CATにおける項目選択ルールの履歴

荘島宏二郎，橋本貴充，大津起夫，石塚智一（大学入試センター）

本研究では，コンピュータ適応型テスト（CAT）において，項目反応理論（IRT）に基づく3つの項目選択ルールの項目選択履歴についてシミュレーション研究を行った．3つの方法は，ベイジアンアプローチ，多段階a層別化法，重み付き情報量最大化法であった．項目提示回数が20前後では，重み付き情報量最大化法がもっとも潜在特性の測定精度と効率が良かった．

1. 導入

近年，コンピュータを用いたテストが普及しつつあり，computer based testing (CBT) と言われ注目されている．CBTをめぐる国内や米国の事情は，池田(2003)，池田・林(2004)，廣瀬（2000 2004）などに詳しい．

CBTの1つのカテゴリとして，コンピュータ適応型テスト (computer adaptive testing, CAT ; van der Linden & Glas, 2002; Wainer, 2000）がある．CBTといった場合には，単にコンピュータを用いたテストという意味が強いが，CATといった場合には，被験者個々人によって実施されるテストが異なる．CATでは，被験者個々人によって仕立て着のようにテストが異なるという意味で Tailored Testing とも言われる（村木・萩原，2002）．我が国におけるCATの開発としては，永岡・植野（1992），菊地（2003a 2003b）などがある．

被験者個々人によって実施されるテストが異なるというのは，被験者によって提示される項目群が異なるということである．被験者それぞれに能力が異なるので，それぞれの能力に見合った項目を提示する仕組みを提供するのがCATである．

一般にCATにおいて，回答パタンの履歴から次に提示されるべき項目が被験者の前に表示されるまでに，2つのプロセスを経る（Sympson & Hetter, 1985）．それらは，(1)項目選択プロセス（item selection process）と(2)項目実施プロセス(item administration process）である．基本的に両プロセスのそれぞれで検閲を満たした項目が次に被験者に提示されるべき項目として採用される．

まず，(1）の項目選択プロセスでは，被験者の回答パタンから最適な項目を項目バンクから選抜するプロセスである．つまり，被験者志向のプロセスである．本研究の興味の対象であるので，後ほど詳細な説明を行う．

次に，(2）の項目実施プロセスであるが，このプロセスは項目バンク志向である．というのは，項目選択プロセスでは，ある基準にしたがって被験者に最適な項目を選択するので，その基準のもとで選択されやすい項目と選択されにくい項目がでてくる．すると，選択されやすい一部の項目の項目内容が被験者に広く知られてしまう危険性がある(Mills & Stoking, 1996)．これは項目内容のセキュリティ問題と密接にかかわる．また，一部の項目が選択される率が高いと項目バンクにある膨大な項目を十分に生かせないという問題点が発生する．したがって，項目実施プロセスでは，仮に項目選択プロセスにおいて検閲を満たした項目であっても露出率（exposure rate）が高い項目に検閲をかけて露出率の低い項目とのバランスをとる．Sympson & Hetter（1985）では，選択されやすい項目は実施されにくくなるように，最終的な露出率をコントロールしている．

本研究の関心は，(1）の項目選択プロセス

における項目選択ルールについてである．また，項目反応理論（item response theory, IRT）の枠組みの中での項目選択ルールに限定する．ところで，IRTに基づく項目選択ルールは，さまざまに提案されておりvan der Linden & Glas (2002) が参考になるが，いったいユーザはどのルールを採用したらよいのかが明らかでない．項目提示終了後の各ルールの特徴に関する記述はvan der Linden（1998）やChang & Ying（1999）でも見られるが，履歴に関する特徴が分かっていない．

2．目的

本研究は，項目選択ルールのシミュレーション研究を通じて，特に項目提示の履歴に焦点化してシミュレーション研究を行うことにより，CATの実務家に判断材料となる資料を供することである．本節では，議論に必要な準備として，関連するIRTの理論的なガイダンスと項目選択ルールに関する先行研究についてレビューを行う．

IRTでは，潜在特性がθである受験者が，項目反応が2値（正答／誤答）の項目j（=1, 2,･･･, n）に対する確率的な反応をU_jとしたとき，当該項目に正答（U_j=1）する確率を

$$\Pr(U_j = 1 \mid \theta) = P_j(\theta) \qquad (1)$$

のように何らかのθの関数として表現する．これを項目反応関数(item response function, IRF）という．IRFは，一般的にロジスティック分布関数を用いた

$$P_j(\theta) = c_j + \frac{1 - c_j}{1 + \exp\{-1.7a_j(\theta - b_j)\}} \qquad (2)$$

がよく用いられ，3パラメタ・ロジスティック（3PL）モデルという．ここでa_jは傾き母数，b_jは位置母数，c_jは下方漸近母数である．上の式において，c_jを考えないとき，2パラメタ・ロジスティック（2PL）モデルといわれ，我が国では最も頻繁に用いられるモデルである．

さて，IRTに基づく項目選択ルールをざっとレビューすると，4つの方法が利用できると思われる．それらは

(a) 最大情報量アプローチ（maximum information approach）
(b) ベイジアン・アプローチ（Bayesian approach ; Owen, 1975）
(c) 多段階a層別化法（multistage a-stratified method ; Chang & Ying 1999）
(d) 最大期待事後重み付き情報量法（maximum expected posterior weighted information; van der Linden, 1998）

である．

これらのうちまず（a）の最大情報量アプローチは，被験者の回答の履歴から，各回で被験者の能力値θの最尤推定値を推定し，その能力値に対して最大の情報量をもつ項目を次の項目として選択するというものである．

IRTにおける情報量とは，θのレベルごとに項目の推定精度を定量的に評価できるものであり，通常のFisher情報量である．すなわち，θの最尤推定量の標準誤差の2乗の逆数と同じである．いま，テストにn個の項目があるとき，情報量は

$$I(\theta) = \sum_{j=1}^{n} I_j(\theta) = \sum_{j=1}^{n} \frac{\{\partial P_j(\theta)/d\theta\}^2}{P_j(\theta)Q_j(\theta)} \qquad (3)$$

で表現される（例えば，Hambleton & Swaminathan, 1985 ; 池田, 1994）．なお，$I(\theta)$を項目情報量，$I(\theta)$をテスト情報量という．

次に，（b）のベイジアン・アプローチは，項目が提示された各回でのθの事後分布を利用する．いま，t回目の項目提示が終了した時点での潜在特性値θ_tの事後分布を$g(\theta_t)$とすると，$g(\theta_t)$は

$$g(\theta_t)=\frac{\prod_{s=1}^{t}P_s(\theta_t)^{u_s}Q_s(\theta_t)^{1-u_s}pr(\theta_t)}{\int\prod_{v=1}^{t}P_v(\theta_t)^{u_v}Q_v(\theta_t)^{1-u_v}pr(\theta_t)d\theta_t} \quad (4)$$

で求めることができる．ここで，$pr(\theta_t)$はθ_tの事前分布，P_sはs回目に提示された項目のIRFである．また，$Q_s=1-P_s$である．さらに，u_sはs回目に提示された項目に対する反応である．

そして，t回目における事後分布から事後期待値（expected a posteriori, EAP）と事後標準偏差（posterior standard deviation, PSD）を利用する．それらはそれぞれ，

$$EAP_t=\int\theta_t g(\theta_t)d\theta_t \quad (5)$$

$$PSD_t=\sqrt{\int(\theta_t-EAP_t)^2 g(\theta_t)d\theta_t} \quad (6)$$

である．(4)，(5)，(6) 式の積分計算は，ガウス・エルミート求積点，等間隔の求積点，あるいはモンテカルロ法などで十分な数値近似を行うことができる．ベイジアン・アプローチは上記のEAPとPSDを用いて

$$|EAP_t-b_j|<0.5\times PSD_t \quad (7)$$

を満たす項目群からランダムに次に提示すべき項目として選抜する．つまり，暫定的な能力値に近い位置母数をもつ項目が選ばれる．その際，PSDの大きさがに考慮されている．しかしながら，この方法は，(a) の方法と似た結果が得られることが知られている（Chang & Ying, 1999）．

(a) と (b) の方法は，項目提示の初期のころから，傾き母数の大きい項目が提示される可能性がある．しかしながら，PSDが大きい段階，つまり，項目提示の初期から，傾き母数の高い（情報量の大きい）項目を提示しても効率の良い測定ができない（Chang & Ying, 1996）．測定における誤差分散が大きくなるからである．したがって，項目提示の初期には傾き母数の小さい項目を，項目提示が進むにつれて傾き母数の大きい項目を提示するのが良いとされる．

(c) の多段階a層別化法（Chang & Ying, 1999，村木・萩原 2002）は，まさにこの点を乗り越えるために提案された．まず，傾き母数aの大きさによって項目バンクを数段階に層別化する．その上でPSDが小さくなるにつれて，したがって，項目提示が進行するにつれて，より高次のステージから次の項目を選抜する仕組みである．なお，村木・萩原（2002）では，各ステージから項目を選抜する際には，EAPとPSDが (7) 式を満たすように選抜している．よって，村木・萩原 (2002) では，ベイジアン・アプローチも選択プロセスのうちに含まれている．

最後に，(d)の最大期待事後重み付き情報量法（以下，頭文字をとってMEPWI法と略記する）である．MEPWI法は，情報量と$g(\theta_t)$との重み付き和である，以下の式

$$\psi_j^2=\int I_j(\theta_t)g(\theta_t)d\theta_t \quad (8)$$

を未提示の項目全てに対して評価し，その中で(8)式が最大であるような項目jを次に提示される項目とする．これは，t回の項目提示の履歴を経た後の暫定的な事後分布$g(\theta_t)$に対して，情報量が最も大きい項目を選択しようという意図である．

この方法は，事後分布の期待値と同じくらいの位置母数を持つ項目が選択されやすく，また，事後分布の標準偏差が小さくなるほど（項目提示の回数が増えるほど），傾き母数が大きい項目が選択されやすい．

3．数値実験

3．1．　手続き

3方法の項目選択履歴に関するシミュレーションを行った．3つの方法は，第2節で紹介した(b)(c)(d)とした．(a)の最大情報量アプローチは，ベイジアン・アプローチと似た結果を返す（Chang & Ying, 1999）ために数値

実験から割愛した．また，全般的に村木・萩原（2002）の手続きを参考とした．

項目反応モデルは2PLモデルとし，項目バンクの項目数は3000とした．項目母数の発生はaはその自然対数を平均0，分散0.3の正規分布$N(0,0.3)$から発生させ，bは$N(0,2)$から発生させ，仮想的な項目バンクを作成した．

次に，θを$N(0,1)$から発生させ，発生させたθを（2）式に代入させた値と，一方でランダムに[0,1]の一様分布から発生させた値と比較して大きければ項目反応を1，そうでなければ0とした．

また，最初に提示される項目は，3つの方法においていずれも傾き母数が0.5，位置母数が0.0の項目とした．

最後に，項目提示終了ルールであるが，通常は，PSDないしテスト情報量がある一定の基準を満たした時点で，項目提示を終了するという方法がよく用いられる．しかし，受験者によって提示項目数が異なるのは不公平感につながるという指摘がある（村木・萩原2002）．よって，ここでも項目数が30に達したところで項目提示を終了することにした．上記の手続きを1000人分繰り返した．

なお，多段階a層別化法に関して，$a \leqq 0.8$である項目をステージ1，$0.8 < a \leqq 1.3$である項目をステージ2，$1.3 < a$である項目をステージ3とした．その結果，ステージ1の項目数は678，ステージ2は1765，ステージ3は557となった．また，PSDが1.0を下回った時点でステージ2の項目から選択されるように移行し，さらにPSDが0.5を下回った時点でステージ3に移行するようにした．なお，該当するステージにおいて候補となる項目がないときには，(7) 式を満たす項目からステージに関係なくランダムに選抜した．また，PSDの計算にはガウス・エルミート求積点を用いた．

3．2．　数値実験結果

以下から示す結果は，項目バンクにどのような性質の項目がどれだけ貯蓄されているかによる．したがって，あくまでも作成した仮想的な項目バンクに依存する結果である．ただし，3方法の結果を比較する上ではよく傾向が表れていると思われる．

まず，項目提示の履歴が進むにつれて，どのような傾き母数の値が選択されていったのかについて図1に示す．図は，1000回の繰り返しの平均である．

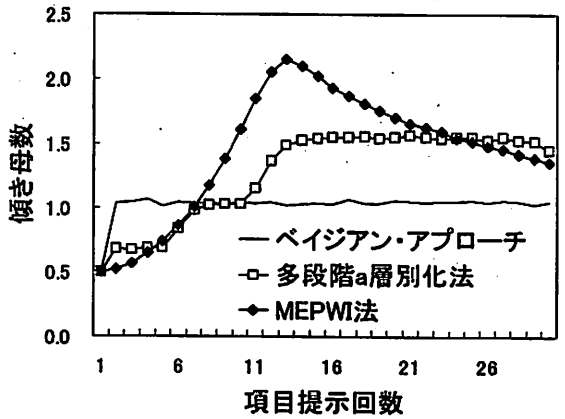

図1　傾き母数の平均値の履歴

図1より，ベイジアン・アプローチは，履歴を通じて傾き母数の平均値は一定であった．それは，ベイジアン・アプローチが（7）式さえ満たせば，傾き母数についてはランダムに選択されるからである．多段階a層別化法は，履歴が進むにつれて高位のステージから項目が選択されていることが伺える．平均的には6〜7回目の項目提示からステージ2に移行し，11〜13回目でステージ3に移行した．さらに，MEPWI法は，履歴の最初の方は，傾きが小さい項目がよく選択され，履歴が進むにつれ傾きが大きい項目が選択されていった様子が伺える．ただし，13回目でピークが見られる．これは，重み付き情報量を最大にするような傾き母数の高い項目が有限だからである．それでも，未提示項目の中では最も重み付き情報量を大きくする項目が常に選択さ

れ続けている．無論，傾き母数の大きい項目が多くあれば，図1におけるMEPWI法の曲線は，高いままに保たれる．

次に平均偏差（mean difference, MD）を考察する．MDは小さい方が推定の精度がよいとされ，

$$MD = \frac{\sum_{r=1}^{1000}(EAP_r - \theta)}{1000} \qquad (9)$$

で評価した．ここで，EAP_rは，r番目の繰り返しにおける事後期待値である．20回の履歴のうち各回で評価した平均を図 2 に示す．

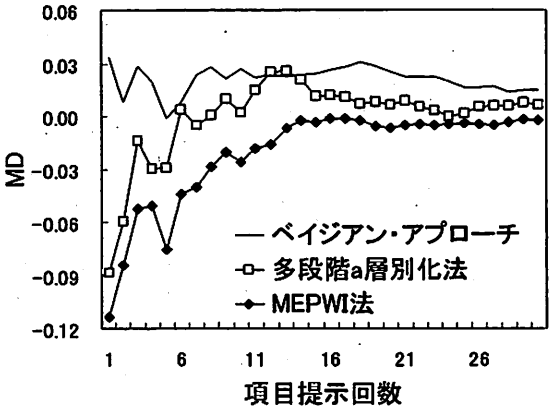

図2　MDの履歴

基本的には，どの方法もθの真値をよく再現できていると思われる．20～30回あたりではMEPWI法の
成績が平均的に一番よかったが，標本変動を考慮すると優劣はつけがたい．また，項目提示の20回以降，MEPWI法と多段階a層別化法は，平均的な真値とのズレが0.001以下である．これは偏差値で考えるとおおよそ，0.01の違いでしかないので，実質科学的に問題がないといってよい．

次は，偏差平方平均平方根（root mean square difference, RMSD）を考察する．RMSDは

$$RMSD = \sqrt{\frac{\sum_{r=1}^{1000}(EAP_r - \theta)^2}{1000}} \qquad (11)$$

で評価し，真値からの平均的な2乗距離の平方根の指標である．MDと同じく値が小さい方が推定の精度がよいとされる．結果を図3に示す．

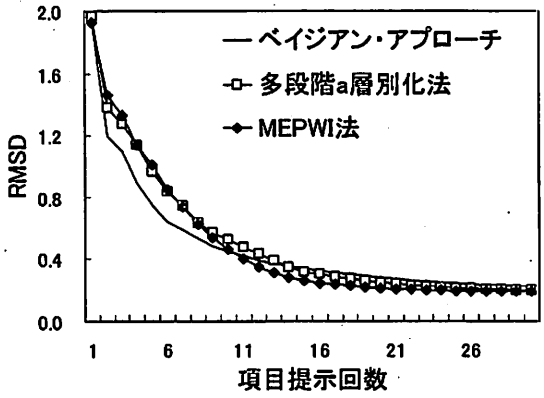

図3　RMSDの履歴

図3より，MEPWI法は，履歴の最初の方では傾き母数が小さい項目を投入する傾向があるために，ベイジアン・アプローチよりも真値からの2乗距離が大きい．しかし，履歴の中盤では，MEPWI法は他の2つの方法に比べて精度が高いといえる．

多段階a層別化法は，PSDが1.0を下回った時点でステージ2，PSDが0.5を下回ればステージ3に移行するというルールのもとで数値実験を行っており，ステージを移行する時期を早めれば，MEPWI法と大差ない成績を収めることができたに違いない．しかし，MEPWI法は，被験者に対して常に最適な項目を選択する．つまり，履歴の中盤から後半にかけて傾き母数の高い項目の出し惜しみをしない．しかしながら，多段階a層別化法は，（7）式を満たせば，ステージの中からランダムに項目が選抜されるので，必ずしも最適な項目が選択されるとは限らず，したがって，最小誤差分散法に収束の速さで勝つことはできないであろうと強く想像できる．

しかし，履歴が進むにつれ，いずれの方法も真値からの距離が小さくなっていくことが分かる．さらに，25項目以後も項目提示を続

けたときは，3 方法にそれほどの差は見られなくなる．MEPWI 法が他の方法と比較して非常に良いのは，数値実験では項目提示が 12～20 項目ほどのときであった．

最後に，PSD の履歴を図 4 に示す．

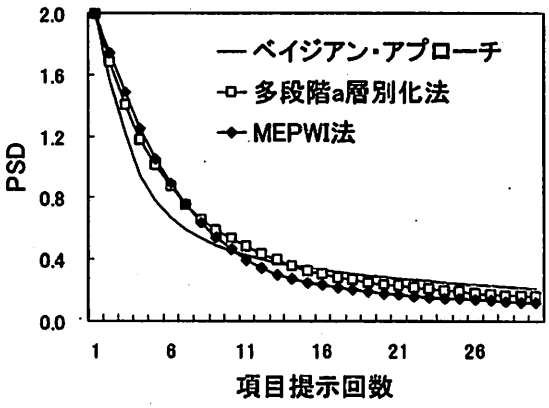

図 4　PSD の履歴

図 4 より，PSD と RMSD が似た傾向があることが分かる．やはり，初回から比較的傾きが高い項目が投入されるベイジアン・アプローチに比べて，中盤から後半で傾きが高い項目が投入される MEPWI 法が精度のよい測定ができている．また，中盤から終盤にかけて，多段階 *a* 層別化法よりも MEPWI 法の方が PSD が小さい測定が可能であった．この先，40 回，50 回と項目提示が続けば，3 者に差がなくなっていくであろうが，15～30 項目程度のときは MEPWI 法が最も PSD が小さい測定が可能である．多段階 *a* 層別化法の成績も悪くないが，項目選択にランダム性を残しているので，MEPWI 法よりも PSD が小さい測定をすることは難しいと思われる．

以上，選択された傾き母数の値，MD，RMSD，PSD の履歴を見てきたが，ベイジアン・アプローチと多段階 *a* 層別化法に比べて 20 項目程度のテストにおいては，MEPWI 法が優れていることが確認された．経験的に言えば，20 項目というのは，CAT としては，長すぎず短すぎずといったテストの長さ（Test Length）であると思われる．なお，項目提示の 5, 10, 15, 20, 25, 30 回の 6 回における上記の要約を表 1 に示した．

表 1　第 5, 10, 15, 20, 25, 30 回における傾き母数・MD・RMSD・PSD

	履歴	BA*	MaS**	MEPWI
傾き母数の平均	5 回	1.022	0.694	0.743
	10 回	1.052	1.034	1.613
	15 回	1.040	1.539	2.022
	20 回	1.065	1.547	1.703
	25 回	1.063	1.553	1.519
	30 回	1.053	1.452	1.357
MD	5 回	−0.001	−0.029	−0.075
	10 回	0.027	0.002	−0.026
	15 回	0.024	0.012	−0.003
	20 回	0.026	0.007	−0.007
	25 回	0.017	0.002	−0.004
	30 回	0.015	0.007	−0.002
RMSD	5 回	0.749	0.968	1.008
	10 回	0.449	0.526	0.462
	15 回	0.339	0.319	0.262
	20 回	0.281	0.246	0.214
	25 回	0.241	0.211	0.193
	30 回	0.222	0.197	0.190
PSD	5 回	0.780	1.015	1.054
	10 回	0.455	0.534	0.464
	15 回	0.344	0.327	0.250
	20 回	0.284	0.239	0.180
	25 回	0.243	0.189	0.144
	30 回	0.212	0.156	0.118

*BA：ベイジアン・アプローチ

**MaS：Multistage *a*-Stratified（多段階 *a* 層別化法）

表 1 より，MEPWI 法の特筆すべき明らかな利点は，履歴の中盤から後半にかけての RMSD の精度の高さ，および PSD の小ささであるが，当然のことながら，履歴が 40 回，50 回と進む状況では，MD，RMSD，PSD の違いは 3 方法間で区別がなくなるだろう．

4. 考察

本研究では，IRT に基づいた CAT の項目選択ルールについてシミュレーション研究を行った．MD，RMSD，PSD の観点からは，項目提示が 20 項目であるときは，MEPWI 法が最も優れている．しかし，項目バンクにある未提示項目全てについて(8)式を評価するのは計算機の負担になる恐れがあるので，(7)式と併用するなどの工夫が必要であろう．

なお，本研究のシミュレーション状況が，必ずしも一般的なテスト実務家の置かれている状況ではないであろう．つまり，項目バンクに 3000 もの項目がない状況も現実的には多いだろう．したがって，CAT 作成者には，項目選択ルールを選ぶ上で，なお色々な状況と相談する必要があると思われる．しかしながら，30 回程度の項目提示状況は，CBT の利用場面での通常の回数であり，そのような状況の下では，常に MEPWI 法が優越するということは 1 つの有力な資料であると思われる．

文献

CHANG, H.-H. & YING, Z. (1996) A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, 20 : 213-229.

CHANG, H. -H. & YING, Z. (1999) a-Stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, 23 : 211-222.

CHANG, H. -H. & STOUT, W. F. (1999) The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika*, 58 : 37-52.

廣瀬英子 (2000) 心理測定尺度のコンピュータ・テスト化に向けての最近の動向 教育心理学研究，48，235-246.

廣瀬英子 (2004) 測定・評価に関する研究動向と展望ーテスト研究と評価研究ー，教育心理学年報，43，99-106.

池田央 (2003) コンピュータを利用した心理・教育テストの動向ー米国を例に 情報処理，44，940-945.

池田央，林 則生 (2004) Computer Based Testing の現状と開発 植野真臣 (編著) 知識社会のための情報・統計科学，長岡技術科学大学 (pp.40-57)

菊地賢一 (2003a) Microsoft Office を利用した CAT システムの開発 日本行動計量学会第 31 回大会発表論文抄録集, 214-215.

菊地賢一 (2003b) MS Word と Excel を利用した汎用的 CAT システムの開発 日本テスト学会第 1 回大会発表論文抄録集，33-34.

MILLS, C. N. & STOCKING, M. L. (1996) Practical issues in large-scale computerized adaptive testing. *Applied Measurement in Education*, 9, 287-304.

村木英治，萩原康仁 (2002) 多段階 *a* 層別化方法のシミュレーションによる検証 日本行動計量学会第 30 回大会発表論文抄録集，246-249.

永岡慶三，植野真臣 (1992) 多元的適応型テストシステムのアルゴリズム 日本教育工学雑誌，15，157-165.

OWEN, R. J. (1975) A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal or the American Statistical Association*, 70 : 351-356.

VAN DER LINDEN, W. J. GLAS, C. A. W. (2000) *Computer Adaptive Testing: Theory and practice*. Dordrecht, The Netherlands: Kluwer Academic Press.

WAINER, H. (2000) *Computerized Adaptive Testing: A Primer*. NJ: Lawrence Erlbaum Associates, Inc.