テスト問題の良否を識別係数で識別してよいか

有田清三郎，斎藤　泰一*，那須　郁夫**

医師国家試験・歯科医師国家試験などの国家試験では各問題の良否をϕ係数を使って，次のように判定している．各問題について成績上位者と下位者をとりだし，成績上位者が多く正答し，下位者が多く誤答した問題を「良い問題」と判定し，逆に上位，下位で正答した人数が変わらないか，あるいは下位のほうがよくできた問題を「悪い問題」として採点から除外する．ϕ係数はこの問題良否の判定基準として使用されている．

しかしながら，問題の良否（問題の質）を1回限りのテストの成績上位者・下位者の正答人数，誤答人数（データ）で評価することが果たして妥当であろうか．我々はこれを確率モデルを使って，コンピュータ・シミュレーションによって検討した．確率モデルを用いた解析の結果，ϕ値は大きな変動幅をもって分布することが示された．また320問からなるテストでϕ値のコンピュータ・シミュレーションを行ったところ，54問が0.20以下の「悪問」と判定された．

以上の結果から，問題の良否の判定をϕ係数の数値のみによって判定することはきわめて危険であることが示された．

（昭和63年12月12日採用）

Unreasonable Validity of the Problem in the Examination by the Discriminating Index of Phi-Coefficient

Seizaburo Arita, Taiichi Saito* and Ikuo Nasu**

To evaluate the quality of problems in the examination, quantitative indexes such as the percentage of the correct answers, the discriminating index of ϕ and so on are commonly used.

As the validity of problems, if the value of the discriminating index of a problem is less than 0.20, the problem is judged as wrong.

However this index is calculated only by the real data based on the number of examinees who correctly answered the problem.

The aim of this paper is to show the unreasonable point of this index by the stochastic model and the computer-simulation. From the numerical results, it was suggested that the validity of the problem should not be determined by the discriminating index of ϕ, but the quality of the problem itself. (Accepted on December 12, 1988) *Kawasaki Igakkaishi 15 (1) : 109—116, 1989*

Key Words ① **Validity of problems** ② **Phi-coefficient** ③ **Discriminating index** ④ **Medical examination** ⑤ **Computer-simulation**

川崎医科大学　数学教室
〒701-01 倉敷市松島577
*　同　　薬理学教室
** 日本大学松戸歯学部　衛生学

Department of Mathematics, Kawasaki Medical School: 577 Matsushima, Kurashiki, Okayama, 701-01 Japan
Department of Pharmacology
Department of Dental Public Health, Nihon University School of Dentistry at Matsudo

1. はじめに

医師国家試験・歯科医師国家試験では多数の多肢選択問題（医師国家試験では320問）が出題されるが，これらの国家試験で各問題の良否判定に，次のような数量的処理が試みられている．すなわちこの医師国家試験320問での成績上位者と下位者を同数とりだし，それぞれを成績上位群，下位群とし，各問題について成績上位群が多く正答し，下位群が多く誤答した問題を「良い問題」と判定し，逆に上位群，下位群で正答した人数が変わらない問題，あるいは下位群のほうがよくできた問題を「悪い問題」として採点から除外する．この問題良否の判定基準として使用されている数量的指標が識別係数（ϕ係数）である．

しかしながら，問題の良否を1回限りのテストの成績上位者・下位者の正答人数，誤答人数というデータだけで評価することが果たして妥当であろうか．我々はこの識別係数を数量的見地から，数学モデルとコンピュータ・シミュレーションによって検討した．

2. テスト問題における識別係数

実際の問題良否判定では，識別係数（ϕ値）は次のような形で使われている．

数百問（たとえば320問）からなるテストで，成績上位群と下位群をきめる．そのテストの数百問から1問ずつとりだし，各問について成績上位群，下位群の正答人数を調べ，この人数から識別係数（ϕ値）を計算する．この計算されたϕ値をあらかじめ決められた「ある値」（たとえば0.20）と比較し，ϕ値がある値より大きければ「良問」，小さければ「悪問」（あるいは「悪問の疑いあり」）と判定する[1)~4)]（**Fig. 1**）．

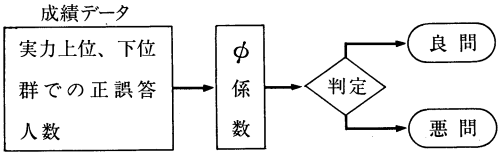

Fig. 1. The flow-chart of discriminating process by the phi-coefficient.

与えられた一つの問題に実力上位の人と実力下位の人それぞれn人が解答を行い，正答した人は実力上位群でa人，下位群でb人であったとする（**Table 1**）．

Table 1. The number of examinees who answered the problem correctly in the higher and lower knowledge level groups.

	正答	誤答	計
実力上位	a	n-a	n
実力下位	b	n-b	n

Table 1で上位群が正答，下位群が誤答の欄に集中すればするほど，この問題は実力の上位，下位を峻別していると考えることができる．識別係数（ϕ値）はこの上位，下位の識別の度合いを表す一つの数量的指標で，次の式によって計算される．

$$\phi = \frac{a-b}{\sqrt{(a+b)(2n-a-b)}} \qquad (1)$$

一般にϕ値は上位，下位の人数を同数（n）にとらなくてもよい[3)]が，ここでは説明をわかりやすくするため同一とした．このϕ値は心理学，教育におけるテストの研究等に多用されている．[5)~7)]

この式の分子（a−b）は上位と下位の正答人数の差，分母の（a+b）は正答人数の合計，（2n−a−b）は誤答人数の合計である．a−bが正，つまり上位での正答人数が下位よ

Table 2. Examples of certain phi-coefficients.

			ϕ
実力上位	100	0	1
実力下位	0	100	
	70	30	0.4
	30	70	
	50	50	0
	50	50	
	30	70	−0.4
	70	30	
	0	100	−1
	100	0	

り多ければ ϕ 値は正となる．また a−b=0 ならば ϕ 値=0，a−b が負ならば ϕ 値も負となる．たとえば n=100 で，a=100，b=0 ならば ϕ=1；a=50，b=50 ならば ϕ=0；a=0，b=100 ならば ϕ=−1 となる（**Table 2**）．

ϕ 値はもともと統計学での四分表での相関係数[4]で，四分点相関係数とも呼ばれており，これは2変量（ここでは「実力の上下」と「正答・誤答」）の関連性を測る尺度である．したがって ϕ 値は一種の相関係数あるいは一致係数（関連度）とみることができる．

実力上位の正答人数 a と下位の正答人数 b を与えると，(1)式で定義された ϕ 値は −1 から1までの値をとる．実力の上位と下位の正答人数の組合せ（a, b）によって ϕ 値は種々の値をとるが，ϕ 値がどのように変わるかを調べるため，n=100 とし，a，b を10きざみで0から100まで動かしたときの ϕ 値を計算した（**Table 3**）．

3. 確率モデルによる識別係数の分布

ここでは，実力上位群，下位群の実力と識別係数の関係を調べるため，確率モデルを導入し，実力レベルに対する識別係数の変動幅を理論的に算出する．

今，上位群の実力が一定でその正答率を p_1，下位群のそれを p_2 とする．また上位群，下位群の人数をともに n 人，ある問題についての上位群，下位群の正答人数をそれぞれ X, Y とする．このとき X, Y はそれぞれ2項分布 B (n, p_1)，B (n, p_2) に従う．

平均正答人数に対する ϕ 値は，

$$\phi = \frac{(\overline{X}-\overline{Y})}{\sqrt{(\overline{X}+\overline{Y})\,(2n-\overline{X}-\overline{Y})}} \qquad (2)$$

となる．$\overline{X}=np_1$，$\overline{Y}=np_2$ を (2) 式に代入すると，ϕ 値は，

$$\phi = \frac{p_1-p_2}{\sqrt{(p_1+p_2)\,(2-p_1-p_2)}} \qquad (3)$$

となる．ϕ 値は成績上位群，下位群の人数 n に関係せず，上位群，下位群の正答率 (p_1, p_2) のみによって決定されることがわかる．これを ϕ (p_1, p_2) と書くと，これは n=100 における ϕ (a, b) の値（**Table 3**）と同じ値をとる．(3)式は上位群，下位群の平均正答人数による ϕ 値であるから，これを ϕ ($\overline{X}$, $\overline{Y}$) と書くことにする．

Table 3. Values of discriminating index according to the knowledge level of two groups.

(実力上位)

(実力下位) b \ a	0	10	20	30	40	50	60	70	80	90	100
0	.000	.229	.333	.420	.500	.577	.655	.734	.816	.905	1.000
10	−.222	.000	.140	.250	.346	.436	.524	.612	.704	.800	.905
20	−.333	−.140	.000	.115	.218	.314	.408	.503	.600	.704	.816
30	−.420	−.250	−.115	.000	.105	.204	.302	.400	.503	.612	.734
40	−.500	−.346	−.218	−.105	.000	.101	.200	.302	.408	.524	.655
50	−.577	−.436	−.314	−.204	−.101	.000	.101	.204	.314	.436	.577
60	−.655	−.524	−.408	−.302	−.200	−.101	.000	.105	.218	.346	.500
70	−.734	−.612	−.503	−.400	−.302	−.204	−.105	.000	.115	.250	.420
80	−.816	−.704	−.600	−.503	−.408	−.314	−.218	−.115	.000	.140	.333
90	−.905	−.800	−.704	−.612	−.524	−.436	−.346	−.250	−.140	.000	.229
100	-1.000	−.905	−.816	−.734	−.655	−.577	−.500	−.420	−.333	−.229	.000

$$\phi = \frac{a-b}{\sqrt{(a+b)\,(2n-a-b)}}$$

しかしながら，実力がある受験者でも，国試でいつも実力と同等の成績をあげることができるわけではないから，正答人数X，Yは実力（p_1, p_2）を中心にある幅をもって変動する．したがって，テストによって正答人数が変動したときのϕ値の変動範囲を求めてみよう．正答人数X，Yは2項分布に従うから，X，Yが平均±2SD（SD：標準偏差）の幅で変動すると，ϕ値は，

ϕ（$\overline{X}$+2 SDx, $\overline{Y}$−2 SDy）から

ϕ（$\overline{X}$−2 SDx, $\overline{Y}$+2 SDy）まで

Table 4. Confidence intervals of phi-coefficients.

(a)

下位＼上位	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.0	0.265 ~0.194	0.369 ~0.292	0.461 ~0.383	0.538 ~0.462	0.615 ~0.535	0.688 ~0.615	0.772 ~0.694	0.850 ~0.783	0.929 ~0.876	1.000 ~1.000
0.1		0.207 ~0.078	0.312 ~0.181	0.403 ~0.282	0.494 ~0.381	0.576 ~0.461	0.658 ~0.560	0.747 ~0.653	0.844 ~0.757	0.933 ~0.876
0.2			0.181 ~0.049	0.289 ~0.154	0.377 ~0.248	0.467 ~0.344	0.559 ~0.445	0.657 ~0.546	0.753 ~0.654	0.848 ~0.784
0.3				0.175 ~0.042	0.272 ~0.133	0.365 ~0.231	0.464 ~0.341	0.566 ~0.441	0.666 ~0.559	0.769 ~0.700
0.4					0.170 ~0.032	0.264 ~0.129	0.360 ~0.231	0.472 ~0.348	0.579 ~0.464	0.691 ~0.617
0.5						0.167 ~0.026	0.271 ~0.129	0.379 ~0.247	0.489 ~0.370	0.615 ~0.538
0.6							0.168 ~0.029	0.280 ~0.153	0.405 ~0.279	0.534 ~0.460
0.7								0.181 ~0.043	0.308 ~0.178	0.457 ~0.384
0.8									0.202 ~0.064	0.365 ~0.295
0.9										0.263 ~0.185

(MEAN ± 1 SD)

(b)

下位＼上位	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.0	0.301 ~0.158	0.408 ~0.253	0.501 ~0.344	0.575 ~0.424	0.655 ~0.495	0.725 ~0.578	0.811 ~0.655	0.884 ~0.749	0.955 ~0.850	1.000 ~1.000
0.1		0.271 ~0.014	0.377 ~0.116	0.463 ~0.222	0.551 ~0.325	0.633 ~0.404	0.706 ~0.511	0.793 ~0.606	0.888 ~0.713	0.962 ~0.847
0.2			0.248 ~-.017	0.356 ~0.087	0.442 ~0.184	0.528 ~0.282	0.616 ~0.387	0.712 ~0.491	0.802 ~0.605	0.879 ~0.752
0.3				0.242 ~-.025	0.341 ~0.064	0.432 ~0.163	0.525 ~0.279	0.628 ~0.379	0.719 ~0.505	0.803 ~0.666
0.4					0.239 ~-.037	0.331 ~0.062	0.425 ~0.167	0.533 ~0.287	0.637 ~0.407	0.728 ~0.580
0.5						0.237 ~-.045	0.342 ~0.058	0.444 ~0.181	0.549 ~0.311	0.654 ~0.499
0.6							0.238 ~-.041	0.343 ~0.090	0.469 ~0.216	0.570 ~0.424
0.7								0.250 ~-.026	0.373 ~0.113	0.494 ~0.347
0.8									0.271 ~-.005	0.401 ~0.259
0.9										0.302 ~0.146

(MEAN ± 2 SD)

変動する．(SD_X, SD_Y) はそれぞれX，Yの SD とする．平均正答人数に対する $\phi(\overline{X}, \overline{Y})$ 値は **Table 2** のそれと同一である．実力上位群を縦軸に，実力下位群を横軸にとり，上位，下位の組合せで ϕ 値がどのような分布をするかを調べるため，ϕ 値の平均±1 SD，平均±2 SD を **Table 4** に示す．

上位群の実力（正答率）を $p_1=0.8$，下位群のそれを $p_2=0.5$ とすると，平均正答人数による ϕ 値は，

$$\phi\ (p_1=0.8,\ p_2=0.5)=0.313$$

となる．平均±1 SDでの ϕ 値は 0.247～0.379，平均±2 SD での ϕ 値は 0.181～0.444 と変動する．

また，$n=100$ のとき（$p_1=0.8$, $p_2=0.5$）における確率モデルによる ϕ 値のヒストグラムを **Figure 2** に示す．**Tables 3, 4, Figure 2** からも ϕ 値は大きな変動をもって分布することがわかる．

4. 識別係数の使われ方

ϕ 係数が使用される理由は **Table 2** から，ϕ 値がある値より大きければ，実力上位が多く正答し，下位が多く誤答する．すなわちこの問題は実力を正しく反映し，上位，下位を識別できる「良問」と考えるからである．また逆に ϕ 値がある値より小さければ，実力上位と下位の正答人数が変わらない，あるいは下位の方がよくでき，多く正答したと考え，上位，下位を正しく識別できない，あるいは実力が逆転して反映された「悪問」と考えるからである．そして，この良否をきめる ϕ 値の境界値が，米国の国試では 0.25，日本ではそれを若干緩和した 0.2 前後の値とされてきた．[1),2)]

しかし，この識別係数が適用できるのは1回限りのテストの成績からではなく，受験者の恒等的な実力があらかじめわかっているときであり，上位群（下位群）とは成績上位群（成績下位群）ではなく実力上位群（実力下位群）である．すなわち，受験者の実力が既知であることを土台としているにもかかわらず，この識別係数をテストによる成績での上位，下位で使用する背景には，次のような一般通念が暗黙のうちに仮定されているからではないだろうか．「実力上位の者は大抵毎回のテストでも成績は上位であり，したがってまたそのテストを構成している問題の大部分は正答する．反対に実力下位の者は毎回のテスト全体の成績が悪く，したがってそれを構成している各問題についても誤答する．」

テストの成績上位者，下位者という観点から考えると，成績と実力は必ずしも1対1に対応

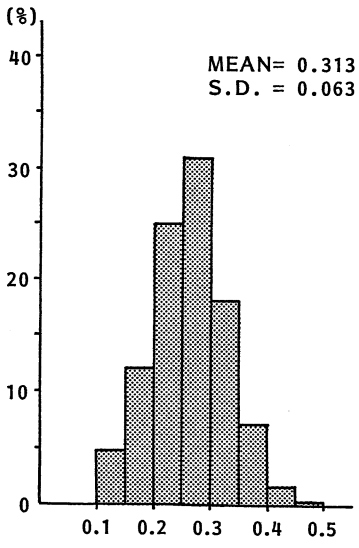

Fig. 2. The distribution of phi-coefficients with higher-score group (80 points) and lower-score group (50 points) by the stochastic model.

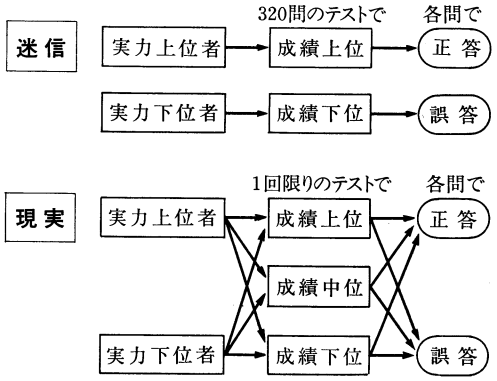

Fig. 3. The relationship between the knowledge levels and the scores of examinees.

せず，また個々の問題の正，誤答についてはなおさら成績と対応しない（**Fig. 3**）.

たとえば8割できた成績上位の者は2割の問題は必ず誤答したのである．逆に4割正答の成績下位者でも4割の問題は正答したのだから，問題ごとに考えれば，上位者が誤答し，下位者が正答した問題もありうるわけである.

「実力が判定できないとき，ただ1回のテストでの成績上位，下位者を実力の上位，下位者にみたてて，そのテストの各問について，識別係数で問題の良否を決めて良いだろうか？」

我々はこの問題を検討するため，320問からなるテストにおいて，コンピュータ・シミュレーションにより，成績上位，下位者の各問の正答，誤答数から識別係数を算出し，どのくらいの悪問がでるかを検討した.

5. コンピュータ・シミュレーション

320問からなるテストを考える．成績上位群，下位群を次のように設定する．すなわち320問のうち256問正答（100点満点で80点）した者を成績上位群，160問正答（50点）した者を成績下位群とし，各群とも100名とする．（上位群，下位群の組合せはこれとは別に（80点，40点）等も準備した.）次に，各問題に問題番号1，2，3，…，320を付け，80点の上位者は320問中のどの問題を256問正答したかをコンピュータでシミュレートした．すなわち，320問から無作為抽出により256問を選び，これを正答した問題と考え，このような操作を100回（100人分）行った．下位者についても同様に320問から160問を無作為に選んだ.

コンピュータ・シミュレーションの結果の一部を **Table 5** に示す．上位者1番目の受験者は320問中に正答できた問題番号は1，3，4，7，…，318の256問であった．順次，上位者2番目，3番目，…，100番目，下位者1番目，2番目，…，100番目の全員の正答問題番号をシミュレートした.

次に，この表で正答人数を問題番号ごとに（すなわち列の方向で）とりだし，上位者100名のうちの正答人数（aとする），下位者100名のうちの正答人数（bとする）をそれぞれ計数した．問題1では上位者で正答した人は1番，2番，4番，7番，…，98番の計82名であった．下位群についても同様に56名であった．このとき問題番号1ではa＝82，b＝56となり，

Table 5. Phi-coefficients by the computer-simulation.

		問1	問2	問3	問4	問5	……	……	問316	問317	問318	問319	問320	
上位群	1人	○	○	×	○	○	×		○	×	○	○	○	
	2人	○	×	○	○	○	×		○	○	×	○	○	
	3人	×	○	○	○	×	×		○	○	○	○	○	80
	4人	○	○	×	○	○	×		○	○	×	○	○	点
	⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮	⋮	⋮	⋮	⋮	
	100人	×	○	○	○	○	×		○	○	×	○	×	
下位群	1人	×	○	○	×	×	○		×	○	○	×	○	
	2人	○	○	×	○	×	○		×	○	○	×	○	
	3人	×	○	○	×	×	○		○	×	○	○	×	50
	4人	×	×	○	○	○	○		○	×	○	○	×	点
	⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮	⋮	⋮	⋮	⋮	
	100人	○	×	○	×	×	○		×	○	○	×	○	
上位正答人数		82		62	84						70		76	
下位正答人数		56		73	41						68		54	
識別係数		0.28		-0.12	0.44						0.01		0.23	

このa，bの値とn＝100を(1)式に代入して識別係数ϕ＝0.28を得た．

「問題の良否」の判定基準すなわち識別係数の境界値を0.2に設定した．コンピュータ・シミュレーションにより，320問中，悪問（すなわちϕ<0.2）と判定された問題数は54であった．

6. 考　察

1) 問題の良否判定に識別係数が使用可能かをコンピュータ・シミュレーションによって検討した．シミュレーションの結果，320問中54問が「悪問」(ϕ値<0.20)と判定された．これは質の上から「良問」と思われる問題でも受験者の正答，誤答数だけで「悪問」と判定される可能性があることを示している．ϕ係数の使用には，「各問ともすべて，下位者がわかれば上位者は必ずわかり，下位者がわかり上位者がわからない問題はない」ことを前提としており，「問題の難易は個人によって異なる」ことを無視している．[8)~10)]

また識別係数は，上位者のみが正答した問題，あるいは上位者が下位者に比べてはるかに多く正答した問題に対して，「良問」の判定をする．このことからも，識別係数は成績上位者の正答問題を基準にして判定していることがわかる．したがって，識別係数は「良問」の基準を問題内容ではなく，成績上位者が多く正答した問題に設定しているといえる．受験者の応答によるデータだけで，上位の者が正答できず，下位の者が正答できた「本来良問」の問題も削除することはきわめて危険である．

2) ϕ係数による問題の良否判定は受験者の正答，誤答数のデータのみに依存しているため，相対的である．受験者の質は試験ごと（医師国家試験では年度ごと）に異なるため良否の基準も年度ごとに異なることになる．本来資格試験の問題の良否に関する基準は出題された時点で設定されるべきで，解答後に設定されるべきではないことは学習評価の基本理念である．[9)] また，受験者の質が年度によってあまり変わらないと仮定してϕ係数を使用したとしても，シミュレーション結果が示すごとく，統計学的な観点から何%かの問題は必ず「悪問」として除外せねばならなくなる．

3) ϕ係数は，そのときの受験者の中で全体の成績がよかった者が一つの問題を間違って解答し，全体の成績の悪かった者がそれに正答することはあり得ないという発想から導き出されたものである．しかし，これが筆答の試験問題であったらどうだろうか．全体の成績が悪い者でも正しい解答を書いていればそれだけの点数をもらえるのである．その解答者がその問題に関してはよく勉強していた，あるいは何らかの機会にそれに遭遇したことがあったのでよく知っていたのであろう．よく知っていたと誉められこそすれ，知っているはずがないとけなされることはないはずである．どちらが教育上の効果があるかはいわずと知れたことである．ϕ係数をこのように応用するのは，本質的に歪んだ判断から出発しているのである．教育という見地から考えなおさねばならない問題である．

以上より，識別係数だけで問題の良否を判定しようとするならば，良問をも削除する危険性があり，識別係数の数値のみによって判定すべきではない．問題良否の検討には識別係数とは別の，実質的な問題内容の吟味検討が必要であろう．

この研究は文部省科研費一般C 62510149の補助を受けた．

文　献

1) Hubbard, J. P.（吉岡昭正訳）：医学教育測定．東京，医歯薬出版．1971, pp. 29—32, 45—53
2) 吉岡昭正：医師国家試験の統計学的分析．医教育 8：247—262, 1971
3) 日本医学教育学会教育開発委員会編：医学教育マニュアル 4．評価と試験：東京，篠原出版．1978, pp. 12—13

4) 文部省科研費医学教育総合班研究編：医学教育における評価と客観試験例題集．東京，篠原出版．1976，pp. 29—31
5) 芝 祐順，渡部 洋，石塚智一編：統計用語辞典．東京，新曜社．1984，p. 104
6) フライス J. L. 著（佐久間昭訳）：計数データの統計学．東京，東京大学出版会．1975，pp. 45—48
7) Lord, F. M. and Novick, M. R.: Statical theories of mental test scores. New York, Addison-Wesley. 1968, pp. 335—354
8) 有田清三郎，斎藤泰一，那須郁夫：問題の良否を識別係数で識別できるか．医教育 15：366，1984
9) 斎藤泰一，有田清三郎，那須郁夫：試験問題の評価に使用されている指標の検討．医教育 16：355，1985
10) 有田清三郎，斎藤泰一，那須郁夫：ϕ 係数の応用における問題点—テスト問題の良否を ϕ 係数で識別してよいか—．昭和61・62年度科学研究費（総合研究（A））研究報告書．1987，pp. 228—229