educational Measurement, 8*, 291-296.

Nedelsky, L. (1954) Ability to avoid gross error as a measure of achievement. *Educational and Psychological Measurement, 14*, 459-472.

Patnaik, D. and Traub, R. E. (1973) Differential weighting by judged degree of correctness. *Journal of Educational Measurement, 10*, 281-286.

Reilly, R. R. and Jackson, R. (1973) Effects of empirical option weighting on reliability and validity of an academic aptitude test. *Journal of Educational Measurement, 10*, 185-194.

Reise, S. P. and Yu, J. (1990) Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27*, 133-144.

Sabers, D. L. and White, G. W. (1969) The effect of differential weighting of individual item responses on the predictive validity and reliability of an aptitude test. *Journal of Educational Measurement, 6*, 93-96.

Samejima, F. (1969) Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement No. 17.*

Thissen, D. M. (1991) *MULTILOG Version 6.0 user's guide*. Chicago, IL. Scientific Software, Inc.