UDC 519.23

テストレットモデルによる英語試験問題の分析

石塚　智一*
中畝 菜穂子*
内田　照久**
前川　眞一***

要　約

2000 年度に実施された大学入試センター試験英語の試験問題について，Graded Response Model (Samejima, 1969, 1972) に基づくテストレットモデルによる分析と，2 パラメタロジスティックモデル（2PL）による分析をこころみた。利用したデータは，この年の大学入試センター試験の受験者のうち，いわゆる 5 教科型の受験者（国語Ⅰ・Ⅱ，理科のＢ科目，地歴のＢ科目または公民，数学ⅠＡ，数学ⅡＢを受験した者）249,256 人の解答である。テストレットオペレーティング特性曲線，テストレット特性曲線，テスト特性曲線や独立項目として扱った 23 の設問の項目パラメタ等をみると 2 つのモデルの間の一致はかなり高く，また，2 つのモデルに基づく能力推定値の間の相関も 0.987 であった。しかし，テストレット情報関数の違いは大きく 2PL の情報量がテストレットモデルの情報量を概ね上回る結果となった。また，能力推定の標準誤差の平均は 2PL がテストレットモデルを下回る結果となった。

1　はじめに[1]

我が国の試験はいわゆる大問形式に基づいて出題されている。大問形式の出題は多くの場合，1 つのテーマを中心として，様々な側面から知識や推論能力を問う形で作題され，単なる断片的な知識だけでなく，思考力を測るのにも適した形式だと言われる。一方では，選ばれたテーマによって出来不出来の個人差が決まってしまうとか，最初の設問への正誤によって続く設問への解答が誘導され易いなどの欠点も指摘される。

これに対して，相互に独立な小問を積み上げることによって試験を作成する方法も考えられる。このような形式では，幾つかの設問の組み合わせで思考力を測定することは困難になるが，個々の設問を工夫することにより思考力を測ることも不可能ではない。また，この方式によれば，設問の数を増やすことによって教科・科目の広い範囲に渡って出題できるので，試験の内容的妥当性が高まり，選ばれたテーマの当たり外れによって出来不出来が左右されることも少なくなる。

この 2 つの出題形式をテスト理論の観点からとらえると，後者はテスト項目への正答・誤答に対して局所独立性を仮定する項目反応理論に適した出題形式と考えられ，逆に言えば，我が

* 評価・追跡研究部門
** 特別試験研究部門
*** 問題設計基盤研究部門

[1] 本研究は，文部省科学研究費（基盤研究 C「CAT 開発へ向けての項目プールの作成－大問反応理論による試験問題の分析－」研究代表者：石塚智一）の補助により実施された。

国のように大問形式で作題された試験には，そのような項目反応理論に基づく分析が馴染まないと考えられて来た。

Wainer and Kiely (1987)は，CBT (Computer Based Testing) における適応形テストの文脈から，幾つかの項目を一塊の項目群として捉えることの利点を論じ，これをテストレット（testlet）と呼んだ。彼等は，テストレットのタイプを，階層型テストレット，線型テストレット，及びその混合型に分類して論じている。ここで我が国の大問形式の出題に関して，大問を 1 つのテストレットと考えると，彼等の線型テストレットに対応すると考えられる。そのように捉えると，我が国の大問形式の出題に対してもテストレットモデルによる分析の援用が可能となる。

我々はここで，2000 年度に実施された大学入試センター試験の英語の試験問題に対してテストレットモデルによる分析を試みるが，テストレットモデルの最大の特徴は，分析に関して必須である項目間の局所独立性の仮定を，項目の集合であるテストレット間の局所独立性で置き換えたことにある。なお，南風原(2000) は，局所独立性の仮定は，実験的独立性（先行する問題の正誤が，後に続く問題の正誤に影響を与えない様に試験問題が作られていること）のみならず，項目の 1 次元性（テストに含まれる全ての項目が，ある特定の 1 次元の能力を測るものであること）の仮定とも密接に関連していることを示しているが，ここでは，全てのテストレットは共通な英語の能力という 1 次元の潜在特性を測定しているものと仮定する。

局所独立性を仮定できない大問形式の試験に項目反応理論に基づく分析を適用すると，能力推定の標準誤差を過小評価し，結果として信頼性が過大評価されるが，能力値の推定には組織的なバイアスが見られないと言われる（Sireci, Thissen, and Wainer, 1991; Lee , 2000)。我々は，同じ試験に対して，2 パラメタ・ロジスティックテストモデルによる分析を行い，テストレットモデルによる結果との比較を行いながら，局所独立からの逸脱の影響について検討を試みる。

2 古典的パラメタによるテストの記述

2.1 試験の概略

分析の対象としたのは 2000 年 1 月に実施された英語の試験である。この試験は 528,812 人が受験しているが，将来における他年度の試験（例えば共通第 1 次学力試験）との比較を考慮し，かつての共通 1 次時代の受験生に近い集団を選びたいと考え，国語は国語Ⅰ・Ⅱを受験，理科・地歴では B 科目，数学では数学ⅠA，数学ⅡB を受験したものに限定した 249,256 人の解答を分析の対象とした。

ここで対象とする英語の試験は，次の 6 つの大問から構成されている。第 1 問は発音・強勢・会話問題，第 2 問は語彙・語法・文法に関する問題及び整序作文，第 3 問がつなぎ語の補充・文整序問題，第 4 問が図表読み取り問題，第 5 問が口語コミュニケーション問題，第 6 問が長文読解問題である。それぞれの大問はいくつかの項目から成り立っており，全項目数は 51 項目である。そのうち，複数の項目に正答できてはじめて加点される問題をそれぞれ 1 つの項目として扱うと，そのような項目の組は 3 組あるので，全項目数は 48 項目となる。ここではまず，大問ごとの正答率及び正答項目数の度数分布から，大問レベルでの今年度の試験の特徴を概観し，その後，各項目について，正答率，点双列相関係数，さらにアルファ係数などの古典的パラメタによって項目レベルでの特徴について述べる。

2.2 大問単位での特徴

各大問の正答率は 51%から 76%であった（表 2.1)。ここでの正答率は，正答項目数を各大問に含まれる項目数で除したものであって，配点によって重み付けられた大問得点率とは異なる。最も正答率が低かったのが第 4 問の図表読み取り問題，最も高かったのが第 5 問の口語コミュニケーション問題であった。図 2.1 を見ると，正答率の低かった第 4 問は比較的左右対称な分布をしていることがわかるが，その他の

大問はグラフの右寄りに山がある負の歪度を持つグラフとなっている。特に正答率が 70%を越えている第 3 問と第 5 問のグラフは右肩上がりのグラフとなっており，これらの大問は受験生にとって容易な問題であったことがわかる。

表 2.1 大問の統計量

変数	平均正答率	標準偏差	項目数	歪度
第1問	0.707	1.439	8	-0.431
第2問	0.665	2.827	18	-0.473
第3問	0.720	1.121	4	-0.755
第4問	0.517	1.382	5	0.011
第5問	0.762	1.210	5	-0.889
第6問	0.654	1.908	8	-0.364

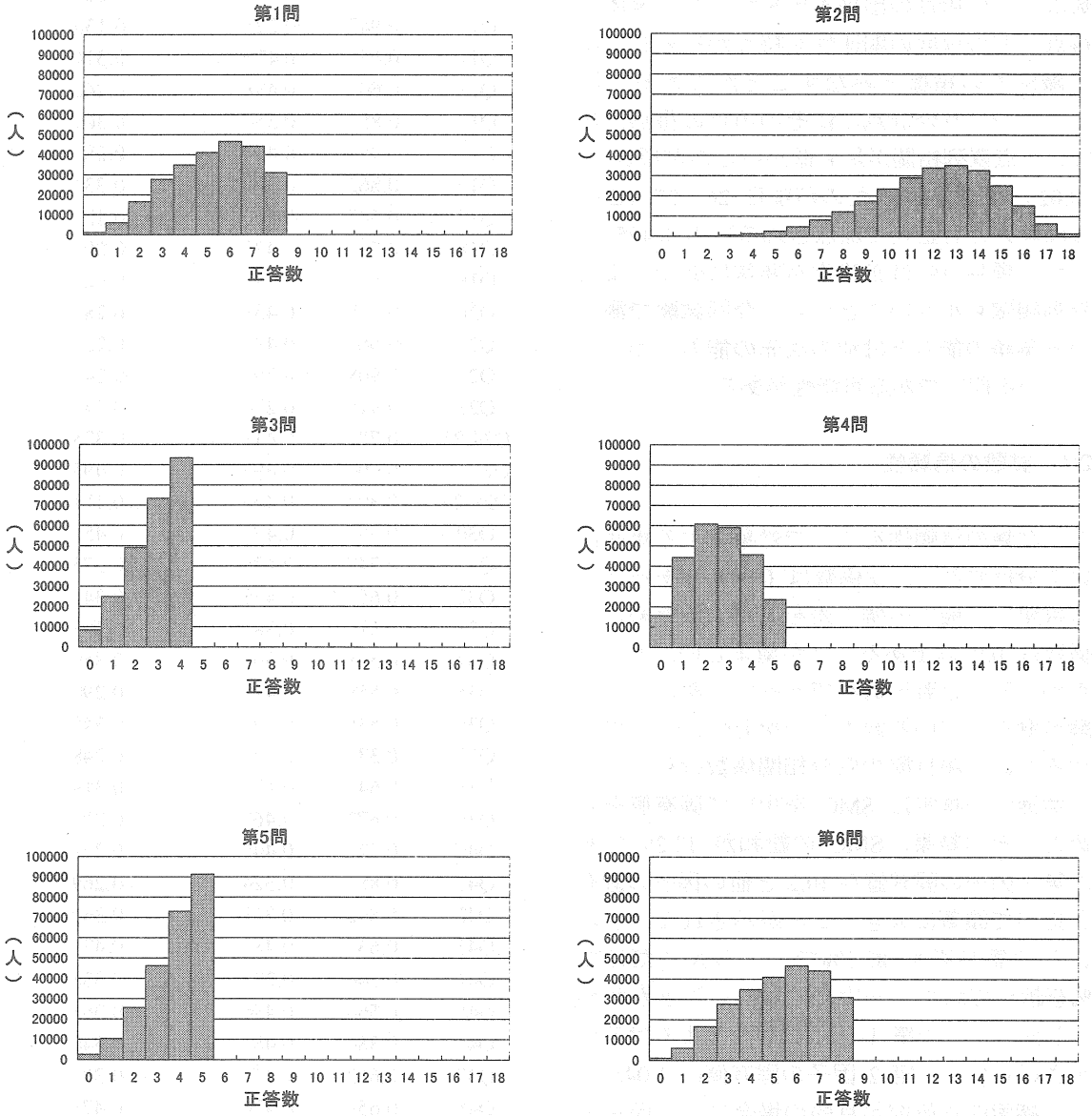

図 2.1 大問ごとの得点分布

2.3 項目単位での特徴

各項目ごとに正答率をみていく（表 2.2)。この正答率は，その項目がどの程度難しかったかを示す困難度の指標となる。正答率の最も低かったものは，項目 19 の 17%，最も高かったものは項目 8 の 96%であった。なお，項目 19 は語彙問題，項目 8 は強勢問題であった。次に各項目の成否と合計点との間の点双列相関係数（項目総点相関）をみていく。この点双列相関係数は，その項目の出来不出来がテスト全体での得点とどの程度の関連性を持つかを表し，項目の識別力の指標とみなすことができる。なお，ここでいう合計点にはその項目の得点は含めない。点双列相関が最も低かったのが項目 19 の 0.023，最も高かったのが項目 26，27（この 2 つは両方とも正答で加点される）の 0.495 であった。項目 19 は正答率が非常に低く，また点双列相関も小さいことから，今回試験で測っている英語の能力とは別の次元の能力について測っている問いである可能性がある。

2.4 試験の信頼性

この試験の信頼性をここで対象とする集団について分析すると，α係数は 0.844 であった。この数値は，能力の個人差を測定するためにはα係数は 0.9 以上あることが望ましい，という観点からみると若干低めであった。次に，この試験が測定している能力の一次元性について吟味するため，項目間の四分相関係数行列について，共通性の推定に SMC を用いて固有値をもとめた。その結果，SMC の総和が 12.29 であり，第 1 因子の固有値が 10.2 と他の因子の固有値と比べて顕著に大きいことが示された。さらに，その寄与率も 83.1%で，この因子だけで 80%を越えており，一次元性が高いことがわかる。しかし，この第 1 因子の固有値よりはるかに小さいものの，第 2 因子の固有値は 1.045 であり，探索的な研究が目的の場合には，因子として採用される可能性もある。(図 2.2)。

以上の信頼性係数及び因子分析の結果より，今回の試験は英語能力の複数の側面を測ってい

表 2.2 古典的項目パラメタ

項目	平均	標準偏差	項目総点相関 *
Q1	0.764	0.425	0.180
Q2	0.834	0.372	0.273
Q3	0.507	0.500	0.196
Q4	0.644	0.479	0.336
Q5	0.855	0.352	0.249
Q6	0.552	0.497	0.181
Q7	0.534	0.499	0.123
Q8	0.968	0.175	0.193
Q9	0.698	0.459	0.338
Q10	0.355	0.478	0.228
Q11	0.947	0.225	0.136
Q12	0.655	0.475	0.314
Q13	0.607	0.488	0.242
Q14	0.857	0.350	0.366
Q15	0.782	0.413	0.280
Q16	0.862	0.345	0.334
Q17	0.507	0.500	0.221
Q18	0.746	0.435	0.296
Q19	0.171	0.377	0.023
Q20	0.712	0.453	0.280
Q21	0.560	0.496	0.227
Q22	0.908	0.290	0.243
Q23	0.918	0.274	0.243
Q24,25	0.785	0.411	0.375
Q26,27	0.600	0.490	0.495
Q28,29	0.304	0.460	0.114
Q30	0.657	0.475	0.452
Q31	0.747	0.435	0.375
Q32	0.657	0.475	0.341
Q33	0.819	0.385	0.303
Q34	0.508	0.500	0.359
Q35	0.558	0.497	0.290
Q36	0.538	0.499	0.345
Q37	0.339	0.474	0.248
Q38	0.643	0.479	0.319
Q39	0.672	0.469	0.279
Q40	0.729	0.445	0.371
Q41	0.881	0.324	0.269
Q42	0.892	0.311	0.360
Q43	0.636	0.481	0.473
Q44	0.946	0.227	0.338
Q45	0.566	0.496	0.363
Q46	0.384	0.486	0.346
Q47	0.662	0.473	0.390
Q48	0.659	0.474	0.473
Q49	0.782	0.413	0.394
Q50	0.606	0.489	0.343
Q51	0.631	0.483	0.237

* ここでの総点は各該当項目の得点を除いてもとめたもの

る可能性も考えられるが，ある程度の一次元性は保証されたと見なし得るので，以下，一次元の能力を測定しているものとして試験の分析を行っていくことにする。

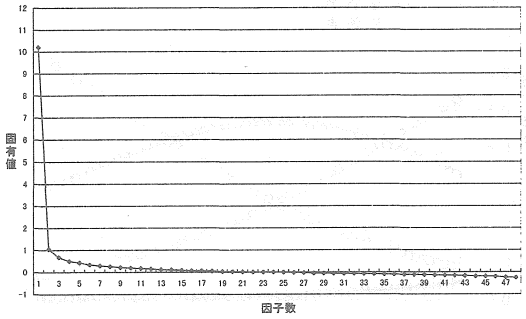

図 2.2　固有値の分布

3　テストレットモデルによる分析

3.1　テストレットの構成

上述したように，この試験は 6 つの大問より成り，各大問がさらに幾つかの中問に分かれるという構成になっている（試験の構成の詳細は付録 1 を参照)。ただ，大問として大きくまとまってはいても，例えば第 2 問 A に含まれる 11 の設問は相互に独立と考えてよく，このような設問は出題形式に捕らわれてテストレットとして括るよりも，独立な項目として扱うことによって IRT (Item Response Theory) により近い分析ができる。このように考えると，局所独立を仮定するのが無理と思われるのは，第 1 問 B に含まれる 2 つの設問，第 1 問 D に含まれる 3 つの設問，第 3 問 A に含まれる 2 つの設問，第 4 問に含まれる 5 つの設問，第 5 問に含まれる 5 つの設問，第 6 問に含まれる 8 つの設問ということになる。第 4 問，第 5 問，第 6 問はそれぞれ中問に分かれているが，いずれも共通の長文ないし会話文に付随する設問となっているため，中問同士が独立と考え難いので，大問ごとに 1 つのテストレットとした。こうして，それぞれ，2，3，2，5，5，8 の項目からなる 6 つのテストレットを構成し，残りの 23 の設問は独立な 2 値項目として扱った。なおテストレットの得点は含まれる設問に対する正答数とした。

こうして構成したテストの信頼性は，0.787 と，項目を独立に扱った場合に比べて低めになっている。テストレットの背後に想定するモデルとしては，Partial Credit Model (Masters, 1982), Nominal Response Model (Bock, 1972)等様々なモデルが考えうるが，ここでは，6 つのテストレットについてそれぞれ「含まれる項目数＋1」の得点段階を持つ Graded Response Model (Samejima, 1969, 1972) のロジスティク関数表現によるものを当てはめて分析した。このモデルは，正答－誤答という二つのカテゴリによって採点されるテストの統計的分析に利用される 2 パラメタ・ロジスティックモデル（2PL）(Lord & Novick, 1968) を，複数の得点段階を持つテストに拡張したもので，2PL を下位モデルとして含むものとなっている。(Muraki, 1990)

3.2　テストレットオペレーティング特性曲線 (Testlet Operating Characteristic Curve)

6 つのテストレットのオペレーティング特性曲線を図 3.1 に示す。$P_{jk}(\theta)$は特定の θ を持つ受験者が j 番目のテストレットで k 点を獲得する確率を示す。すなわち，j 番目のテストレットに含まれる項目数を m_j で表して

$$P_{jk}(\theta)=P^{+}_{jk}(\theta)-P^{+}_{j,k+1}(\theta)$$

ここで

$$P^{+}_{j0}(\theta)=1$$

$$P^{+}_{j,m_j+1}(\theta)=0$$

および

$$P^{+}_{jk}(\theta)=\frac{1}{1+\exp\left[-Da_j\left(\theta-b_{jk}\right)\right]}$$

ここで，a_j は項目識別力に相当するパラメタ，b_{jk} は j 番目のテストレットにおける k 番目のカテゴリの困難度を定めるパラメタである。また，D はロジスティック関数を累積正規曲線に近似させるための定数で D=1.7 である。

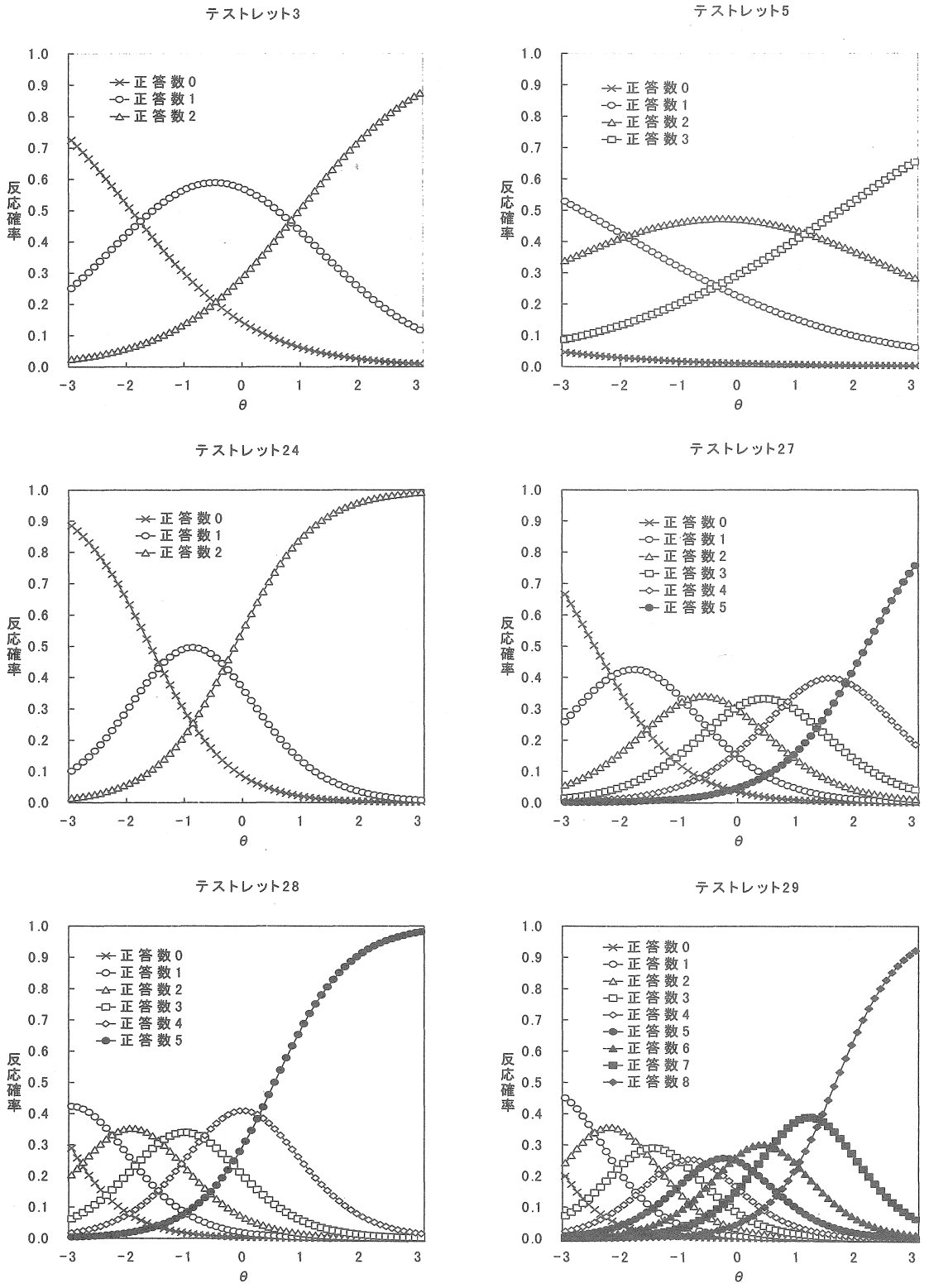

図 3.1 テストレットオペレーティング特性曲線

1 番目のテストレット（T3）はやや易しい設問（正答率 0.64）と中程度の困難度の設問（正答率 0.51）の組み合わせによって構成され，中程度の能力の者が，1 問のみに正答する様子が中央の山型の曲線に現れている。これに対して 2 番目のテストレット（T5）は極めて易しい設問（正答率 0.97），中程度の困難度の設問 2 問（正答率 0.55, 0.53）から成り，能力の低い受験者でも 0 点となる確率が極めて低くなっている。識別力が低い（項目総点相関 0.18, 0.12, 0.19）ので，能力の高い者でも 3 点を取る確率はあまり高くならず，能力の広い範囲に渡って 2 点の者が分布している。

3 番目のテストレット（T24）は識別力がある程度高く（項目総点相関 0.45, 0.38）易しめ（正答率 0.66, 0.75）の 2 問からなる。識別力がある程度高いため，能力が中程度からやや低いものでも，どちらか一方に正答する確率はあまり高くならず，能力の低いところでは両方に誤り，能力の高いところでは両方に正答する様子が現れている。

4 番目のテストレット（T27）はやや易しい設問（正答率 0.64），難しい設問（正答率 0.34），と中程度の設問 3 問（0.51, 0.56, 0.54）からなる。識別力がある程度高い（項目総点相関 0.36, 0.29, 0.35, 0.25, 0.32）ので，能力が高くなるのに応じて，1 点刻みで高くなる得点を取る確率が増えていく様子が見られる。

5 番目のテストレット（T28）は易しい設問 5 問（正答率 0.67, 0.73, 0.88, 0.89, 0.64）からなり，能力の低い者でも 0 点となる確率は低い。また，能力の中程度以上では満点となる確率が一番高い。

最後に 6 番目のテストレット（T29）は極めて易しい設問（正答率 0.95）と難しい設問（正答率 0.38）および易しい 6 つの設問（正答率 0.57, 0.66, 0.66, 0.78, 0.61, 0.63）から成る。易しい設問が含まれているので，能力の低い者でも 0 点の確率は低い。また，難しい設問が含まれているため，また，他のテストレットに比べて含まれる設問の数が多いため，能力がある程度高くならないと満点の確率は高くならないが，識別力がある程度高い（項目総点相関 0.34, 0.36, 0.35, 0.39, 0.47, 0.39, 0.34, 0.24）ので，能力が 1 を越える辺りから急速に満点の確率が高くなっている。

これらの作図に利用した項目・カテゴリパラメタを表 3.1 に示す。表 3.1 には，2 値項目として扱った 23 項目のパラメタも合わせて示した。最後の列「b の平均」はカテゴリ困難度の平均で，その項目全体の困難度を代表するものと考えることができる。なお，パラメタの推定には PARSCALE (Muraki & Bock, 1996) を用いた。

3.3 テストレット特性曲線 (Testlet Characteristic Curve)

テストレットオペレーティング特性曲線はテストレットの基本的な特性を表すものだが，一瞥のもとにその意味を読み取るのは容易ではない。そこで，

$$\overline{P}_{j.}(\theta)=\frac{1}{m_j}\sum_{k=0}^{m_j}P_{jk}(\theta)k$$

に当たるものを図 3.2 に示した。これをテストレット特性曲線（Testlet Characteristic Curve）と呼ぶ。これは θ を固定した場合の期待得点率曲線（Expected Score Curve of Testlet）に当たるものである。図 3.2 には同時に行った 2 パラメタ・ロジスティックテストモデル（2PL）による結果も合わせて示した。2PL において期待得点率に当たるものは，

$$\overline{P}_{j.}(\theta)\equiv\frac{1}{m_j}\sum_{g\in j}P_g(\theta)$$

ただし

$$P_g(\theta)=\frac{1}{1+\exp\left[-Da_g\left(\theta-b_g\right)\right]}$$

によって与えられる。

表 3.1 項目・カテゴリパラメタの推定値

項目番号	カテゴリ数	*a*	*b*1	*b*2	*b*3	*b*4	*b*5	*b*6	*b*7	*b*8	*b*の平均
T1	2	0.288	-2.517								-2.517
T2	2	0.523	-2.086								-2.086
T3	3	0.549	-1.963	0.930							-0.517
T4	2	0.498	-2.368								-2.368
T5	4	0.293	-9.080	-2.401	1.723						-3.253
T6	2	0.556	-1.053								-1.053
T7	2	0.348	1.090								1.090
T8	2	0.407	-4.452								-4.452
T9	2	0.494	-0.888								-0.888
T10	2	0.344	-0.804								-0.804
T11	2	0.866	-1.645								-1.645
T12	2	0.497	-1.726								-1.726
T13	2	0.780	-1.782								-1.782
T14	2	0.309	-0.061								-0.061
T15	2	0.497	-1.466								-1.466
T16	2	0.063	14.858								14.858
T17	2	0.449	-1.335								-1.335
T18	2	0.323	-0.473								-0.473
T19	2	0.620	-2.550								-2.550
T20	2	0.662	-2.561								-2.561
T21	2	0.751	-1.318								-1.318
T22	2	0.937	-0.387								-0.387
T23	2	0.170	2.915								2.915
T24	3	0.880	-1.637	-0.184							-0.911
T25	2	0.555	-0.822								-0.822
T26	2	0.594	-1.776								-1.776
T27	6	0.798	-2.491	-1.152	-0.103	0.917	2.160				-0.134
T28	6	0.931	-3.568	-2.428	-1.494	-0.599	0.501				-1.518
T29	9	1.056	-3.774	-2.656	-1.822	-1.155	-0.575	0.013	0.708	1.625	-0.954

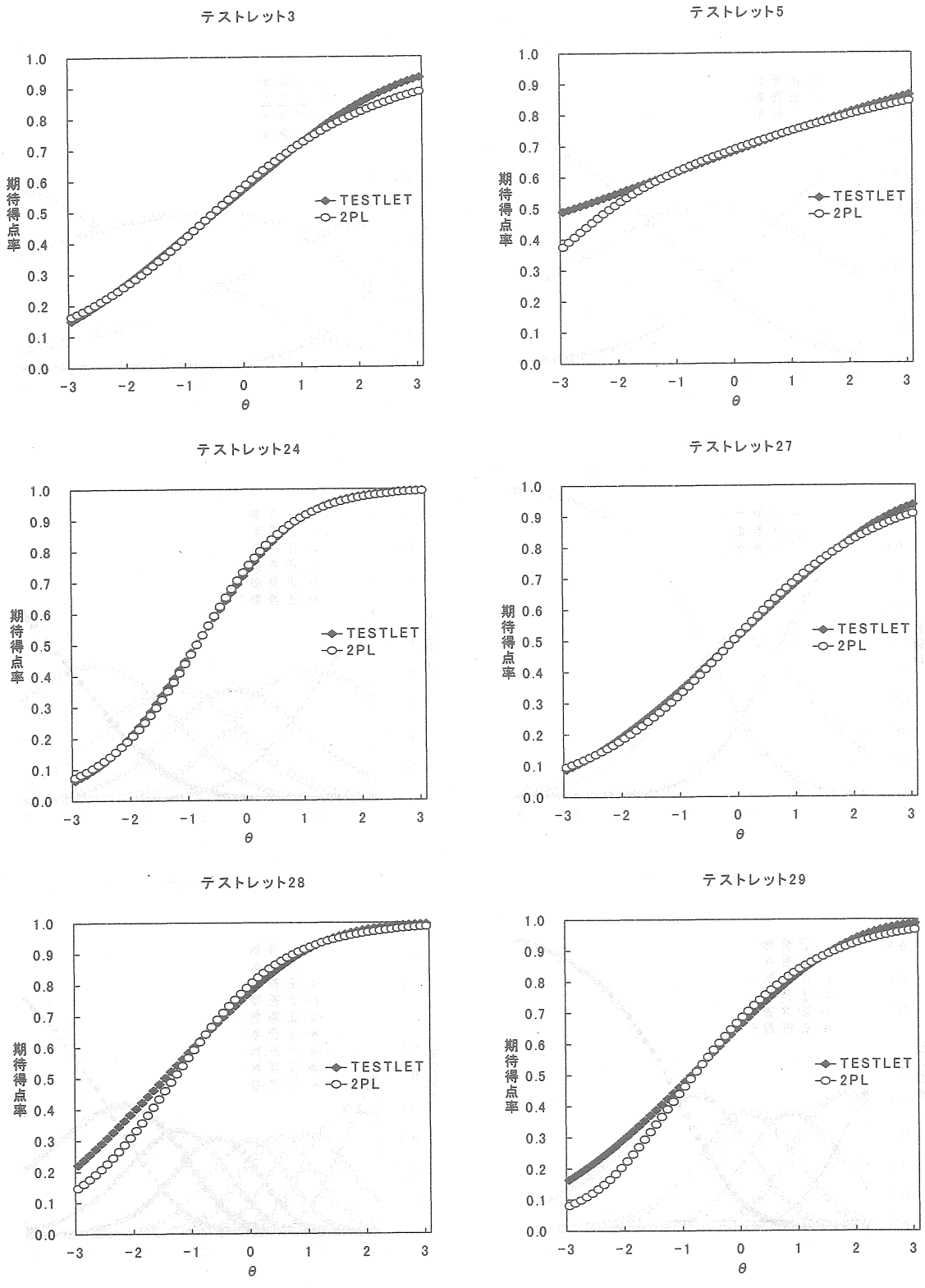

図 3.2　テストレット特性曲線（期待得点率）

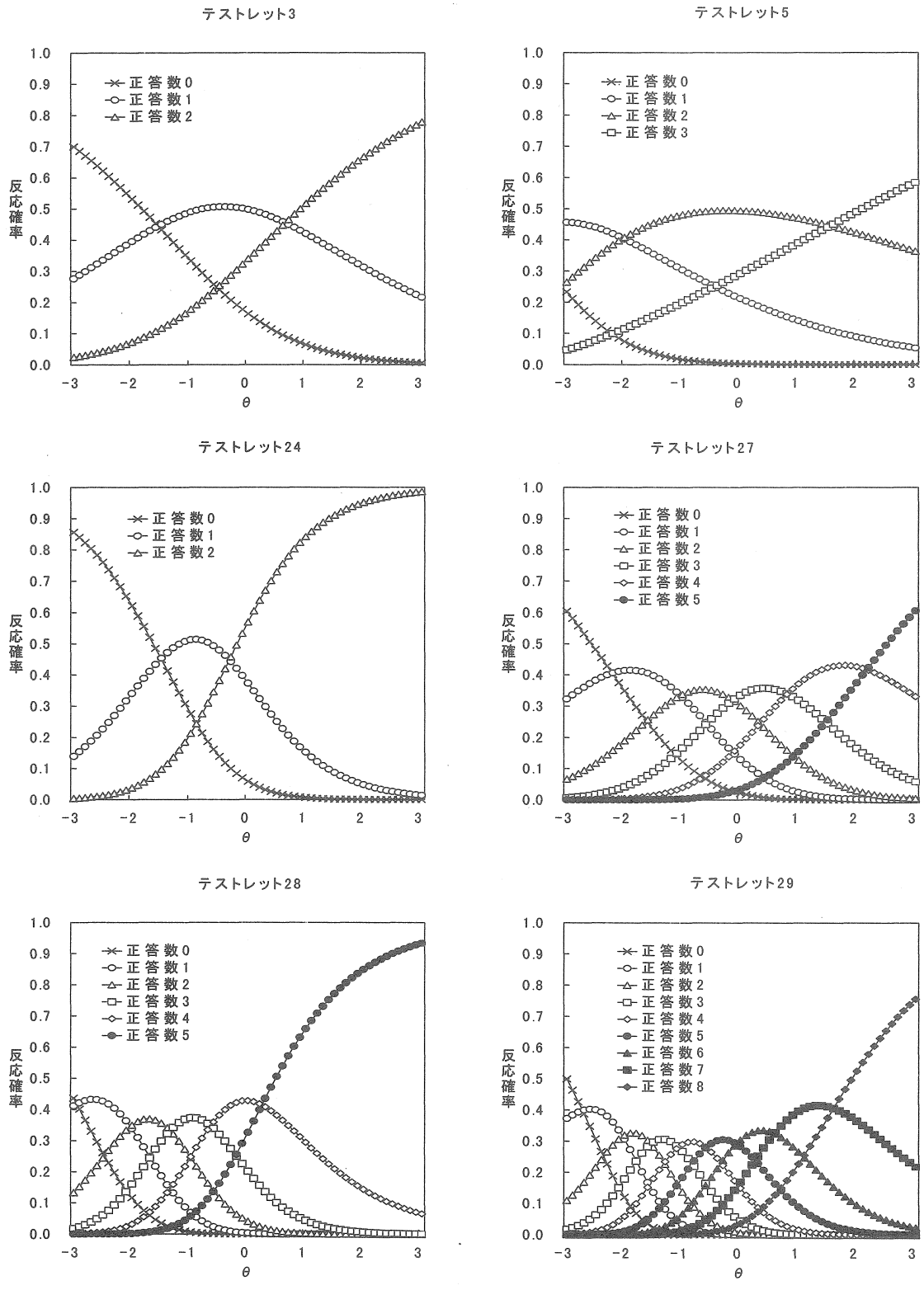

図 3.3 2PLによるテストレットオペレーティング特性曲線

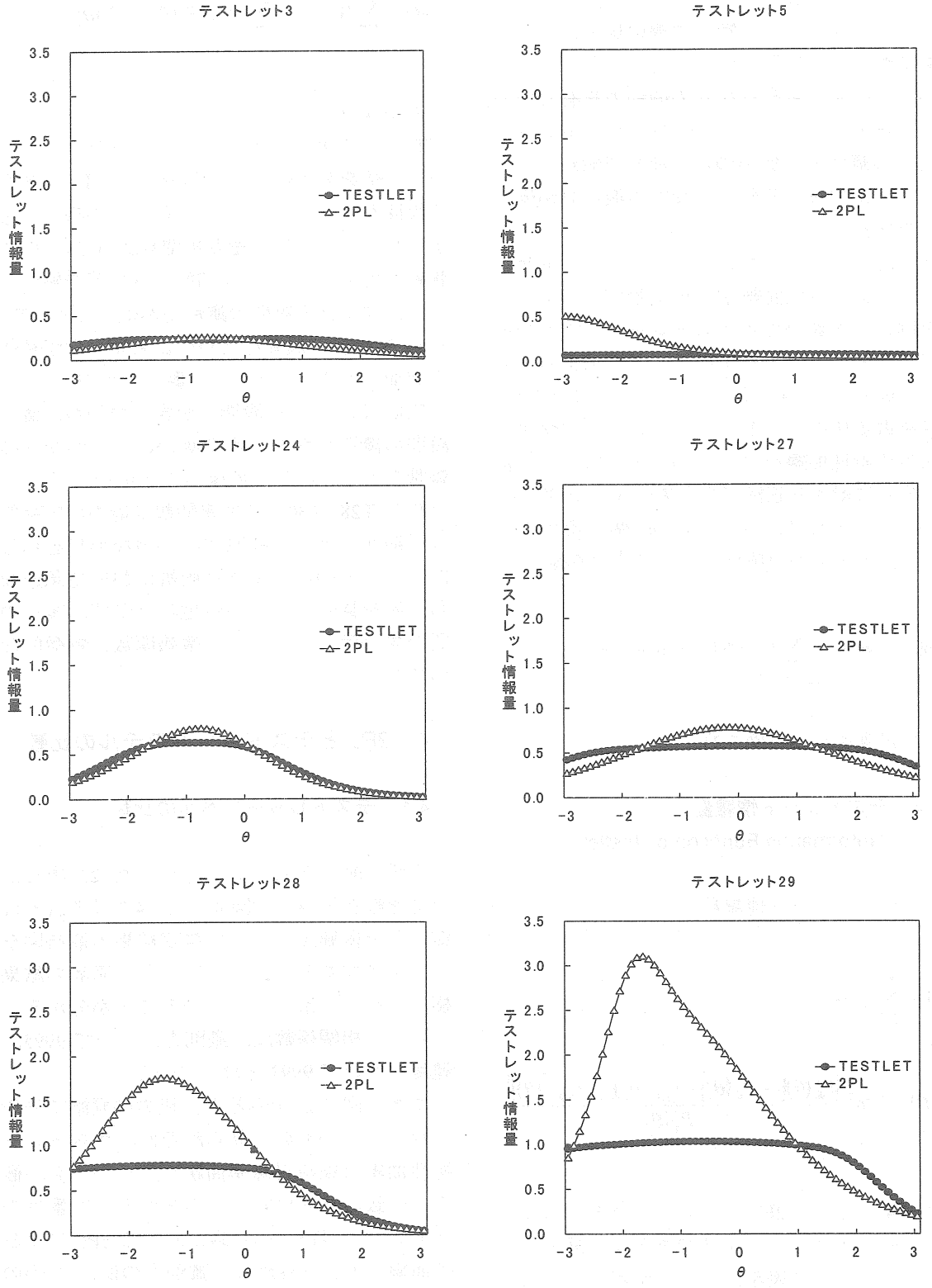

図 3.4 テストレット情報関数

2PL との比較については次章で述べるので，ここではテストレット特性曲線の様子について簡単に触れる。2 番目のテストレット（T5）を除いて，いずれもそれなりの識別力を有していることが分かる。また，2 番目のテストレット（T5）の識別力が低いのは，前節で触れたように，それに含まれる個々の設問の識別力が低いためである。

ところで，2PL において，テストレットオペレーティング特性曲線 $P_{jk}(\theta)$に対応するものは θを固定した場合の各テストレットの得点分布を計算することにより求めることが出来る。例えば，m_j の項目からなるテストレットにおいて k 点を取る確率は，そのテストレットに含まれる項目の特性曲線を $P_g(\theta)$, $g \in j$, これらの m_j 個の項目に対する正誤反応を表す 0-1 の並びを u_g, $g \in j$ とすれば，これらの m_j 個の項目のうち，ちょうど k 個の項目にだけ正答する確率

$$\mathrm{Prob}(X_j = k|\theta) = \sum_{\left(\sum_{g\in j} u_g\right)=k} P_g(\theta)^{u_g}\left(1-P_g(\theta)\right)^{1-u_g}, k = 0,1,\ldots,m_j$$

として計算される（図 3.3）。

3.4 テストレット情報量 (Information Function of Testlet)

次のテストレット情報量，

$$I_j(\theta) = \sum_{k=0}^{m_j} A_{jk}(\theta)$$

ただし

$$A_{jk}(\theta) = D^2 a_j^2 \frac{\left\{P_{jk}^+(\theta)\left[1-P_{jk}^+(\theta)\right] - P_{j,k+1}^+(\theta)\left[1-P_{j,k+1}^+(\theta)\right]\right\}^2}{P_{jk}(\theta)}$$

を図 3.4 に示す。2PL でこれに当たるものは上式で $m_j = 1$ とおいたものであるが，通常はそれを以下のように表現したものが用いられることが多い。

$$I_j(\theta) = \sum_{g\in j} I(\theta, u_g) = \sum_{g\in j} D^2 a_g^2 P_j(\theta)\left[1-P_j(\theta)\right]$$

図 3.4 を見ると，1 番目のテストレット（T3）と 2 番目のテストレット（T5）のテストレット情報量の低いことが分かる。T3 は含まれる設問が 2 つであるうえ，それらの識別力もあまり高くない（項目総点相関 0.20, 0.34）ので情報量も大きくならない。T5 は 3.2 節で触れたように，含まれる設問の識別力が低いため，T3 よりも 1 つ多い設問を含むのにも関わらず貧弱な情報量しか持たない，3 番目のテストレット（T24）は，2 つの設問しか含まないが，個々の設問の識別力がある程度高いので，かなりの情報量を示している。最後の 3 つのテストレット（T27, T28, T30）の情報関数は能力の中央部の広い範囲に渉って高原状の平坦な形状を示し，能力の一方の端，または両端において急速な落ち込みを見せる。この台地状の形状は多くの設問を含むテストレットの情報関数に特徴的なものである。

4 2PL とテストレットモデルの比較

4.1 テストレットレベルの比較

まず，独立項目として分析した 23 項目について比較を試みる。図 4.1，図 4.2 はそれぞれ，識別力と困難度について推定結果の散布図を描いたものである。2 つの図から，両者の結果が極めて高い一致を示していることが分かる。ちなみに，相関係数は，識別力について.9995，困難度について.9993 となっている。

また，図 3.2 をみると，能力の両端でやや不一致の大きいものも見られるが，テストレット特性曲線（期待得点率曲線）についても，概ね良い一致の見られることが分かる。2 番目のテストレット（T5）の左端における 2PL による特性曲線の大きな外れは，識別力の低い 2 つの設問と共に含まれる，極めて易しい設問の影響と考えられる。6 番目のテストレット（T29）の左端の外れも，それに含まれる極めて易しい設問の影響と考えられる。この場合は同時に含まれ

る他の設問の識別力がある程度高いので，T5 のような奇妙な形状にはなっていない。同じことは，5 番目のテストレット（T28）の左端の外れについても言える。このテストレットにも正答率 0.88 と 0.89 というかなり易しい設問が含まれている。このように一部の外れを除けば，他の部分では概ね良い一致が見られ，また，図 3.1 と図 3.3 を比較してみても，テストレットモデルに基づくテストレットオペレーティング特性曲線と 2PL に基づくそれとの間に系統的な不一致はみられない。

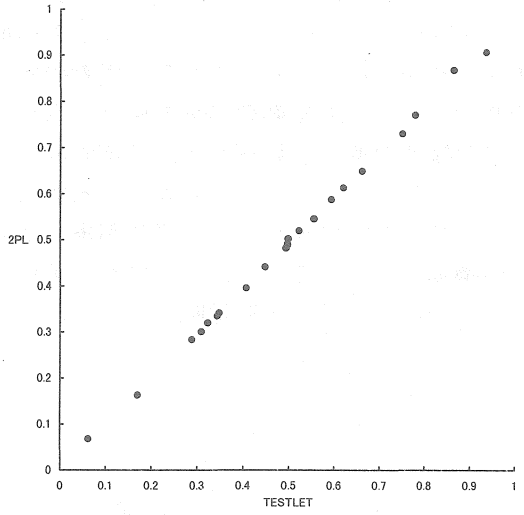

図 4.1 識別力パラメタの散布図

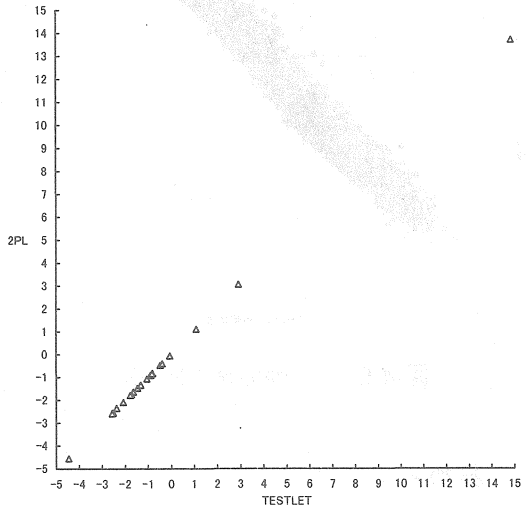

図 4.2 困難度パラメタの散布図

一方，テストレット情報曲線の形状は大きく異なる。テストレットモデルの情報曲線は概ね平坦な形状を示し，1 つ目のテストレットを除き，2PL の情報曲線がテストレットモデルの情報曲線を上回る傾向を示している。2 番目のテストレット（T3）における左端の外れは，テストレット特性曲線の外れと同様，識別力の低い 2 つの設問と共に含まれる極めて易しい設問の影響と考えられる。

項目パラメタの推定の精度を比較してみると，識別力については，テストレットモデルが 29 の独立項目及びテストレットの標準誤差の平均が 0.0037 であるのに対して，2PL では 48 項目の標準誤差の平均が 0.0042 とテストレットモデルの精度が高い。困難度は，テストレットモデルの困難度がテストレットとカテゴリに分解して推定されるので単純な比較は困難だが，2PL の結果をみると，識別力よりも 1 桁程度精度が落ちるようである。いずれにしても，25 万人弱の受験者による推定なので，精度の高い推定値が得られたと考えてよいだろう。

4.2 テストレベルの比較

図 4.3 にテスト特性曲線を示す。テストレットレベルの期待得点率における傾向がそうであったように，テスト特性曲線も 2 つのモデルの間で概ね一致した傾向を示す。

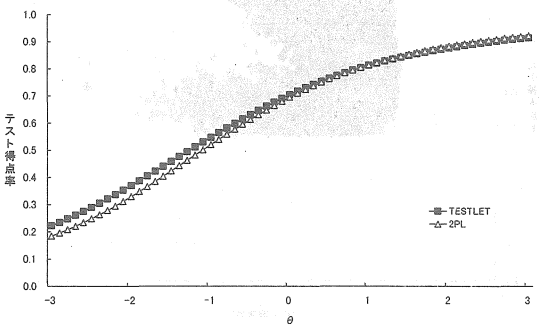

図 4.3 テスト特性曲線

これに対してテスト情報曲線（図 4.4）は，2PL のものが尖りが大きく，概ねテストレットモデルの情報量を上回り，Sireci et al. (1991) や

Lee (2000) の知見を裏付ける結果となっている。

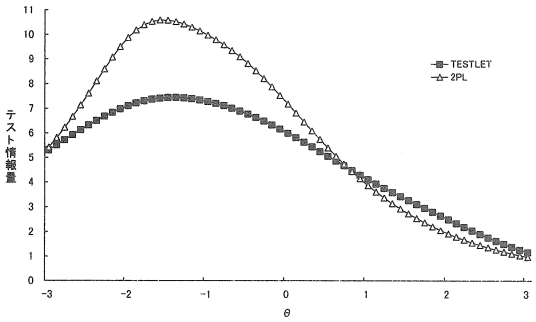

図 4.4 テスト情報関数

また，能力推定の標準誤差を 249,256 人について平均してみると，テストレットモデル 0.43 に対して 2PL 0.36 となっており，情報関数の知見と一致する。ところで，この 2 つの標準誤差の間の相関は 0.397 で，推定モデルの如何を問わず標準誤差が大きくなり易かったり，小さくなり易かったりする解答パタンの存在が示唆される。図 4.5 はこれを散布図に描いたものである。

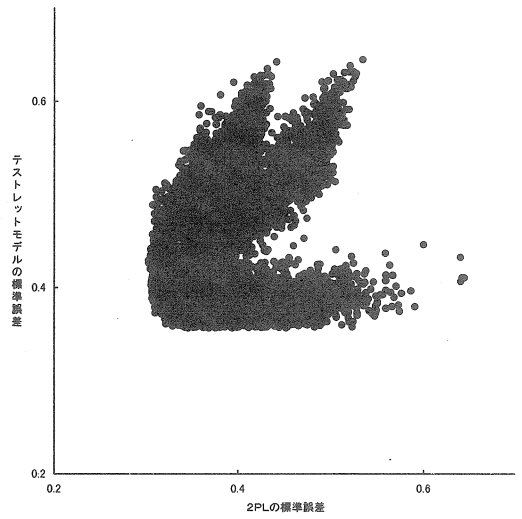

図 4.5 能力推定標準誤差の散布図

散布の形状はおよそ 3 肢に分かれ，下辺に横たわる肢はテストレットモデルの誤差は小さいが，2PL の誤差が大きくなるような解答パタンである。今回は個々のパタンの検討にまで立ち入っていないが，恐らく，これに属するパタンは第 4 問，第 5 問，第 6 問の辺りで希少な動きを示しているものと推測される。残る 2 肢は，両者の大小が連動しているものだが，中でも左辺に立ちあがる 1 肢は 2PL の誤差よりも，テストレットモデルの誤差が大きくなるパタンである。これは，1 番目の肢とは逆に第 4 問，第 5 問，第 6 問の辺りで標準的な動きを示し，残る 23 個の独立項目の一部に希少な動きの見られるものではないかと推測される。

最後に能力推定値の散布図について考察を加える。図 4.6 は 2 つのモデルに基づく能力の推定値を散布図に描いたものである。両者の相関は 0.987 と高い。推定誤差の散布図に見られるような組織的な偏りも見られず，ほぼ直線状に散布している。このことは，両者による能力の推定の一致度がかなり高く，いずれが優れているかの議論は別にして，共にそれなりの実用性を備えたものであることを意味しているといえよう。

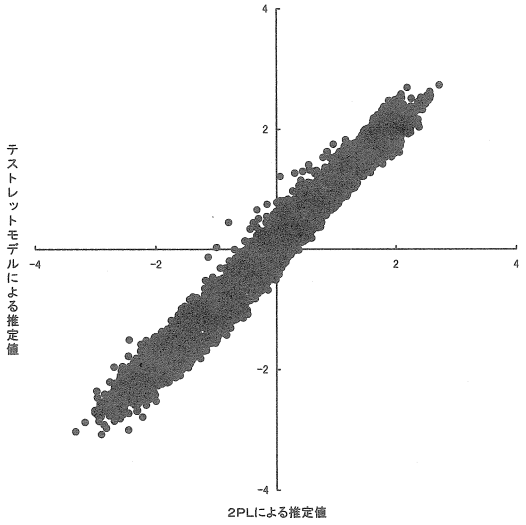

図 4.6 能力推定値の散布図

5 おわりに

2000 年度に実施された大学入試センター試験・英語の試験結果についてテストレットモデルと 2PL によって分析を加え，両者の結果を比

較してきた。全体の知見をまとめると，能力推定の標準誤差については，両者の違いが大きく，Sireci et al. (1991)やLee (2000)の知見を支持する結果となったが，項目パラメタや能力値の推定には顕著な偏りが見られず，両者に大きな違いが見られなかった。このことから，推定精度の問題を除けば，いずれのモデルを用いても実用上大きな問題とならないと考えてよいことが示唆される。

ただ，今回の分析は単年度の，しかも1つの科目の試験について行ったものに過ぎないので，ここでの知見を一般化するためには，他の年度について，また，他の科目についてさらなる分析を加える必要があるのはいうまでもないであろう。

引用文献

Bock, R. Darrell, Estimating Item Parameters and Latent Ability When Responses are Scored in Two or More Nominal Categories, *Psychometrika*, 37, 1, 29-51, 1972.

Lee, Guemin, Estimating Conditional Standard Errors of Measurement for Tests Composed of Testlets. *Applied Measurement in Education*, 13. 2, 161-180, 2000.

Masters, Geoff A. Rasch Model for Partial Credit Scoring, *Psychometrika*, 47, 2, 149-174, 1982.

Muraki, Eiji, Fitting a Polytomous item response model to Likert-type data. *Applied Psychological Measurement*, 14, 59-71, 1990.

Muraki, Eiji & Bock, R. Darrel, *PARSCALR™ IRT based Test Scoring and Item Analysis for Graded Open-ended Exercises and Performance Tasks*, Version 3, Scientific Software International, IL, 1996

Sireci, Stephen G., Thissen, David, & Wainer, Howard, On the Reliability of Testlet-Based Tests, *Journal of Educational Measurement*, 28, 3, 185-201, 1991.

Samejima, Fumiko, Estimation of Latent Trait Ability Using a Response Pattern of Graded Scores, *Psychometric Monograph*, 17, 1969.

Samejima, Fumiko, A General Model for Free-response Data, *Psychometric Monograph*, 18, 1972.

Wainer, Howard, & Kiely, Gerard L., Item Clusters and Computerized Adaptive Testing: A Case of Testlets, *Journal of Educational Measurement*, 24, 3, 185-201, 1987.

南風原朝和 個人正答確率に基づく局所独立性の概念の明確化 －実験的独立性および一次元性との関連を中心に－（未出版）, 2000

付　録

付録1. 試験の構成

問題番号	設問		解答番号
第1問 (発音・強勢・会話問題)	A	1	1
		2	2
	B		3-4
	C	1	5
	D	1	6
		2	7
		3	8
第2問 (語彙・語法・分布に関する問題および整序作文)	A	1	9
		2	10
		3	11
		4	12
		5	13
		6	14
		7	15
		8	16
		9	17
		10	18
		11	19
	B	1	20
		2	21
		3	22
		4	23
	C	1	24
			25
		2	26
			27
		3	28
			29
第3問 (つなぎ語補充・文整序問題)	A		30
			31
	B		32
	C		33
第4問 (図表読み取り問題)	A	1	34
		2	35
		3	36
	B		37
			38

問題番号	設問		解答番号
第5問 (口語コミュニケーション問題)	A		39
	B		40
			41
	C		42
	D		43
第6問 (長文読解問題)	A	1	44
		2	45
		2	46
		4	47
		5	48
	B		49-50-51

付録 2. テストレット・項目対応表

問題	テストレット	項目
第1問	T1	Q1
	T2	Q2
	T3	Q3-Q4
	T4	Q5
	T5	Q6-Q8
第2問	T6	Q9
	T7	Q10
	T8	Q11
	T9	Q12
	T10	Q13
	T11	Q14
	T12	Q15
	T13	Q16
	T14	Q17
	T15	Q18
	T16	Q19
	T17	Q20
	T18	Q21
	T19	Q22
	T20	Q23
	T21	Q25
	T22	Q27
	T23	Q29
第3問	T24	Q30-Q31
	T25	Q32
	T26	Q33
第4問	T27	Q34-Q38
第5問	T28	Q39-Q43
第6問	T29	Q44-Q51

Testlet Analysis of the English Test Scores of the National Center Test

Tomoichi Ishizuka*
Naoko Nakaune*
Teruhisa Uchida**
Shin-ichi Mayekawa***

Abstract

A testlet response model and 2 parameter logistic test model (2PL) were applied on the test scores of English in the battery of National Center Test administered in the year 2000. The adopted testlet model was based on the graded response model of Samejima (1969, 1972). We chose 249,256 examinee out of 528,812 total applicants in order to make our result comparable to the other studies.

We compared testlet operating characteristic curves, testlet characteristic curves, test characteristic curves, and item parameters of 23 independent dichotomous items derived under both testlet and 2PL models, and the result showed a fairly satisfying agreement. However, the testlet information function differed largely. The amount of information of 2PL exceeded the amount of information of testlet model almost everywhere. The arithmetic average of the standard errors of estimation of ability was also smaller in 2PL than in the testlet model.

Key Words : Item Response Theory (IRT), Testlet, Logistic Test Model, Graded Response Model, National Center Test, English Test

* Evaluation and Follow-up Study Section
** Special Examination Section
*** Model Based Measurement Section