Japanese Association of Educational Psychology

教育心理学研究, 2004, 52, 61—70

テストが複数の出題形式を含むときの項目母数の推定

荘島宏二郎* 豊田秀樹**

我が国の一般的なテストは，正誤問題（多肢選択式を含む）・テストレット・論述式問題の組み合わせであることが多い。正誤問題には2値モデル，テストレットはGRMやGPCM，論述式問題はCRMを適用することが可能である。したがって，テストが複数のIRTモデルを適用すべき項目から構成されているときでも，それぞれの項目に適切なIRTモデルを当てはめて，それぞれの項目母数を推定できれば便利である。本研究では，テストが複数のIRTモデルを含むときの項目パラメタの推定方法と被験者母数の推定方法を提案した。また，提案された方法を用いて，実データに対する分析例を示した。最後に，本方法の適用についての注意点を議論した。

キーワード：項目反応理論，3パラメタ・ロジスティックモデル，段階反応モデル，連続反応モデル，EMアルゴリズム（5項目）

目　　的

我が国の一般的なテストは，正誤問題・テストレット・論述式問題の組み合わせであることが多い。正誤問題は穴埋め式・多肢選択式の問題であり，反応を0，1に符号化することができる問題を指す。テストレットは，複数の小問題の1つのまとまりである。また，論述式問題は，比較的長い文章で被験者に回答させ，例えば0点から10点までの枠内で，採点者が妥当だと思う得点が付与される。

項目反応理論（item response theory, IRT ; Lord, 1980）は，テストデータを効果的に分析し，長期に渡って効率よく運用する上で，最も有効な理論体系となりつつある。そして，正誤問題や多肢選択式問題は2値モデルを適用できる。ただし，多肢選択式問題については，名義反応モデル（nominal response model, NRM ; Bock, 1972）も適用可能である。また，テストレットは段階反応モデル（graded response model, GRM ; Samejima, 1969），部分的採点モデル（partial credit model, PCM ; Masters, 1982），一般化PCM（generalized PCM, GPCM ; Muraki, 1992）を用いることで対応できる。また，論述式問題は，連続反応モデル（continuous response model, CRM ; Samejima, 1973）を適用することができる。

CRMは，項目得点が10点や20点など，レンジが大きいときに有効である。理論的には，例えば，項目得点が10点のとき，$(0, 1, \cdots, 10)$という11のカテゴリを考えて，GRMやGPCMなどを適用することができる。実際にPARSCALE（Muraki & Bock, 1987）では，カテゴリ数が15まで分析可能である。しかし，カテゴリ数がKのとき，GRMやGPCMは位置母数を$K-1$個推定する必要があり，推定値の安定性を考えると望ましくない。一方で，CRMはどんなに得点のレンジが大きくても項目母数は常に3つだけである。したがって，理論的に論述式問題にGRMやGPCMが適用可能であってもCRMを適用した方が良いと思われる。なお，カテゴリ数が大きいとき，順序カテゴリ得点を連続得点と見なしても大きな間違いはない（例えば，Bentler & Chou, 1987 ; 狩野・三浦, 2002）。

さて，IRTの理論的な発展は米国の研究者に負うところが大きい。したがって，IRTは，米国のテスト状況にうまく適合している。石塚（2003）が紹介しているが，米国の大規模試験は1917年のArmy Alphaに始まると言われ，その基本方針は12ほどあり，現在の米国のテスト文化に少なからず影響を与えていると思われる。我が国のテスト文化と比較して大きな違いは“Minimum of writing in making responses（最小限の筆記量）”を基本方針としていることである（カッコ内の邦訳は石塚（2003）による）。つまり，受験者に多くを書かせないのである。現在の米国のSAT（Scholastic Assessment Test）やTOEFL（Test of English as a Foreign Language）などの大規模試験も短問形式であることを考えてみると，Army Alphaの影響は大きいと思われる。

一方で，我が国のテストの実務家は，正誤問題で受験者の学力を測定できるとは考えておらず，実際にテ

* 大学入試センター　〒153-8501　目黒区駒場2-19-23
shojima@rd.dnc.ac.jp
** 早稲田大学文学部

NII-Electronic Library Service

Japanese Association of Educational Psychology

ストレットや論述式の問題は好まれている。科挙は論述式問題であった(石塚, 2003)ことから，我が国のテスト文化に科挙の影響が見てとれるかもしれない。そういう意味でも，初めてテストレットに対応できるIRTモデル(GRM)，初めて論述式問題に対応できるIRTモデル(CRM)を考案したのが日本人研究者(Samejima)であることは興味深い。

我が国でIRTを運用するには，1つのテストに複数の項目形式を含む場合にIRTを適用することである。現在，テストが様々な形式の項目が混在する場合，ソフトウェアMULTILOG 6.0 (Thissen, 1991)を用いることができる。ただし，MULTILOGでは，項目母数を推定する際の手続きを明らかにしていない。ただ，周辺最尤推定法(marginal maximum likelihood, MML)を用いているとのみ書かれている。また，MULTILOGでは，CRMの項目母数を推定することができない。

したがって，本研究では，まず，様々な形式の項目が混在するとき，EMアルゴリズムを用いた周辺最尤推定法(marginal maximum likelihood with EM algorithm, MML-EM)を用いて項目母数を推定する方法を詳述し，広く研究者間でその知見を共有することを目的とする。このとき，様々なタイプの変数が含まれているときの因子分析モデルを提案したOgasawara(1998)が参考になる。また，そこでは欠損値の取り扱いについても述べる。さらに，テストが複数の項目形式を含むときの被験者母数の推定方法として最尤推定値，事後期待値，事後モーダル値を得る方法を述べる。最後に，応用上重要であると思われ，かつ，MULTILOGでは推定することができないCRMも含めた複数のIRTモデルを適用すべき項目から構成されている大学の学期末試験において，その経年的な運用という比較的小さな分析事例を示し，本方法の有効性を確認する。また，本方法の具体的な適用に関して予想される問題事項について議論する。

方　法

項目母数の推定

多くのIRTモデルは被験者の潜在特性θと項目得点を結ぶ数学関数としての項目反応関数(item response function, IRF)が定義され，IRFは，θの尺度上に項目独自の項目パラメタに数値的な個性を持ちながら位置づけられている。

いま，あるテストがM種類のIRTモデルを含み，項目j ($=1, 2, \cdots, n$)はm番目のIRTモデルによって規定されているとする。また，項目jは，項目パラメタλ_jをもつとする。そのとき，θの潜在特性をもつ被験者が項目jに対して得点x_jをとる確率が項目jの得点x_jに対するIRFであり，ここでは

$$\Pr(X_j=x_j|\theta)=P_{xj,mj}(\theta) \tag{1}$$

と書くことにする。添え字mjは，項目jがm番目のIRTモデルに規定されていることを示す。ここで，x_jは名義得点・順序得点・連続得点の区別をしていない。

いま，N人の被験者のn個の項目に対する反応データを$X=\{x_{ij}\}(N\times n)$とする。また，被験者iの項目jに対する反応が欠損ならば0，そうでないならば1である2値変数d_{ij}を考える。さて，被験者母数$\theta=[\theta_1,\ldots,\theta_N]'$と項目母数$\Lambda=[\lambda_1,\ldots,\lambda_n]$の下で$X$が得られる確率は

$$p(X|\theta,\Lambda)=\prod_{i=1}^{N}\prod_{j=1}^{n}\prod_{m=1}^{M}P_{xj,mj}(\theta_i)^{k_{mj}\times d_{ij}} \tag{2}$$

となり，同時にθとΛに関する尤度関数である。ここで，k_{mj}は，もし項目jがm番目のIRTモデルに規定されているならば1，そうでないならば0である2値のindicatorである。k_{mj}によって，どの項目が何番目のIRTモデルに適用されているのかが適切に選択される。また，d_{ij}によって，欠損されたデータに関するIRFは尤度から除外される(Mislevy, 1986 ; 前川, 1991)。

さて，(2)式を最大にするΛを数値的に解く場合，局外母数θをθに関する密度関数で積分消去した事後確率$p(\Lambda|X)$をΛの目的関数として，この関数を最大化する$\hat{\Lambda}$を推定する方法が一般的である。この問題はEMアルゴリズム(Dempster, Laird, & Rubin, 1977)を用いた周辺最尤推定法(Bock & Aitkin, 1981 ; Mislevy, 1984, 1986)を適用することが可能である。EMアルゴリズムは，E-stepとM-stepを繰り返して最適化を行う数値解法である。

E-stepでは，局外母数である被験者母数の事後分布と，期待対数尤度を得ることを目的とする。いま，$\Lambda^{(t)}$を第t期のM-stepによって得られた項目パラメタの推定値とすると，被験者iの反応パタンx_iの下で，第$(t+1)$期の被験者iの特性値θ_iの事後分布は，ベイズの定理により

$$p(\theta_i|\Lambda^{(t)},x_i)=\frac{p(x_i|\theta_i,\Lambda^{(t)})p(\theta_i)}{\int_{\Theta i}p(x_i|\theta_i,\Lambda^{(t)})p(\theta_i)\mathrm{d}\theta_i} \tag{3}$$

である。ここで

$$p(x_i|\theta_i,\Lambda^{(t)})=\prod_{j=1}^{n}\prod_{m=1}^{M}P_{xj,mj}(\theta_i)^{k_{mj}\times d_{ij}} \tag{4}$$

であり，$p(\theta_i)$はθ_iの事前分布である。Θ_iはθ_iの母数空間である。

例えば，仮に項目数が6であり，項目1，2に2値モ

NII-Electronic Library Service

Japanese Association of Educational Psychology

デル，項目3，4にGRM，項目5，6にCRMを適用すべきだとする。また，2値モデルを1番目のモデル，GRM，CRMを，それぞれ，2，3番目のモデルとすると，そのとき，k_{mj}を要素とする3×6のダミー変数行列Kは

$$K=\{k_{mj}\}=\begin{bmatrix}1&1&0&0&0&0\\0&0&1&1&0&0\\0&0&0&0&1&1\end{bmatrix} \quad (5)$$

となる。また，2値モデルのIRFをP_{ij}，GRM，CRMのそれをG_{ij}，C_{ij}とおくと，(4)式は，k_{mj}を操作したことにより

$$p(x_i|\theta_i,\Lambda^{(t)})=\prod_{j=1}^{n}(P_{ij}^{k_{1j}}G_{ij}^{k_{2j}}C_{ij}^{k_{3j}})^{d_{ij}}$$
$$=(P_{i1}^{d_{i1}}P_{i2}^{d_{i2}}G_{i3}^{d_{i3}}G_{i4}^{d_{i4}}C_{i5}^{d_{i5}}C_{i6}^{d_{i6}}) \quad (6)$$

となる。よって，項目ごとにモデルを選択して尤度を構成していることになっている。なお，G_{ij}は，どのカテゴリが選択されたかによって，当該カテゴリの項目カテゴリ反応関数が割り当てられる。

次に，EMアルゴリズムを適用した際の目的関数は

$$\ln p(\Lambda|X)=\sum_{i=1}^{N}\int_{\Theta_i}\ln p(X|\theta,\Lambda)p(\theta_i|\Lambda^{(t)},x_i)\,\mathrm{d}\theta_i$$
$$+\ln p(\Lambda)+\mathrm{const.} \quad (7)$$

となる（Tanner, 1993；Wang & Zeng, 1998）。積分計算は，θの尺度上に適当な数のガウス・エルミート求積点，もしくは，等間隔の求積点を設けることにより，十分な近似が得られる。そして，(7)式を最大にするΛをもって$\Lambda^{(t+1)}$とする。

実際は，項目間の局所独立の仮定により

$$\ln p(\lambda_j|X)$$
$$=\sum_{i=1}^{N}\int_{\Theta_i}\sum_{m=1}^{M}k_{mj}d_{ij}\ln P_{xj,\,mj}(\theta_i)p(\theta_i|\Lambda^{(t)},x_i)\,\mathrm{d}\theta_i$$
$$+\ln p(\lambda_j)+\mathrm{const.} \quad (8)$$

をλ_jに関して，M-stepでは項目別に最大化することができる。ただし，E-stepについては，M個のIRTモデルと反応データに基づく尤度を全て用いて，被験者母数の暫定的な事後分布と期待対数尤度を更新していく必要がある。

なお，(8)式における$\sum_m k_{mj}d_{ij}\ln P_{xj,\,mj}(\theta_i)$は，先述の例で言うと，例えば項目1については

$$\sum_m k_{m1}d_{i1}\ln P_{x1,\,m1}(\theta_i)$$
$$=d_{i1}(k_{11}\ln P_{i1}+k_{21}\ln G_{i3}+k_{31}\ln C_{i4})$$
$$=d_{i1}\ln P_{i1}:=d_{i1}\ln P_1(\theta_i|\lambda_1) \quad (9)$$

となり，2値モデルが選択されることになる。また，θ軸上にQ個の求積点（$Y_1,\cdots Y_Q$）を取ると，項目1の目的関数は

$$\ln p(\lambda_1|X)=\sum_{i=1}^{N}\sum_{q=1}^{Q}\{d_{i1}\ln P_1(Y_q|\lambda_1)\}p(Y_q|\Lambda^{(t)},x_i)$$
$$+\ln p(\lambda_1)+\mathrm{const.} \quad (10)$$

になる。例で挙げた項目1について説明したが，他の項目についても同様である。(10)式中，$p(Y_q|\Lambda^{(t)},x_i)$は重みの定数であるので，同式の中カッコ内で適用すべきモデルのみが残り，また，項目ごとに独立に最適化することになる。

各IRTモデルにおける最適化計算は，例えばNRMはBock (1972)，GRMはSamejima (1997, Pp.94-97)，GPCMはMuraki (1992)，CRMはWang & Zeng (1998)を，それぞれ参照してほしい。なお，CRMの項目母数はEMサイクル内で非反復推定することが可能である（Shojima, 2003）。

被験者母数の推定

次に，項目パラメタが既に得られているとき，項目によって規定されているIRTモデルが異なるという条件の下で，被験者母数を推定するには，以下の尤度

$$p(x_i|\Lambda,\theta_i)=\prod_{j=1}^{n}\prod_{m=1}^{M}P_{xj,\,mj}(\theta_i)^{k_{mj}\times d_{ij}} \quad (11)$$

の対数をθ_iについて最大化すればθ_iの最尤推定値が得られる。ここでθ_iに関する事前分布$p(\theta_i)$を用いるならばEAP（expected a posteriori）スコアは

$$\theta_{EAP}=\int_{\Theta_i}\frac{\theta_i p(x_i|\Lambda,\theta_i)p(\theta_i)}{\int_{\Theta_i}p(x_i|\Lambda,\theta_i)p(\theta_i)\mathrm{d}\theta_i}\mathrm{d}\theta_i \quad (12)$$

のように算出できる。また，MAP（modal a posteriori）スコアは

$$\frac{\partial\ln p(x_i|\Lambda,\theta_i)}{\partial\theta_i}+\frac{\partial\ln p(\theta_i)}{\partial\theta_i}=0 \quad (13)$$

をθ_iについて解けばよい。

分　析

テスト

調査法・初等統計学・実験計画法の通年の講義を行った直後の学年末定期試験を都内の大学生164名に2002年1月に実施した。このテストは13項目から構成され（荘島, 2003），項目1から項目10は5肢選択式の正誤問題であり，正答は1，誤答は0とした。項目11は3つの小問題からなるテストレットであり，得点は0，1，2，3を与えたが，カテゴリ3に反応した被験者が少なかったのでカテゴリ3とカテゴリ2を合わせてカテゴリ2とした。すなわち，最終的に得点は0，1，2を割り振った。項目12は2つの小問題からなるテストレットであり，0，1，2の得点を与えた。項目13は論述式問題であり，0－30点を割り振った。このデータ

NII-Electronic Library Service

Japanese Association of Educational Psychology

を「本データ」と呼ぶ。なお，項目13の164名全員の回答を8人全ての採点者が点数を付け，それらを項目13-20とする。大学の定期試験の論述試験の採点に8人の採点者を配すことは珍しい。ここでは，後の分析で4人ずつに分け，項目1-12の母数の推定値の安定性を調べる目的で採点者を8人用意した。付録に8人の採点者の採点基準のプロトコルを付した。

項目反応モデル

項目1から項目10は2値モデル，項目11と項目12はカテゴリ3のGRMを適用した。問題の性質上，項目11，12には当て推量パラメタは仮定せず，識別力パラメタと2つの困難度パラメタ b_{j1}，b_{j2} $(j=11,12)$ を仮定した。項目13から項目20はCRMを適用した。ここで，各種のIRTモデルのうち，比較的なじみの薄いと思われるCRMの紹介を行う。正規密度型のCRMは，得点のレンジが0から k_j である項目 j において，潜在特性値が θ である被験者が得点 x_j をとる確率（密度）を

$$f(X_j=x_j|\theta)$$
$$=\frac{a_j}{\sqrt{2\pi}a_j}\exp\left\{-\frac{1}{2}a_j^2\left(\theta-b_j-\frac{1}{a_j}\ln\frac{x_j}{k_j-x_j}\right)^2\right\} \quad (14)$$

と表現するモデルである(Samejima, 1973)。ここで a_j は識別力パラメタ，b_j は困難度パラメタ，$a_j(>0)$ は x_j が変化するにつれてどのていど困難度に影響を与えるかを表現したパラメタである。a_j が小さいほど x_j の変化が θ の変化に影響する。なお，項目13-20において，(14)式と $k_j=30$ より，$x_j=0$ もしくは $x_j=30$ を許さないので，$x_{ij}=0$ だった被験者 i には0.01点とし，$x_{hj}=30$ だった被験者 h は29.99点とした。

分析手続き

4通りの分析を行った。

分析1 項目1-16を用いた。項目1-10の正誤問題には，2パラメタ・ロジスティック（2PL）モデルを適用した。また，項目パラメタの事前分布には，16項目の全ての識別力パラメタの対数に平均0，分散0.5の正規分布を課し $(\ln a\sim N(0, 0.5))$，全ての困難度パラメタの事前分布を $N(0, 2)$ を課した。なお，項目11と12には，2つの困難度パラメタがあるが，いずれの困難度パラメタにも $N(0, 2)$ の事前分布を与えた。これらの事前分布はIRTの専用ソフトウェアBILOG 3 (Mislevy & Bock, 1990) のデフォルトである。また，CRMの α パラメタには，$\alpha>0$ により，識別力パラメタと同じ事前分布を課した $(\ln \alpha\sim N(0, 0.5))$。また，被験者母数の事前分布には標準正規分布を用いた。分析1では，主として項目内容に照らした際の母数の推定値の内容的な妥当性を考察することを目的とする。

分析2 項目1-10の正誤問題に，3PLモデルを適用した。項目1から項目10は5肢選択問題であるので，当て推量を0.2ほど見込めるからである。一方，$N=164$ は3PLモデルの母数を推定するには少ない。分析2では，その損失を比較・考察する。項目パラメタの事前分布は，分析1と同様にし，当て推量パラメタの事前分布には $Beta(5, 17)$ を課した。

分析3 一般的に項目母数の推定をするために標本数 $N=164$ では少ない。分析3では，分析1の2値の項目の推定値を安定させるために，項目1-10の事前分布として，TABLE 1において該当する項目の推定値を平均，標準誤差を標準偏差にもつ正規分布を仮定した[1]。

分析4 論述式項目は，項目1-12の客観式テストとは異なり，採点者の個性が反映される。客観式項目(項目1-12)の母数の推定値もその影響を受ける。そこで，分析1とは別の採点者4名に入れ替えて項目母数の推定値の安定性を検証する。すなわち，項目1-12，17-20を用いた。その他の状態は分析1と同じである。

結果・考察

分析1-3の項目パラメタの推定値と標準誤差を順にTABLE 2-4に示す。分析4については，図的な考察を行う。

項目によってまちまちであるが，おおむね推定の精

TABLE 1 正誤問題の事前分布のための2PLモデルの項目母数の推定値と標準誤差

項目	a	$S.E.$	b	$S.E.$
1	0.329	0.099	-1.668	0.520
2	0.357	0.096	-0.722	0.292
3	0.554	0.110	0.591	0.188
4	0.476	0.125	-1.605	0.375
5	0.584	0.123	-0.943	0.205
6	0.618	0.119	0.370	0.161
7	0.376	0.102	0.193	0.222
8	0.723	0.146	-0.543	0.138
9	0.417	0.104	1.832	0.457
10	0.358	0.094	0.491	0.259

[1] 1996年(6年前)に同一内容を同じ教師が通年の授業を行った直後の全く別の大学生226名に対して実施した学年末定期試験のデータ。このデータを「事前データ」と呼ぶ。項目数は50であり，すべて5肢選択形式である。本データ中の10個の5肢選択項目は「事前データ」の項目からそのまま選ばれており，事前データに含まれる項目の推定値の情報を利用できるようにあらかじめ計画されて「本データ」の試験は実施された。TABLE 1の推定値と標準誤差は事前分布を仮定しないもとでBILOGによって得たものである。

NII-Electronic Library Service

Japanese Association of Educational Psychology

TABLE 2　分析1の項目パラメタの推定値と標準誤差

項目	a	S.E.	b	S.E.		
1	0.312	0.004	-1.838	0.194		
2	0.240	0.004	0.029	0.150		
3	0.389	0.008	0.705	0.089		
4	0.374	0.006	-0.870	0.085		
5	0.261	0.004	0.014	0.128		
6	0.532	0.014	0.901	0.069		
7	0.381	0.008	0.762	0.096		
8	0.245	0.003	-1.111	0.181		
9	0.429	0.011	1.841	0.223		
10	0.198	0.002	-0.926	0.235		
項目	a	S.E.	b_1	S.E.	b_2	S.E.
11	0.480	0.007	-0.749	0.054	2.184	0.199
12	0.846	0.022	0.475	0.019	4.097	0.752
項目	a	S.E.	b	S.E.	α	S.E.
13	2.577	0.026	-0.274	0.001	2.780	0.006
14	2.295	0.022	-0.158	0.001	2.826	0.008
15	1.844	0.016	-0.192	0.002	2.732	0.011
16	2.581	0.026	-0.058	0.001	2.596	0.005

TABLE 3　分析2の項目パラメタの推定値と標準誤差

項目	a	S.E.	b	S.E.	c	S.E.
1	0.437	0.027	-0.518	0.427	0.289	0.014
2	0.661	0.067	1.412	0.277	0.339	0.006
3	1.099	0.146	1.043	0.062	0.231	0.003
4	0.957	0.085	0.206	0.047	0.302	0.005
5	1.161	0.196	1.039	0.076	0.351	0.004
6	0.929	0.066	1.081	0.069	0.147	0.003
7	0.756	0.052	1.172	0.122	0.200	0.004
8	0.702	0.070	0.775	0.175	0.393	0.007
9	0.914	0.136	1.982	0.230	0.153	0.002
10	0.631	0.120	1.768	0.732	0.453	0.010
項目	a	S.E.	b_1	S.E.	b_2	S.E.
11	0.477	0.007	-0.748	0.054	2.201	0.201
12	0.838	0.021	0.482	0.020	4.124	0.755
項目	a	S.E.	b	S.E.	α	S.E.
13	2.559	0.026	-0.271	0.001	2.770	0.006
14	2.252	0.021	-0.155	0.001	2.810	0.008
15	1.846	0.016	-0.189	0.002	2.726	0.011
16	2.578	0.026	-0.055	0.001	2.588	0.005

度が高い。おそらく，CRMを組み込んだことによる影響であると推察される。というのは，CRMの項目情報関数はθによらず一定であり (Samejima, 1973)，常に$I_j(\theta)=a_j^2$の高い値を保つ。また，TABLE 2-4におけるCRMの識別力パラメタの値はかなり高い値で推定されている。これは，E-stepにおいて，被験者パラメタの事後分布の位置が精度よく定まることを物語り，その結果，M-stepでの項目パラメタの推定精度を高めているのだと推察される。以下に，個々の分析に関して考察を行う。

TABLE 4　分析3の項目パラメタの推定値と標準誤差

項目	a	S.E.	b	S.E.		
1	0.271	0.003	-1.964	0.129		
2	0.172	0.004	-0.535	0.073		
3	0.421	0.006	0.617	0.023		
4	0.302	0.005	-1.283	0.089		
5	0.204	0.005	-0.775	0.042		
6	0.592	0.007	0.577	0.014		
7	0.356	0.005	0.444	0.030		
8	0.265	0.006	-0.614	0.017		
9	0.380	0.004	2.016	0.110		
10	0.078	0.004	0.373	0.078		
項目	a	S.E.	b_1	S.E.	b_2	S.E.
11	0.478	0.007	-0.755	0.054	2.188	0.201
12	0.839	0.021	0.474	0.020	4.111	0.756
項目	a	S.E.	b	S.E.	α	S.E.
13	2.594	0.026	-0.278	0.001	2.782	0.006
14	2.300	0.022	-0.162	0.001	2.826	0.008
15	1.840	0.016	-0.196	0.002	2.731	0.011
16	2.597	0.026	-0.062	0.001	2.597	0.005

分析1の考察

TABLE 2より，2値の項目1-10のうち，最も易しい問題は項目1であり($b_1=-1.838$)，最も難しい問題は項目9であった($b_9=1.841$)。また，これら10項目の識別力は，いずれも中程度以下の値であった。また，テストレットの項目11, 12において，今回の被験者たちにとって，カテゴリ2以上に正答するのは，$b_{11,2}=2.184$，$b_{12,2}=4.097$より，極めて難しいと言える。

項目13-16の項目パラメタの違いから，採点者の採点に関する個性が読み取れる。項目13は採点者1の割り振った項目得点から項目パラメタを推定したものであるが，甘目に採点したというプロトコル通り，若干ではあるが，困難度が低めに推定された。識別力に関しては，4人のうち項目15(採点者3)の識別力がわずかに低い。採点者3のプロトコルを見てもいい加減に採点したとは思えないが，相対的に他の3人の採点者よりは潜在特性を見抜く眼力に欠けると言える。また，αパラメタは4人とも，2.5以上であった。これは$\ln x_j/(k_j-x_j)$が+1だけ変化すると，困難度が1/2.5ほど増えることを示し，4人の採点者の高得点/低得点の割り振り方に関する感覚がほぼ等しいことを表している。

推定された項目母数の値は項目内容に照らして不自然なものではなく，使用したプログラムが適切に機能

NII-Electronic Library Service

Japanese Association of Educational Psychology

していたことが確認された。

分析2の考察

主に，分析1との比較を中心に考察する。分析2は，2値項目1-10に3PLモデルを適用したというものであった。項目母数を1つ多く仮定することによって，項目母数の推定値の安定性がやや失われた。すなわち，TABLE 2に比べてTABLE 3の識別力と困難度の標準誤差が大きくなっている。また，当て推量を仮定したことによって，TABLE 2よりも識別力が高めに推定された。BILOGなどの専用ソフトウェアでも経験的に確認されることであるが，IRFに下方漸近線を仮定することによって，その分，傾きが急勾配になるからである。その際に時として項目の位置母数である困難度が劇的に変化する項目も現れる。ここでは，例えば，TABLE 3の項目10の困難度パラメタは1.768であるが，2PLモデルのもとでは-0.926であった(TABLE 2)。ただし，標準誤差が0.732と大きいので推定値の値はそれほど信用できないであろう。

分析3の考察

主に，分析1との比較を中心に考察すると，分析3の項目1-10の項目パラメタの標準誤差が明らかに小さい。これは，項目1-10の事前分布に事前データの分析結果の情報を組み込んだためである。分析1の項目パラメタの事前分布は，BILOGのデフォルトであり，一般に事前分布の分散が大きいという意味で制約が緩い。それに比べて，分析3で与えた事前分布は，すでに同規模データから母数が推定されている項目の事後分布であり，分散が小さいという意味で制約が厳しいからである。その結果，TABLE 4の項目パラメタの推定値はTABLE 1で与えられた推定値に引きずられる方向で安定的に推定されている。以上のことから，事後分布を事前分布として用いる本方法は，比較的少ない被験者に対しても経年的な運用をする際に効果的であることが示されたといえる。

分析4の考察

分析4では，4人の採点者を差し替えた。FIGURE 1は，分析4と分析1で得られた項目1-12の識別力パラメタ a_j $(j=1,\cdots,12)$ の散布図である。ほぼ，一直線上にのっており，採点者を差し替えても識別力パラメタの値が安定しているといえる。また，FIGURE 2は，分析4と分析1で得られた項目1-12の困難度パラメタ $b_j(j=1,\cdots,10), b_{j1}$，$b_{j2}(j=11,12)$ の散布図である。FIGURE 2は，FIGURE 1よりも更に一直線に並んでいる。採点者を替えたことによる母数の推定値のバラツキは小さく，その意味で項目母数の推定値の安定性が高いと言うことができる。

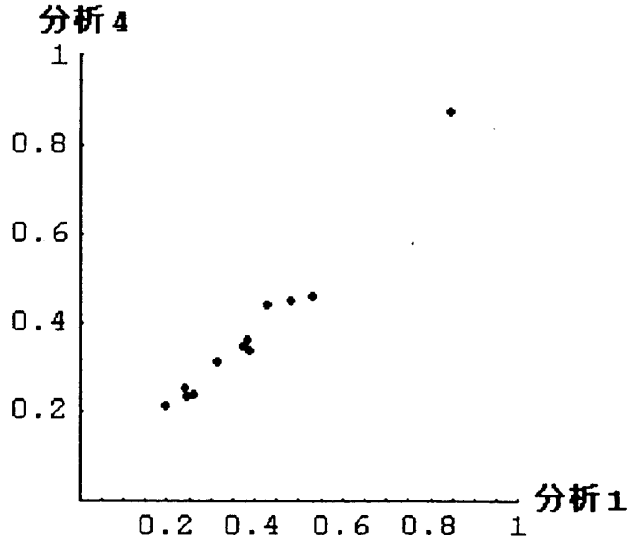

FIGURE 1 分析1と分析4の項目1-12における識別力パラメタの散布図

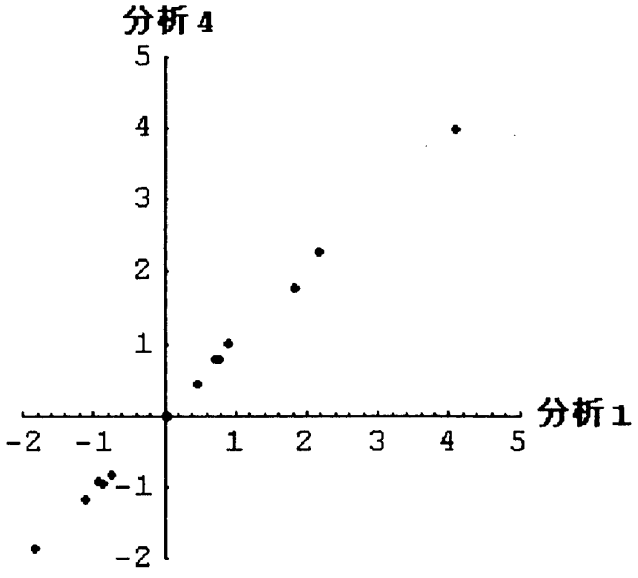

FIGURE 2 分析1と分析4の項目1-12における困難度パラメタの散布図

討　論

本方法の適用に関して，留意すべき点が幾つかあると思われる[2]。それらは，

1．異なる項目形式は，異なる潜在特性を測定している可能性

2．用いられる項目反応モデルの比較可能性

であろう。

1．について，英語テストを例に挙げれば，正誤問題

NII-Electronic Library Service

Japanese Association of Educational Psychology

は文法問題や発音問題などで多用されるかもしれない。また，論述式問題やテストレットは，多少なりとも統語論的な問題で用いられるだろう。したがって，それらの問題は，英語能力の下位能力である文法能力や語彙能力，長文読解能力など，様々な特性のどれを重点的に測定しているかで異なっている可能性がある。

しかし，正誤問題であっても長文読解問題で用いられ，かつ，文脈の中でしか正解を特定できないのであれば，それは，長文読解能力を少なからず測定しているし，論述式問題であっても問題作成者は，どれだけの語彙を駆使しているかに注目しているかもしれない。つまり，測定している特性は，項目形式にのみ依存しているのではなく，むしろ，その多くは項目内容に依存している。したがって，内容的妥当性を吟味することが必須であると考える。

そして，項目内容によっては，多次元モデルを用いることも考えられる。荘島 (2003) は，構造方程式モデリング (structural equation modeling, SEM ; Jöreskog & Sörbom, 1993 ; 豊田, 1998, 2000, 2003) のソフトウェアを用いて同様な試みを行っている。しかし，多次元モデルを用いるということは，1人の被験者に対して，複数の潜在特性値 θ が推定されるということである。複数の θ の出力は，テストの実務家を困惑させるという意味でも，多次元モデルは現実的にまだ応用段階に至っていない。実際の大学入試でも，個別大学・個別学部の受け入れ方針 (admission policy) にしたがって，「英語・数学・国語」の得点を「2：2：1」などのような重み付き和得点を算出し，合否を決める場合が多い。本研究で1次元モデルに特化して議論を進めたのは，現況の試験制度や日常のテストを鑑みたときに現実的だったからである。

2.について，本研究の実データ解析例で用いられた2PLモデル，GRM，CRMは相互に比較可能である。つまり，それらはGRMの特別な場合として説明することができる (Samejima, 1973) からである。なぜ，本研究でこれらのモデルを取り上げたのかといえば，それら3つのモデルの相性がよく，推定された項目母数についても比較する上で齟齬が少ないと判断したからである。GRMの代わりにGPCMを用いたならば，同一の潜在特性上に位置づけられているので用いることは可能であっても，推定された項目母数同士を単純に比較することができない。そういう意味では，どのようなモデルの組み合わせであっても常に好ましいとは言えないだろう。しかしながら，項目形式によっては，NRMやGPCMが相応しいことも考えられ，相性がいいからという理由で多値型の項目にはGRMを適用することが常に相応しいわけではないことは明らかである。そのとき，項目形式によっては，NRMやGPCMを適用する必要があるが，そのときの有用性は本研究の分析例では示されておらず，今後の研究課題となろう。

なお，項目母数の推定値は，「被験者母数の推定値」と「情報量」を知る上で必須である。項目内容に照らし合わせて項目母数を比較する作業も重要であるが，基本的に項目母数の役割は，上記の2点である。本研究では，各IRTモデルが同一の潜在特性上に布置しているので，被験者母数を推定する上では，異なるIRTモデル同士の項目母数が比較可能でなくても問題がない。また，θ を固定したときの情報量の大きさは，異なるIRTモデル間で常に比較可能である。しかし，これらの点については今後の議論が待たれることでもあり，本研究のみで結論を急ぐべきではないだろう。

最後に，我が国の一般的なテストは，正誤問題，テストレット，論述式問題の組み合わせであることが多く，そのようなテスト文化の中では2値の正誤問題だけでは潜在特性を測定することが難しいと考えるテストの実務家は多い。そのときこそ，本研究で述べられた方法は広くIRTユーザに薦めることができる。本研究により，IRTが1名の教員が行う学年末程度の小規模テストの実施にも対応できることが示された。

また，豊田 (2002) のように，IRTは心理テストにも応用されてきている。本研究によって，小規模テストにIRTが適用可能であることが示されたことは，比較的小規模の心理質問紙データにも対応できることを示唆している。1つの調査の中では，複数の心理質問紙でバッテリを組むことも多く，その場合，質問紙の間で項目形式が必ずしも統一されていない。そのような場合にも本研究の方法は有効に働くはずである。また，討論で述べられたことについては，引き続き議論をしていく必要がある。

[2] この点は，査読者の方々に有益なご指摘をいただきました。この場をお借りして感謝いたします。ありがとうございました。

引用文献

Bentler, P. M., & Chou, C. -P. 1987 Practical issues in structural modeling. *Sociological Methods and Research*, **16**, 78-117.

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

NII-Electronic Library Service

Japanese Association of Educational Psychology

37, 29-51.

Bock, R. D., & Aitkin, M. 1981 Marginal maximum likelihood estimation of item parameters : Application of an EM algorithm. *Psychometrika*, **46**, 443-459.

Dempster, A. P., Laird, N. M., & Rubin, D. B. 1977 Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 1-38.

石塚智一 2003 我が国における伝統的試験と標準化された試験 国立大学入学者選抜研究連絡協議会第24回大会セミナー資料『我が国における公的試験の標準化について』1-3.

Jöreskog, K. G., & Sörbom, D. 1993 *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Hillsdale, NJ : Lawrence Erlbaum Associates.

狩野 裕・三浦麻子 2002 AMOS, EQS, CALISによるグラフィカル多変量解析―一目で見る共分散構造分析― 現代数学社

Lord, F. M. 1980 *Applications of item response theory to practical testing problems*. Hillsdale NJ : Erlbaum.

Masters, G. N. 1982 A Rasch model for partial credit scoring. *Psychometrika*, **47**, 149-174.

前川眞一 1991 パラメタの推定 芝 祐順(編) 項目反応理論 東京大学出版会 Pp.87-130.

Mislevy, R. J. 1984 Estimating latent distributions. *Psychometrika*, **49**, 359-381.

Mislevy, R. J. 1986 Bayes modal estimation in item response models. *Psychometrika*, **51**, 177-195.

Mislevy, R. J., & Bock, R. D. 1990 *BILOG 3: Item analysis and test scoring with binary logistic models*. Chicago, IL : Scientific Software, Inc.

Muraki, E. 1992 A generalized partial credit model : Application of an EM algorithm. *Applied Psychological Measurement*, **16**, 159-176.

Muraki, E., & Bock, R. D. 1987 *PARSCALE: Analysis of graded responses and ratings*. Chicago, IL : Scientific Software, Inc.

Ogasawara, H. 1998 A factor analysis model for a mixture of various types of variables. *Behaviormetrika*, **25**, 1-12.

Samejima, F. 1969 Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph*, No.17.

Samejima, F. 1973 Homogeneous case of the continuous response model. *Psychometrika*, **38**, 203-219.

Samejima, F. 1997 Graded Response Model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern itam resporse theory*. New York : Spring-Verlag. Pp.85-100.

荘島宏二郎 2003 複数の項目反応モデルの母数の同時推定 豊田秀樹(編著) 共分散構造分析[技術編] ―構造方程式モデリング― 朝倉書店 Pp. 222-233.

Shojima, K. 2003 A noniterative item parameter solution in EM cycles of the continuous response model. Manuscript submitted for publication.

Tanner, M. A. 1993 *Tools for statistical inferrence : Methods for the exploration of posterior distributions and likelihood functions*. New York : Springer-Verlag.

Thissen, D. 1991 *MULTILOG User's Guide-Version 6*. Chicago, IL : Scientific Software.

豊田秀樹 1998 共分散構造分析[入門編] ―構造方程式モデリング― 朝倉書店

豊田秀樹 2000 共分散構造分析[応用編] ―構造方程式モデリング― 朝倉書店

豊田秀樹(編著) 2002 項目反応理論[事例編] ―新しい心理テストの構成法― 朝倉書店

豊田秀樹(編著) 2003 共分散構造分析[技術編]―構造方程式モデリング― 朝倉書店

Wang, T., & Zeng, L. 1998 Item parameter estimation for a continuous response model using an EM algorithm. *Applied Psychological Measurement*, **22**, 333-344.

謝 辞

本研究を作成するにあたり，飯塚久哲さん(UCC上島珈琲株式会社)，米村大介さん(株式会社インテージ)，および早稲田大学文学研究科の尾崎幸謙さん，室橋弘人さん，神笠泰宏さん，齋藤朗宏さん，中村健太郎さんに協力していただきました。この場をお借りして感謝いたします。 (2003.2.6 受稿, '04.1.24 受理)

NII-Electronic Library Service

Japanese Association of Educational Psychology

付　録

採点者1のプロトコル（項目13）

1原則あたり10点を与え，名称が正しく言える程度に応じて0から5点を与え，内容が正しく述べられた程度に応じて0から5点を与えた。おおむね，甘めに採点した。

採点者2のプロトコル（項目14）

1原則につき10点で，名前ができていたら2点，具体例で3点，内容で5点を与えた。「具体例」はその原則の例として正しければ3点を与えた。「内容」はその原則の定義だけの場合は最低点で1点を与え，その原則を守ることによって，どうして実験の精度が上がるのかについて，他の統計用語等も用いて回答していれば最大で5点とした。

採点者3のプロトコル（項目15）

3原則の名前，各5点（計15点）。正確でなくても意味の通じるものはその程度に応じて1-2点減点した。3原則の説明：「無作為」「反復」は4点，「局所管理」は7点（計15点）とした。説明が例のみのものは説明として不十分として2点減点，漢字の間違い，筆が滑ったものなど推測できる範囲の間違いは1つにつき1点減点。その他適時，説明不足と感じるものは1-2点減点した。

採点者4のプロトコル（項目16）

3原則それぞれに，各10点，そのうち5点を名前に，5点を説明に割り振った。名前は，正しく言えていれば満点，それに類する言葉，あるいは説明については減点法で採点。説明は，それぞれ説明3点，例2点とし，説明で点がつかないものは，例は見なかった。説明は，反復では誤差の分散について，無作為化では系統誤差の混入について，局所管理ではブロック因子の意味について触れているもののみ採点対象とした。あとは，名前と同様に減点法で採点。

採点者5のプロトコル（項目17）

各原則ごとに10点を割り当てた。内訳は，名前に3点，内容に7点である。名前は多少間違っている場合には減点した。内容は理論で4点，例で3点とした。理論では，原則ごとにポイントを1つ作り，そこに2点を与え，後の2点は印象で決めた。例は理論が書けていた者は3点となることが多かったが，理論が間違っていても書いてあれば2点を与えた。

採点者6のプロトコル（項目18）

3原則の1つごとに10点を割り振った。内訳は，原則の名前が4点，実際に何をすべきかの説明が2点，何を目的としているかの説明が2点，具体例が2点だった。覚えている言葉をつないでみたような答案に対しては，採点を行った後に減点した。逆に，理解できているのは分かるが答案が採点のポイントから外れている人には加点した。

採点者7のプロトコル（項目19）

各原則ごとに「原則の名称が正しく書けているか」と「原則の内容が正しく書けているか」「原則を説明する例は適切か」の3点に注目した。各原則に対する配点は10点。そのうち，原則の名称が正確に書けていれば4点。原則の内容に対して4点，例に対して2点を与えた。ただし，採点は非常に甘めで，若干の差異には目を瞑った。

採点者8のプロトコル（項目20）

0-30点のうち，各原則に10点ずつ配分した。さらにその10点を，原則の名称に3点，内容に7点配分した。

NII-Electronic Library Service

Japanese Association of Educational Psychology

Item parameter estimation when a test contains different item response models

Kojiro Shojima (National Center for University Entrance Examinations) and Hideki Toyoda (Waseda University)

Japanese Journal of Educational Psychology, 2004, 52, 61—70

In Japan, tests typically consist of true-false questions, testlets, and essay questions. A dichotomous response model is most suitable for true-false questions ; a graded response model (GRM) or a generalized partial credit model (GPCM), for testlets ; and a continuous response model (CRM), for essay questions. When tests contain more than one type of question, these differences make it necessary to use the applicable item response model in order to estimate item parameters. In the present research, we propose an estimation method for such tests, and use actual data to show how it is a applied. The discussion describes practical applications of this method.

Key Words : item response theory, 3-parameter logistic model, graded response model, continuous response model, EM algorithm.

NII-Electronic Library Service