Bulletin of the School of Education,
Nagoya University (Educational Psychology)
1992, Vol. 39, 15 − 21.

項目困難度の分布の偏りがＩＲＴ項目パラメタの発見的推定値に与える影響

野　口　裕　之

問　題

項目反応理論で項目パラメタを推定する方法のひとつに“発見的方法（heuristic method)”がある。この方法は，古典的テスト理論の項目統計量と項目反応理論の項目パラメタとの間に，

$$a = \rho / \sqrt{(1-\rho^2)} \qquad (1)$$

$$b = -\phi^{-1}(\pi)/\rho \qquad (2)$$

という関係が導かれる（Lord & Novick, 1968, p. 378）ことを利用する。ここで，a及びbはそれぞれ項目反応理論の識別力及び困難度パラメタを，ρは項目得点とテストが測定する潜在特性との双列相関係数，πは項目通過率を，さらに，$\phi^{-1}(\cdot)$は標準正規分布関数の逆関数をそれぞれ表わす。ただし，これらの関係は常に成立するのではなく，

i）テストが１次元性を持つ
ii）各項目の項目特性曲線が正規累積型である
iii）テスト項目間で局所独立の仮定が成り立つ
iv）潜在特性値の分布が正規型である

という条件が満足された時に成立する。

項目パラメタの推定法には，この発見的推定法の他に，同時最尤法，周辺最尤法，ベイズ推定法，非線形因子分析に基づく方法などがあり，むしろ，発見的方法以外の方がよく用いられている。わが国で発見的推定法を実際の研究に用いた例には，芝（1978），松井（1992），野口（1992）などがあるが，外国雑誌（*Psychometrika, APM, JEM, JES*）を含めても項目反応理論に関する研究の中で用いられた例はきわめて少ない。

発見的推定法があまり用いられない理由としては，

i）潜在特性が正規分布するとみなせる被験者集団にのみ適用が可能である
ii）項目反応理論の項目パラメタと古典的テスト理論の項目統計量との関係を与える式が標本ではなく母集団において成立するものである
iii）項目パラメタ推定値の標準誤差が求まらない
iv）一致性などの統計的性質が確認されていない
v）計量的方法の研究者の関心を引かない
vi）容易に利用できるコンピュータプログラムがない

などが考えられる。

しかしながら，発見的推定法のプログラムを開発したり，他の推定法との比較を行なうなどの研究もいくつか報告されてきた。

Urry（1974）では，(1) (2)式の関係をグラフに表わして，このグラフを用いて項目パラメタを近似的に求める方法を提案した。項目得点と潜在特性との双列相関係数は項目得点とテスト得点との双列相関係数で代用したが，テスト項目が等質（$KR_{20} \geqq .90$）で十分長いテスト（$n \geqq 80$）ならば十分に問題なく，さらにこの方法を実際に用いるには被験者数が2000名を超えることが望ましいとした。

Schmidt（1977）では，Urry（1974）の方法が系統的に識別力パラメタを小さめに，困難度パラメタを絶対値で大きめに推定するというバイアスを持つことを指摘した。そして，潜在特性θの推定値の信頼性係数に基づき，希薄化修正の公式を利用してバイアスを除去する方法を提案した。

Jensema（1976）では，項目パラメタ値を予め設定した上でコンピュータシミュレーションに基づいて被験者の項目反応データを発生させ，それから得られた項目パラメタ推定値と真値との相関係数を計算した。その結果，真値と発見的推定法との間の相関は，識別力で.798，困難度で.963であり，真値と同時最尤法との間の相関は，識別力で.863，困難度で.971であり，計算時間を考慮すると発見的推定法が驚く程良い結果をもたらすとした。

Ree（1979）は，テストの長さを80項目に，被験者数を2000名に固定　した上で，被験者特性値の分布が一様分布，負に歪んだ分布，正規分布の３つの状況を設定し，

発見的推定法と同時最尤法とを比較した。その結果，発見的推定法は正規分布の下で最も良く機能し，その場合は同時最尤法とほとんど変わらないとした。

Swaminathan & Giffort (1983) は，テストの長さに10，15，20，80項目，被験者数に50，200，1000名，特性値の分布に標準正規分布，一様分布，負に歪んだ分布の全部で4×3×3の36通りの状況を設定し，同時最尤法と発見的推定法との比較を行なった（プログラムにはそれぞれLOGISTとANCILLESが用いられた）。その結果，識別力パラメタについては，短いテストではいずれも推定が良くないこと，発見的推定法では過大評価の傾向があるが長いテストでその程度は小さくなること，等が，そして困難度パラメタでは，発見的方法も同時最尤法の何れも大変良い結果が得られること，特に短いテスト・少数被験者の場合には同時最尤法の方がよいこと，等が，被験者特性値の分布については，発見的推定法の方が分布の歪みの影響を受けやすいこと，等が明らかにされた。そして，全体として見ると，少数項目の場合は同時最尤法が非常に優位に立つが，項目数及び被験者数が増加するとその違いはほとんどなくなり，発見的推定法はコンピュータの計算時間の点で非常に有利になるとした。

以上の研究では，項目数及び被験者数が増加する（例えば，80項目で2000名など）と，被験者特性値分布が歪みを持たない場合には，発見的推定法は同時最尤推定法とほとんど同じ結果を与えており，コンピュータの計算時間の点でかなり有利であることが示された。

ところで，以上の研究では項目数，被験者数，被験者特性値の分布については様々な条件を設定して検討されたが，項目パラメタとりわけ困難度パラメタの分布については一様分布が設定されることがほとんどであった。しかしながら，実際のテストでは項目困難度の分布が一様分布に限られているという訳ではない。例えば，1990年度に実施された日本語能力試験（実施主体は国際交流基金と日本語教育学会）については，項目困難度の分布が表1に示したようになっている。全部で6種類のテストが示されているが，それらの何れのテストも困難度の平均がやや易しい方に寄っているが，特に1級・2級共に聴解のテストで困難度の最大値がそれぞれ.087，.388と被験者集団の中心付近の値を示し，困難度の分布が易しい方に偏り，切断型を示している。野口（1992）ではこの点について特に検討することなく項目反応理論に基づく分析を進めている。聴解という問題項目の性質上このような分布とならざるを得ないとしても，そのような状況で発見的推定法を項目パラメタの推定に用いることの適切性について検討しておく必要がある。日本語能力試験に限らず他のテストでも，困難度パラメタの分布が一様分布とはならない場合ものあることも十分に考えられる。従って，困難度パラメタの分布型が発見的推定法にどのような影響を与えるのかについて検討する必要がある。

目　的

本研究は，項目反応理論の項目パラメタの発見的推定法に関して，テストに含まれる項目の困難度の分布に偏りがある場合に推定結果にどのような影響が生ずるかについて，困難度の分布が一様分布をする場合と比較し検討することを目的とする。その際に，テストに含まれる項目数及び被験者数についても考慮する。

方　法

本研究では，項目パラメタ推定の正確さを問題にするため，テストに含まれる項目のパラメタの真値を予め設定しておき，被験者の項目反応データをコンピュータシミュレーションを用いて発生し，このデータを基に項目パラメタの推定を行ない，真値と推定値との比較検討を実施する。

1．データ

テスト項目：項目困難度が易しいものから難しいものまで一様に分布する“一様型テスト”と項目困難度が中程度から易しい方に分布する“切断型テスト”の2種類を用意する。それぞれ，項目数は20，

表1　日本語能力試験の項目困難度パラメタの分布

		項目数	平　均	標準偏差	最小値	最大値
1級	文字・語彙	65	−1.051	1.031	−4.540	1.841
	聴　解	24	−1.619	.819	−3.098	.087
	読解・文法	46	− .328	1.282	−2.950	3.864
2級	文字・語彙	70	− .633	1.473	−3.300	6.357
	聴　解	21	−1.171	.860	−2.742	.388
	読解・文法	41	− .273	1.623	−2.873	3.765

40，80の3通りとする。項目の識別力は日本語能力試験の場合を参考にして，全ての項目について値を0.6とした。全ての項目を同じ値としたのは，できるだけ困難度の分布の変化のみが結果に反映されるように配慮したからである。これら6通りのテストに含まれる各項目の困難度は表2に示した通りである。

被験者：被験者の特性値の分布は母集団で標準正規分布するものとし，実際に用いる被験者集団には当該母集団からのランダムサンプルを充てる。被験者数は1000，2000，4000の3通り設定した。

2．プログラム

被験者集団：各被験者の特性値として標準正規乱数を発生させ，これを当該被験者集団の人数分繰り返した。

項目反応パタン：各被験者の各項目に対する反応（正答ならば1，誤答ならば0）は一様乱数を発生させ，その値Rが当該被験者の特性値と各項目のパラメタ値とから計算される項目特性関数の値Pよりも小さければ“正答（1）”，大きければ“誤答（0）”とし，この手続きを被験者数×項目数　分繰り返した。

発見的推定法：項目パラメタの推定には，芝（1978）で用いられたプログラムを本研究に合わせて修正したものを用いた。このプログラムは，（1）及び（2）式を用いる点では従来の研究と同じであるが，項目得点と潜在特性との双列相関係数を求めるのに，項目間テトラコリック相関係数行列を因子分析した結果得られた第Ⅰ因子の因子負荷の値を用いている点に特徴がある。

3．評　　価

項目困難度の分布（2）×項目数（3）×被験者数（3）の全部で18通りの項目パラメタ推定値のデータセットが得られるが，各データセット毎に，識別力パラメタ推定値の平均及び標準偏差，困難度パラメタ推定値の平均及び標準偏差，困難度パラメタの真値と推定値との相関係数，そして，識別力及び困難度パラメタの推定値と真値との間の平均二乗残差の平方根を計算し，これらの値を用いて項

表2　項目困難度の分布

	一様型テスト			切断型テスト		
困難度	20項目	40項目	80項目	20項目	40項目	80項目
−2.5						
−2.4		1	2		1	3
−2.3	1			1		
−2.2		1	2		1	3
−2.1						
−2.0	1	1	2	1	2	3
−1.9						
−1.8		1	2		2	3
−1.7	1			2		
−1.6		1	2		2	4
−1.5						
−1.4	1	1	2	2	2	4
−1.3						
−1.2		1	2		2	4
−1.1	1			2		
−1.0		2	2		2	4
−0.9	1	1	3	2	2	4
−0.8		1	2		2	4
−0.7	1	1	3	2	2	4
−0.6		1	2		2	4
−0.5	1	2	3	2	2	4
−0.4		1	2		2	4
−0.3	1	1	3	2	2	4
−0.2		1	2		2	3
−0.1	1	1	3	1	2	3
0.0		2	2		2	3
0.1	1	1	3	1	2	3
0.2		1	2		1	3
0.3	1	1	3	1	1	3
0.4		1	2		1	3
0.5	1	2	3	1	1	3
0.6		1	2			
0.7	1	1	3			
0.8		1	2			
0.9	1	1	3			
1.0		2	2			
1.1	1					
1.2		1	2			
1.3						
1.4	1	1	2			
1.5						
1.6		1	2			
1.7	1					
1.8		1	2			
1.9						
2.0	1	1	2			
2.1						
2.2		1	2			
2.3	1					
2.4		1	2			
2.5						

表中の数字は当該困難度の項目の含まれる個数を示す

目パラメタ推定の評価を行なう。

結　　果

各データセット毎に計算される，識別力パラメタ推定値の平均及び標準偏差，困難度パラメタ真値の平均及び標準偏差，困難度パラメタ推定値の平均及び標準偏差，困難度パラメタの真値と推定値との相関係数，そして，識別力及び困難度パラメタの推定値と真値との間の平均二乗残差の平方根の値は表3に示した通りである。識別力パラメタの真値は全て0.6に設定されている為，識別力パラメタ真値の平均及び標準偏差，識別力パラメタの真値と推定値との相関係数は示していない。

識別力パラメタ推定値の平均は，一様型テストについては20項目1000名から順に .570，.577，.603，.588，.584，.594，.580，.596，.581であり，切断型テストでは同じく順に .617，.596，.616，.604，.603，.607，.607，.609，.590　である。一様型テストでは真値よりもやや小さく，切断型テストでは真値よりもやや大きく値を推定する傾向が見られる。

困難度パラメタ推定値の平均は，一様型テストについては20項目1000名から順に .032，-.031，-.041，-.045，-.017，-.019，.048，-.014，-.040であり，切断型テストでは同じく順に -.828，-.834，-.829，-.756，-.737，-.785，-.774，-.733，-.766である。困難度パラメタの真値の平均が，一様型テストでは20，40，80項目全てで .000であり，切断型テストでは20項目で -.835，40項目で -.750，80項目で -.740であるから，何れの場合も推定値の平均と真値の平均とは非常に近い値が得られている。

困難度パラメタの真値と推定値との相関係数は，一様型テストについては20項目1000名から順に .997，.997，.999，.995，.996，.999，.993，.996，.998であり，切断型テストでは同じく順に .984，.994，.992，.988，.995，.997，.982，.995，.997である。何れの場合も極めて高

表3　項目パラメタの推定結果

		â平均	âSD	b平均	bSD	ĥ平均	ĥSD	bĥ相関	√a残差	√b残差
一様型テスト	20項目									
	1000名	.570	.041	.000	1.304	.032	1.373	.997	.050	.129
	2000名	.577	.051	.000	1.304	−.031	1.412	.997	.056	.153
	4000名	.603	.038	.000	1.304	−.040	1.339	.999	.038	.081
	40項目									
	1000名	.588	.056	.000	1.202	−.045	1.314	.995	.057	.171
	2000名	.584	.044	.000	1.202	−.017	1.281	.996	.047	.133
	4000名	.594	.027	.000	1.202	−.019	1.235	.999	.028	.070
	80項目									
	1000名*	.580	.064	−.030	1.170	.048	1.302	.993	.067	.212
	2000名	.596	.045	.000	1.193	−.014	1.267	.996	.045	.129
	4000名	.581	.032	.000	1.193	−.040	1.277	.998	.037	.120
切断型テスト	20項目									
	1000名	.617	.073	−.835	.745	−.828	.782	.984	.076	.140
	2000名	.596	.027	−.835	.745	−.834	.748	.994	.027	.079
	4000名	.616	.040	−.835	.745	−.829	.789	.992	.043	.108
	40項目									
	1000名	.604	.060	−.750	.754	−.756	.789	.988	.060	.125
	2000名	.603	.038	−.750	.754	−.737	.772	.995	.038	.083
	4000名	.607	.035	−.750	.754	−.785	.790	.997	.036	.077
	80項目									
	1000名	.607	.063	−.740	.795	−.774	.844	.982	.064	.167
	2000名	.609	.041	−.740	.795	−.733	.832	.995	.042	.089
	4000名	.590	.031	−.740	.795	−.766	.838	.997	.033	.080

＊ パラメータ推定が不可能な項目があった為，79項目になっている。

い正の相関が得られている。

識別力パラメタの推定値と真値との間の平均二乗残差の平方根の値は，一様型テストについては20項目1000名から順に .050, .056, .038, .057, .047, .028,.067, .045, .037 であり，切断型テストでは同じく順に .076, .027, .043, .060, .038, .036, .064, .042, .033 である。いずれの場合も平均的には真値に近い推定値が得られている。被験者数が増加すると推定値が真値に近づく傾向が見られるが，項目数の増加はかならずしも推定値を真値に近づけるのに貢献していない。一様型テストと切断型テストとの間に結果の違いは特には見られない。

困難度パラメタの推定値と真値との間の平均二乗残差の平方根の値は，一様型テストについては20項目1000名から順に .129, .153 ,.081, .171, .133, .070, .212, .129, .120 であり，切断型テストでは同じく順に .140, .079, .108, .125, .083, .077, .167, .089, .080 である。識別力パラメタと同様にいずれの場合も平均的には真値に近い推定値が得られている。なお，識別力パラメタと困難度パラメタとで尺度の原点及び単位が異なる為，これらの数値を直接比較することはできない。また，被験者数が増加すると推定値が真値に近づく傾向が見られるが，項目数の増加はかならずしも推定値を真値に近づけるのに貢献していない。一様型テストと切断型テストとの間では，同一項目数・被験者数で一様型テストが切断型テストよりも大きな値を示す場合にはその逆の場合よりも差が大きい傾向が見られる。標本誤差を考慮する必要があるが，80項目では一貫して一様型テストの方が切断型テストよりも大きな値を示している。

考　察

本研究の結果は各条件毎に1回ずつのみしかシミュレーション実験を実施していない為，標本誤差の評価ができないという難点を持っている。この点に配慮しながら考察を進める。

各条件における識別力パラメタ推定値の平均と真値との差をとると，一様型テストについては20項目1000名から順に -.030, -.023, .003, -.012, -.016, -.006, -.020, -.004, -.019 であり，切断型テストでは同じく順に .017, -.004, .016, .004, .003, .007, .007, .009, -.010 である。大部分の場合に一様型テストよりも切断型テストの方が絶対値が小さく，平均的には切断型テストの方が良い推定値を与えている。

また，各条件における困難度パラメタ推定値の平均と真値の平均との差は，一様型テストについては20項目1000名から順に .032, -.031, -.040, -.045, -.017, -.019, .048, -.014, -.040 であり，切断型テストでは同じく順に .007, .001, .006, -.006, .013, -.035, -.034, .007, -.026 である。この場合も識別力パラメタの場合と同様に，大部分の場合に一様型テストよりも切断型テストの方が絶対値が小さく，平均的には切断型テストの方が良い推定値を与えている。

以上のことから直ちに切断型テストの方が一様型テストよりも良い項目パラメタ推定値が得られるとは言えない。なぜならば，大切なのは，個々の項目について正確なパラメタ推定値が得られることであり，項目パラメタ推定値の平均が正確に真値の平均に一致することではないからである。

困難度パラメタの真値と推定値との相関係数によって，個々の項目のパラメタ推定値と真値との相対的な一致の度合いを全体として見ることができるが，この場合は全ての条件においてわずかではあるが，一様型テストの方が切断型テストよりも高い値を示している。しかしながら，いずれの場合も .98 を超える高い値を示しており，しかもこの高さは2つの変数のデータの範囲の広い事に起因するとも考えられる。

識別力及び困難度パラメタの推定値と真値との間の平均二乗残差の平方根の値は，同一項目数・被験者数で一様型テストが切断型テストよりも大きな値を示す場合にはその逆の場合よりも差の大きい傾向が見られ，80項目の場合には一貫して一様型テストの方が切断型テストよ

表4　困難度パラメタの推定値と真値との平均二乗残差の平方根（修正後）

	一様型テスト			切断型テスト		
	20項目	40項目	80項目	20項目	40項目	80項目
1000名	.129	.106	.143	.107	.125	.099
		(5)	(7)	(1)		(1)
2000名	.153	.105	.094	.079	.067	.071
		(1)	(4)		(1)	(2)
4000名	.081	.070	.088	.069	.056	.080
			(3)	(1)	(1)	

上段は平方根の値，下段は削除した項目数を表わす

りも大きな値を示している為，これまでの結果の考察を合わせると，切断型テストの方が一様型テストよりも良い推定結果が得られたように思われる。しかしながら，一様型テストの場合にはごく一部に極端に真値と異なるパラメタ推定値が得られる場合がある。ここで，あらためて困難度パラメタ推定値と真値との差の絶対値が0.3を超える項目を除いて，困難度パラメタの推定値と真値との間の平均二乗残差の平方根の値を計算してみたものが表4である。表4によると，困難度パラメタの推定値と真値との間の平均二乗残差の平方根の値は，一様型テストについては20項目1000名から順に .129, .153, .081, .106, .105, .070, .143, .094, .088 であり，切断型テストでは同じく順に .107, .079, .069, .125, .067, .056, .099, .071, .080 である。修正前と比べると一様型テストでの改善の大きいことがわかる。切断型テストでもごく一部で若干の改善が見られる。この場合でもなお一様型テストよりも切断型テストの方で値が小さい。削除された項目の困難度を検討すると，ほとんどのものが最も難しいか最も易しい項目もしくはそれに近い困難度を持つ項目である。従って，両端付近の項目のパラメタ推定は，一様型テスト，切断型テストによらず良い値が得られ難く，そのため両端付近の項目をより多く含む一様型テストで平均二乗残差の平方根の値が大きくなる傾向がでたものと思われる。このことは，また，切断型テストでも項目困難度の分布が極端な偏りをもたなければ，発見的推定法を用いてパラメタの推定を行なうことが特に問題を生じないことを意味する。ただし，本研究にはシミュレーションを各条件毎に1回しか実施していない為，結論の一般化に限界がある。さらに，シミュレーションを繰り返したり，識別力パラメタの値を変化させる，困難度の分布により多様性をもたせる，被験者特性値の分布に変化を持たせる，など設定条件を変えて検討する必要がある。

結　　論

項目困難度の分布に偏りを持つテストでもその偏りが極端なものでなければ，項目困難度が易しいものから難しいものまで一様に含まれるテスト同様に発見的推定法を用いてパラメタの推定を実施しても特に大きな問題は生じないものと言える。むしろ，項目数や被験者数の影響の方が大きいものと思われる。

文　　献

Jensema, C. 1976 A simple technique for estimating latent trait mental test parameters. *Educational and Psychological Measurement,* 36, 705-715.

Lord, F. M. & Novick, M. R. 1968 Statistical Theories of Mental Test Scores. Reading MA, Addison-Wesley.

松井　仁　1992　インクブロットテストへの項目反応モデルの応用　教育心理学研究，40，29-36.

野口裕之　1992 項目反応モデルによる分析　外国人日本語能力試験企画小委員会編　1990年度日本語能力試験実施報告書

Ree, M. J. 1979 Estimating item characteristic curves. *Applied Psychological Measurement,* 3, 371-385.

Schmidt, F. L. 1977 The Urry method of approximating the item parameters of latent trait theory. *Educational and Psychological Measurement* , 37, 613-620.

芝　祐順　1978　語彙理解尺度作成の試み　東京大学教育学部紀要，17，47-58.

Swaminathan, H & Gifford, J. A. 1983 Estimating of parameters in the three parameter latent trait model. In Weiss,D.J. (Ed.) New Horizons in Testing Latent Trait Theory and Computerized Adaptive Testing, Academic Press.

Urry, V. W. 1974 Approximations to item parameters of mental test models and their uses. *Educational and Psychological Measurement,* 34, 253-269.

(1992年8月29日　受稿)

ABSTRACT

How does the true item difficulty parameter distribution in a test affect the item parameter estimates by heuristic method?

Hiroyuki NOGUCHI

The purpose of this paper is to study how the true item difficulty parameter distribution in a test affects the item parameter estimates by heuristic method. Two types of tests were constructed. The one is "uniformly type test" in which the true item difficulty parameters have uniformly distribution from easy to hard. The other is "trancated type test" in which the true item difficulty parameters have negatively skewed distribution. Test length were fixed at 20, 40 and 80 items and the number of subjects were fixed at 1000, 2000 and 4000. The three factors —— test length (3 levels), difficulty parameter distribution (2 levels) and number of subjects (3 levels) —— were completely crossed to simulate 18 testing situations. The subjects' abilities were randomly sampled from the unit normal distribution. The item response data of each subject were generated using the uniform random number generator in computer and the item characteristic functions. The item parameters were separately estimated by heuristic method for all eighteen data sets. The estimated item parameters were compared to the true item parameters in all data sets.

It was concluded that heuristic method practically lead to good item parameter estimates, if the skewness of true item difficulty distribution in a test was not extreme but moderate. The number of items and/or subjects rather than true item difficulty parameter distribution affect the estimates of item parameters.

Key words : item response theory, item perameter estimation, heuristic method, item difficulty distribution, simulation study.