J-CATの項目プールデザインにおける項目選択方法の比較

―能力推定効率と項目露出の観点から―

今井　新悟　　菊地　賢一　　平村　健勝

要　旨

J-CAT（日本語コンピュータ適応型テスト）[1]の異なる項目選択法における能力推定効率と出題項目の偏り解消の課題について、項目プールの実データで比較した。両課題が相反する関係にあり、バランスの最適化を図るべきことを確認した。項目困難度と項目識別力を評価する項目選択法より推定能力値と項目困難度の絶対誤差を最小にする方法が優れていることを、能力推定計算の単純性、能力推定の効率劣化の回復の容易さの点から論じた。

【キーワード】コンピュータ適応型テスト　J-CAT　項目プール　項目選択　項目露出

Comparison of Item Selection Methods in Designing the J-CAT Item Pool

IMAI Shingo, KIKUCHI Kenichi, HIRAMURA Takekatsu

【Abstract】We compared the efficiency and the effect of avoiding bias in the process of item selections by analyzing actual data from the J-CAT (Japanese Computerized Adaptive Test). Since these two goals are conflicting, it is recognized that optimizing the balance between the two is necessary. We argued that minimizing the absolute error between the estimated ability and the item difficulty is preferable to a method which evaluates item difficulty and item discrimination, by referring to the fact that the former is easy in both computing and recovery from the deterioration of inefficiency in the ability estimation.

【Keywords】computerized adaptive test, J-CAT, item pool, item selection, item exposure

１．背景と目的

コンピュータ適応型テスト（CAT）はオーソドックスな紙媒体テスト（paper & pencil test）に比べ、少ない出題数・短時間でかつ能力測定精度が高い。また、出題される問題（以降項目と呼ぶ）の難易度や受験者集団の能力の高低に影響されない測定が可能になっている。このような優れた利点があるCATであるが、一方で項目を多数集めて項目プール[2]を構築したり、コンピュータのシステムを開発したりと、敷居の高い方法でもある。本稿ではプレースメントテストや個人の能力判定などの用途を想定したJ-CAT（Japanese computerized adaptive test）の項目プール構築の際の項目選択の方法について考察する。

２．項目選択の方法

適応型テストの項目選択の方法についてはこれまでの研究でいくつもの提案があり、現在でも新たな提案が続いている。各種条件が変わることにより、項目選択方法の優劣も変化するため、唯一絶対の項目選択方法が確定しているわけではない。いずれの方法でも、その評価は以下の二点を中心に議論される。一つは、能力推定の精度である。少ない項目でより精度の高い推定が適応型テストの目指すところであるから、能力推定値が能力真値に収束する速さの課題とも言える。二つめは、項目露出（item-exposure）のリスクをどれだけ軽減・回避できるかである。特定の項目が繰り返し出題されてしまうことにより、項目の漏洩が起こる危険性が高まる。項目露出の問題はすなわち出題項目の偏りの問題であり、露出過剰（over-exposure）のほかに露出不足（under-exposure）の問題も生じる。頻繁に出題されてしまう項目があれば、その逆に、ほとんどあるいは全く出題されない項目もあるということである。出題の偏りを少なくすれば、露出過剰と露出不足がともに改善される。項目選択の方法を考える際には以上二つの課題を解決しなければならないのだが、実はこの二つの課題はトレードオフの関係にある。能力推定の精度・速度を上げれば、偏りが大きくなってしまう。項目選択ではこのバランスの取り方が課題になる。

能力推定の精度・速度を向上させるためには、項目情報量が大きい項目を選択する方法が一般的である。項目応答理論の２パラメータモデルでは項目識別力が高く、項目困難度が推定能力値に近いときに項目情報量が最も高くなる（Hambleton and Swaminathan 1985）。項目識別力の高い項目が項目情報量も高くなる。よって項目識別力の高い項目から選択されることになり、項目選択に偏り（特定の項目が出題されやすくなること）が生じて、項目露出が生じてしまうことになる（Chang and Ying 1999）。これを回避する方法として色々な方法が提唱されてきた。

Sympson and Hetter（1985）は項目情報量で項目を選択し、その項目が使用される回数を記録し、使用回数にフィルターをかけて制限することで露出過剰を抑える方法を提唱した。しかし、この方法では、識別力が低い項目はもちろん、中程度の項目でも選択されず

に使われずじまいになってしまい、能力推定の効率も下がることが指摘されている（Hau & Chang 2001）。使用回数を制限しても、識別力の高い項目が早くその回数制限に達するため、項目選択の偏りは解消されない。

Chang and Ying（1999）、Hau and Chang（2001）では、能力推定の効率と出題項目の偏りのバランスをとるという、相反するジレンマが解決できるとする多段階*a*-層別化法（multistage *a*-stratified method）を提唱している。この方法ではまず、項目プールを項目識別力のパラメータ値に従い、いくつかの層に分ける。テストの最初の方の層には項目識別力の低い項目を、後方の層になるに従って項目識別力が高い項目を割り振る。そして各層でいくつかの項目を選択して出題する。よって、テストの初期には項目識別力の低い項目が出題され、テストが進むにつれて識別力の高い項目が出題される。各層内の項目選択は、推定能力値に最も近い項目困難度パラメータを持つ項目が選択される。普通、テスト前半で項目識別力の高い項目を使ってしまい、識別力の高い項目が枯渇してしまうことによる能力推定精度の劣化が起こるが、層別化は、その劣化を緩和できる。ただし、「項目情報量を使う方法と同程度かそれ以上の能力推定の効率」（Hau and Chang 2001）を保証するものではないだろう。

また、現実に層化しようとすれば、層化するに足る適度な分散を持った項目識別力を持つ大量の項目が必要になる。さらに、Hau and Chang（2001）のシミュレーションでは、項目識別力を一律に0.5、1.0、1.5、2.0にし、４層に分けて各層に同数の項目を想定しているが、実際の項目プール構築で、1.5や2.0の非常に高い項目識別力を持つ項目を他の層と同数程度集めるのは至難である。さらに、各層での打ち切り条件の設定が困難になるという懸念がある。というのは、項目識別力の低い層と項目識別力の高い層とでは、能力推定の精度が異なる。このような条件の異なる層での打ち切り条件の妥当性が問われる。Hau and Chang（2001）では、単純に各層で一定の出題数を層内の打ち切り条件としている。よって、出題数はすべての受験者で一定である。この方法はできるだけ少ない出題数で推定精度を高めようという適応型テスト本来の目的・理念を反映していない。

３．シミュレーション

本稿では、J-CATの開発初期に暫定的に使用した、項目数が82項目でかつ項目識別力パラメータが全体に低めという現実的かつ厳しい条件の項目プールのデータを用い、現実的な項目選択の方法を検証する。

3.1　正誤判定と能力推定

Rを使ってシミュレーションを行う[3]。式(1)は項目応答理論の２パラメータモデルにおいて、能力値がθである受験者がある項目において正答する確率を表している。P（θ）は、通

常、式⑴のようなロジスティック関数で与えられ、exp（ ）は指数関数、Dは定数であり、1.7とすることが多い。なお、aとbは、それぞれ、識別力パラメータと困難度パラメータと呼ばれる。識別力はある問題項目がどれだけ受験者の能力判別ができるかのパラメータであり、困難度パラメータはある問題項目に50％の確率で正答できる受験者の能力に等しい。

仮に能力真値（θ）を0.0に設定し、受験者が正解する確率のと0.0～1.0の間で発生させた疑似乱数を比較することによって正答するか、誤答するかを判定する。

⑴　$$\mathrm{P}(\theta)=\frac{1}{1+\exp(-Da(\theta-b))}$$

能力値推定は芝（1991）のベイズEAP（Expected a posteriori）による。連続量はコンピュータで計算できないので，各離散点で数値積分する。つまり、グラフを短冊状に分けてその各短冊の面積を足し合わせるものである。

3．2　項目選択の方法

問題項目選択はOwen（1975）の方法に基づく方法および絶対誤差を利用した方法の二種類の方法とし、それぞれシミュレーションを1000回繰り返して比較する。

式⑵はOwen（1975）の問題項目選択の方法で用いられる式で、その問題を出題した場合に期待される推定誤差を表す。受験者からの回答が得られた後、項目識別力、項目困難度、能力推定値を与えて計算し、$l(a, b)$が最小になる問題項目を次に出題する項目として選択する。それにより、推定誤差が最も小さくなると思われる問題を出題することができる。

⑵　$$l(a,b)=V_{n-1}(1-(1+V_{n-1}{}^{-1}a^{-2})^{-1}(\phi(\mathrm{D}))^2/(A\Phi(D)))$$

ただし、

$A=\Phi(-D)$

$D=(b-M_{n-1})/(a^{-2}+V_{n-1})^{1/2}$

$\phi(\cdot),\Phi(\cdot)$：標準正規分布の密度関数、分布関数

M_{n-1},V_{n-1}：n－１問目の能力推定値、分散

a：項目識別力パラメータ

b：項目困難度パラメータ

方法A：項目識別力と項目困難度を用い、$l(a, b)$が最小となる項目を選択する。なお、本稿の２パラメータモデルでは、Owen（1975）のcを０とした。

既に述べたように、推定能力値が項目困難度に近く、かつ項目識別力が高い項目で項目情報量が大きくなる。そのような項目を優先的に選択することにより、能力推定値の収束が速くなると予測される。項目情報量は困難度と識別力のうち、識別力の影響が困難度のそれよりも大きいため、結局識別力が高い項目が最優先で選択され、項目選択の偏りが生じると予測される。

方法B：最小絶対誤差法。EAP推定能力値と項目困難度の絶対誤差が最も小さくなる項目を選択する。

項目選択において識別力の影響を受けないことから、方法Aで生じるであろう項目選択の偏りが緩和されることが予測される。また、この方法は、項目選択のための計算が単純で、システムへの負荷が軽減される。一方、能力推定値の収束は方法Aに比べて遅くなり、出題項目数が増えることが予測される。以上のことをシミュレーションで確認する。なお、項目選択の際に、一度使われた項目は二度と使われないようになっている。また、最初にカウントされない問題項目が数問ある。それらの回答パターンが能力推定に使われることもない。これらの正答率は（能力推定に使われる）最初の出題問題の困難度を決めるために使われる。

3．3　データ

J-CATの開発初期のアイテムプールのデータを用いる。このテストは４つのセクションからなるが、その一つ、聴解のセクションを用いる。このプールには82の項目が格納されており、それぞれの項目にプレテストの回答パターンをBILOG-MG[4]によってIRTの２パラメータモデルで算出した項目困難度と項目識別力のパラメータが付与されている。パラメータの基礎統計量（平均、標準偏差、最小値、最大値）は次の通りである。

表１　項目パラメータの基本統計量

	パラメータ	
	項目識別力	項目困難度
平均	0.591	-0.632
標準偏差	0.223	1.429
最小値	0.254	-4.095
最大値	1.513	3.146

4．結果と分析

4．1　推定値差分

1回前の能力推定値と当該能力推定値の差分を推定値差分とする。推定値差分が一定量まで落ちるのに何問の出題が必要か、推定値差分の条件を0.4、0.3、0.2、0.1の4種類に設定して、それぞれの推定値差分に到達するときの出題数の1000回のシミュレーションの平均を以下の表に示す。これを用いて方法間の比較を行う。

表2　推定値差分と出題項目数

	推定値差分			
方法	0.4	0.3	0.2	0.1
A	4.504	6.421	11.290	30.941
B	7.523	10.766	17.594	38.560

予測通り、方法Aの出題項目数は他の方法に比べて少ない。推定値差分の条件が0.4から0.1に向けて厳しくなるにつれて、方法の違いにおける出題数の差は大きくなるが、差の比は小さくなる。

4．2　能力真値との平均絶対誤差

一定数の項目を出題したときの、設定した能力真値と推定能力値の差を比較する。項目数の条件を5項目、10項目、15項目、20項目、30項目、50項目に設定して項目選択の1000回のシミュレーションで、式(3)を使って各方法で設定された能力真値0.0と推定能力値の差を取り、その絶対値を平均化して足し合わせた平均絶対誤差（MAE）[5]を求める。これにより、出題項目数と項目選択方法を条件とした能力推定精度の相対的な比較を行う。

(3)　$$MAE = \frac{\Sigma\left|\hat{\theta} - \theta\right|}{n}$$

ただし、$\hat{\theta}$:能力推定値、θ：能力真値、n:1000

表3　能力の真値と推定値の平均絶対誤差（MAE）

	出題項目数					
方法	5	10	15	20	30	50
A	0.377	0.330	0.299	0.271	0.245	0.218
B	0.439	0.377	0.327	0.314	0.271	0.217

いずれの方法でも出題項目数が増加するにつれて平均絶対誤差が減少するが、項目数が少ない場合に比べて、項目数が多い場合の方が、方法間での差が小さくなる。これは出題項目数をある程度確保できるのであれば、方法間の違いが解消されることを示している。

出題項目数が少なく設定されるテストにおいては、方法Aを採用する利点があるが、出題項目数をある程度確保できるテストでは、出題項目の偏りが懸念される方法Aを採用する利点が薄れる。項目プールのサイズが小さい場合には、特に項目露出の危険性が高くなるので、出題項目の偏りを小さくする項目選択法を優先すべきであろう。一方で、方法Bで方法Aと同程度の能力推定精度を得たい場合にはどの程度出題項目数を増やさなくてはならないだろうか。例えば、方法Aで10問出題された時点でのMAEは0.330であった。方法Bでこれと同程度の値を得るためには、5問弱増やして15問弱の出題が必要になる。出題項目の偏りの回避と能力推定精度の向上はトレードオフの関係にあるが、方法Bで方法Aと同程度の能力推定精度を得るための出題項目数の増加はJ-CATの実データの場合、現実的な範囲に収まっていると言えるため、方法Bを採用するのが賢明であろう。

次に、実際にテストをオペレーションする際に必ず検討されることにテストの長さ、つまり打ち切り条件（stopping rule）の設定がある。より実用的な情報を得るために、以下では、能力推定値の収束の推移を勘案しながら打ち切り条件の設定について考察し、改めて、方法Aと方法Bを比較して、方法Bが優位といえるかを検証する。

4．3　打ち切り条件

打ち切り条件を推定値差分と出題数の上限との組み合わせで以下のように設定する。推定値差分は0.4、0.3、0.2の条件を、出題数上限は5項目、10項目、15項目、20項目、30項目、50項目の5種類の条件を設け、それぞれの条件で方法Aと方法Bの場合で、出題される項目数の平均を以下の表に示す。

表4　推定値差分が0.4の場合

	出題数上限					
方法	5	10	15	20	30	50
A	4.442	4.504	4.504	4.504	4.504	4.504
B	5	7.427	7.523	7.523	7.523	7.523

表5　推定値差分が0.3の場合

	出題数上限					
方法	5	10	15	20	30	50
A	5	6.406	6.421	6.421	6.421	6.421
B	5	9.483	10.718	10.766	10.766	10.766

表６　推定値差分が0.2の場合

	出題数上限					
方法	5	10	15	20	30	50
A	5	9.723	11.217	11.282	11.290	11.290
B	5	10	14.765	17.244	17.594	17.594

例えば、推定値差分0.4で出題数上限が５項目での方法Bの出題数の平均は、上限いっぱいの５項目である。これは、シミュレーションのすべてにおいて出題数上限の５項目の出題があったからであり、テストが５項目の出題で終了したときに、推定値差分が0.4より大きかったことを意味する。出題数の上限を10項目に増やすと、出題数の平均が7.427となる。さらに上限を上げて15項目にしても出題数の平均は微増するのみで、それ以降は出題上限数を上げても出題数の平均に変化がない。これは、シミュレーションのすべてにおいて出題数がその上限に達しなくなったことを意味する。よって、概ね15項目以上の出題数上限を設けても意味がないということになる。以上のことを踏まえた上で出題数上限をさらに細かく設定して、出題数平均の変化を詳しく見てみる。ここでは、推定値差分0.3の場合を見る。

表７　推定値差分が0.3の場合

	出題数上限								
方法	7	8	9	10	11	12	13	14	15
A	6.192	6.357	6.395	6.406	6.414	6.419	6.421	6.421	6.421
B	6.982	7.928	8.784	9.483	10.025	10.415	10.585	10.663	10.718

上限となる項目数が少ないところでは、出題上限数の増加にほぼ比例して、平均出題数も増加することが分かる。そして、上限数が増えるに従い、平均出題数の増加率は鈍り、上限数が13項目から14項目に移るところで、平均増加数の増加が0.1を切る。このあたりに出題上限数を設定しておけば、推定値差分が0.3に達するのを概ね妨げないといえる。そして、これ以上出題数を増やした場合、推定値差分は小さくなるものの、その変化は緩やかになり、テストの経済性の観点からは望ましいものではなくなる。

次に、方法Aと方法Bにおける能力の真値と推定値の平均絶対誤差（MAE）を表３で14項目に最も近い15項目の場合で確認する。MAEについては、方法Aが0.299、方法Bが0.327であり、方法Bが方法Aの1.09倍である。つまり、MAEは方法Bが方法Aより１割程度推定の精度が劣るといえる。この程度の劣化であれば、許容範囲内と言えよう。

5．まとめ

J-CATにおける項目露出制限と能力推定の精度を高めるという要求課題について、初期段階での実際の項目プールの例に基づき、項目選択のシミュレーションを通して考察した。この二つの要求課題は相反するものであり両方を同時に満たすことできない。そして、以下の理由により、項目困難度と項目識別力を共に評価する項目選択方法よりも能力推定値と項目困難度との差が最も小さくなる項目選択法である方法Bが相対的に優れていると結論付けた。第一に、出題の偏りが少ないことである。出題の偏りを抑えて項目露出の危険度を低くすることができる。第ニに、能力推定の精度が相対的に劣化するとしても、あまり多くない出題数増加でその精度が回復できることである。これに加えて第三に、能力推定の計算が相対的に単純であることである。アルゴリズムが単純であり、システムへの負荷が少ないという運用上の利点がある。

謝辞

本研究は科学研究費補助金基盤研究（A）課題番号18202012（平成18～21年度）の助成を受けて開発したJ-CATのデータを使って行った。本稿著者以外の研究分担者は伊東祐郎氏、中村洋一氏、本田明子氏、赤木彌生氏、中園博美氏であり、ここに記して感謝する。

注

1．このテストの詳細は、今井・赤木・中園（2012）およびJ-CAT Project（以下のURL）を参照。
http://www.intersc.tsukuba.ac.jp/～imai/j-cat-project/home.html

2．項目プールは項目バンクとも呼ばれるが、本稿では、実際にシステムに入っていて使われている項目の集合体を項目プールと呼び、将来使う可能性のある項目すべてを集めたものを項目バンクとして区別する。

3．Rは統計計算とグラフィックスのためのフリーソフトである。
http://www.r-project.org/参照。

4．BILOG-MGはScientific Software International社のIRT分析のためのソフトウエアである。

5．真値と測定値のずれを評価するためにRMSEが用いられることが多いが、平均絶対誤差よりも保守的な値になる。

参考文献

今井新悟・赤木彌生・中園博美（2012）『J-CATオフィシャルガイド：コンピュータによる自動採点日本語テスト』ココ出版

大友賢二（1996）『項目応答理論入門』大修館書店

芝祐順（編）（1991）『項目反応理論－基礎と応用』東京大学出版会

柴山直・野口裕之・芝祐順・鎌原雅彦（1987）「最適化テスト方式による語彙理解力の測定」『教育心理学研究』35（4）：363-367

Chang, H.H. and Ying, Z.（1999）“A-stratified multistage computerized adaptive testing” *Applied Psychological Measurement*, 23（3）：211-222

Chen, Shu-Ying and Robert D. Ankenman（2006）“Effects of Practical Constraints on Item Selection Rules at the Early Stages of Computerized Adaptive Testing” *Journal of Educational Measurement*, 41（2）：149-174

Hambleton, R.K. and Swaminathan, H.（1985）*Item Response Theory. Principles and Applications*. Boston：Kluwer-Nijhoff.

Hau, Kit-Tai and Chang, Hua-Hua（2001）“Item selection in computerized adaptive testing：Should more discriminating items be used first?” *Journal of Educational Measurement*, 38（3）：249-266

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

Sympson, J.B. and Hetter, R.D.（1985）“Controlling item-exposure rates in computerized adaptive testing” *Proceedings of the 27th Annual Meeting of the Military Testing Association*：973-977