

UDC 371. 279. 6

昭和62年度共通第1次学力試験「理科」の各科目に関する 項目反応理論に基づくテスト特性の比較

鈴木規夫*

前川眞一**

目 次

1. 研究の目的と方法	7.3.1 合成テストに関する項目パラメタと学力の分布
2. 項目反応理論	7.3.2 項目特性曲線
3. データ	7.3.3 テスト特性曲線
4. 項目パラメタの推定	7.3.4 テスト情報関数と項目情報関数
5. 項目の選別 I	7.4 前年度の「理科」各科目のテスト特性
6. 1次元性の検討と項目の選別 II	8. 討議とまとめ
7. テストの等化	8.1 学力について
7.1 テストの等化の方法	8.2 テスト特性の比較について
7.2 単独テスト間の1次元性の確認	
7.3 等化後の各科目のテスト特性	

1. 研究の目的と方法

本研究は、昭和62年度の共通第1次学力試験の「理科」の各科目の項目データを利用し、項目反応理論に基づき各科目のテスト特性について比較分析を行うと共に、テストやそれを構成する項目の評価に関して1つの手続きを提示することを目的としている。

「理科」については、本年度から従来の2科目選択制から1科目選択制となった。どの科目を選択するかは、試験当日の受験者の判断に委ねられているが、選択による有利・不利をなくすためには、どの科目を選択しても難易度に差異が生じないことが望まれる。しかし、実際には科目間に難易差がないよう問

題を作成することは大変難しく、たとえ、問題作成者が努力しても、おのずとそれには限界がある。そのため、受験者の科目得点と(潜在)学力の関係には、科目によって差異が生じていると推測される。従って、各科目の受験者から学力の等しい者を選び出すには、この差異を除去しなければならない。この問題については、塗師(1982)や清水(1983)によって共通第1次学力試験の「社会」および「理科」に関する分析結果が報告されている。項目反応理論を応用したいくつかの等化法(Haebara, 1980, Stocking & Lord, 1983, 芝, 1978 野口, 1983)も提案されているが、本研究では、項目特性曲線(ICC)を利用したテストの等化法(前川・鈴木, 1988)を利用して「理科」の各科目のテスト結果を同一尺

* 大学入試センター研究開発部情報処理研究部門

** 大学入試センター研究開発部進学適性研究部門

大学入試センター研究紀要 No. 17, 1988, p. 219~247 (昭和62年11月24日受付)

©1988 THE NATIONAL CENTER FOR UNIVERSITY ENTRANCE EXAMINATION

度上に表し、その結果によって比較を行う。

ところで、テストの等化を行うためには、①等化の対象となるテストを同一受験者集団が受験するか (Single-group Design)、②同一受験者集団ではないがそれと同等とみなせる集団が受験するか (Equivalent-group Design)、あるいは、③受験者集団は異なるが、対象となるテストの中に共通な項目が含まれているか (Anchor-test Design)、のいずれかの条件が必要である (Hambleton & Swaminathan, 1985)。ところが、「理科」の各科目は選択制のため、選択集団は異なり、それぞれの集団が同等とみなすことに無理がある。また、本年度は1科目選択制のため、そのままの形では③のアンカーテスト法を使用することができない。従って、本研究では、数学 I を共通項目として③の方法によって「理科」の科目間の等化を試みる。

2. 項目反応理論

一般に、テストを実施したとき、項目に正答したときに 1 を与え、誤答したときに 0 を与えるならば、そのテストに対する反応パターンは 0 と 1 の組み合わせによって表される。そのとき、ある項目に対し、学力の低い者は正答する確率は低く、学力が高くなるにつれ正答する確率は高くなり、やがて 1 に近づくと考えるのは自然である。このような考え方に立ったテスト得点のモデルに項目反応理論 (Lord & Novick, 1968, Lord, 1980) がある。項目反応モデルは、ある潜在特性 (latent trait) または学力 (ability) をもつ被験者がある項目に対し、どの程度の確率で正答するかを関数で近似したものである。近似する関数としては、階段関数、正規分布関数、ロジスティック関数等があるが、本研究では項目反応モデルとして、2パラメタ・ロジスティ

ク・モデル式

$$P_i(\theta) = \frac{1}{1 + \exp\{-1.7a_i(\theta - b_i)\}}$$

を採用する。

図 1 は、ロジスティック・モデルによる項目特性曲線を 2 つの項目について描いたものである。

モデルによれば、学力 θ をもった者の項目 i に対する正答確率は θ の関数 $P_i(\theta)$ で与えられ、各項目は 2 つの項目パラメタ a_i, b_i によって特性化される。項目パラメタ a_i は項目識別力と呼ばれ、項目パラメタ b_i は項目困難度と呼ばれている。図 1 の 2 つの項目特性曲線をみると、項目 B は、項目 A より項目困難度が大きい。すなわち、項目 B の方が項目 A より難しい項目である。また、項目識別力をみると、項目 A の方が項目 B より大きい。すなわち、項目 A の方が項目 B よりも大きな識別力をもつ。

モデルは、次の特徴をもっている。①推定された項目パラメタの値は集団の性質によらず不変であり、②推定された学力は特定の選択された項目によらない。すなわち、どのような集団から項目パラメタを推定しても、推

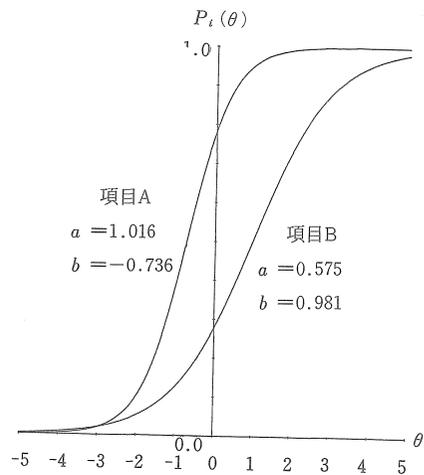


図1. 項目特性曲線

定誤差を除けば同じ結果が得られ、項目プールからどのような項目が選ばれたとしても、推定誤差を除けば同一被験者からは本質的に同じ学力の推定値が得られるということを示している。ただ、項目反応論においては学力を表す尺度の原点と単位は任意であるため、異なる集団を用いて推定された項目パラメタの値を比較するためには第7節で述べる等化の手続きが必要となる。

このような特徴をもつために、いくつかの仮定が設けられている。その1つに学力 θ の1次元性の仮定がある。もし、テストが1次元の学力を測定しているのであれば、テストを構成している項目の内容は一貫しており、どの項目からも項目反応曲線で示したような反応を得ることができる。もちろん、テストの結果に与える要因はこの1つの学力に限らず他の別の学力やテストを受けたときの状況（動機、不安、テスト慣れ）等が微妙に関係してきている。しかし、その学力のテストに与える影響が支配的と考え、主として1つの学力によってテストの結果は決定されると考えることにする。この点については、モデルを適用する前に因子分析によって項目の選択を行い、内部一貫性の高い項目からなるテストについて分析を行うといった方法が用いられることが多いが、ここでは、項目パラメタや学力推定値の不変性の性質を利用して1次元性の確認を行う。

また、もう1つの仮定として、局所独立の仮定がある。これは、同一の学力をもつ者に対して、異なる項目に正答する確率は互いに独立であるとする仮定である。この仮定は、潜在特性空間が完全に1次元であるという仮定と等価である (Lord & Novick, 1968)。本研究では、この局所独立の仮定に関する確認は行わずに分析した。

3. データ

分析の対象としたデータは、昭和62年度共通第1次学力試験の「理科」に含まれる「物理」、「化学」、「生物」、「地学」の各科目および共通項目として利用する「数学I」の単独テストの項目データである。また、テストの等化を行うため、数学Iを共通項目とした「数学I—物理」、「数学I—化学」、「数学I—生物」、「数学I—地学」の合成テストの項目データについても分析の対象とする。単独テストあるいは合成テストは、いくつかの項目から構成されているが、これらの項目に対する受験者の反応を、完全正解なら1、それ以外では0というように、数値化した項目データを用いる。各単独テストおよび合成テストの項目数を表1の左欄に示す。また、分析に用いた項目の項目番号と実際の解答において利用した解答番号との対応を表2に示す。

分析は、届出選択科目を受験した者、英語以外の外国語を受験した者および一部の教科に欠席した者を除いた、全ての教科を本試験で受験した者を対象とし、その中から、「理科」の各科目を受験した者についてそれぞれ満点を除き5,000人を抽出し、これらの集団を対象とした。ちなみに、物理の受験者は、

表1. 数学Iおよび理科テストの項目数

テスト名	項目数	選別後の項目数	除外した項目番号
物理	23	12	1, 2, 5, 11, 12, 14, 15, 18, 19, 22, 23
化学	31	27	11, 12, 14, 15
生物	33	27	10, 17, 20, 25, 27, 31
地学	34	26	1, 6, 7, 10, 25, 26, 27, 28
数学I	17	13	11, 13, 15, 16
数学I—物理	40	25	
数学I—化学	48	40	
数学I—生物	50	40	
数学I—地学	51	39	

105,074人, 化学は112,960人, 生物は109,102人, 地学は20,013人であった。

表2. 項目一覧

項目番号	解 答 番 号				
	物 理	化 学	生 物	地 学	数 学 I
1	1	1	1	1	1ーア
2	2	2	2	2	1ーイウ
3	3	3	3	3	1ーエオカ
4	4	4	4	4	1ーキ
5	5	5	5	5	1(クーン)
6	6	6-7	6	6	2ーア
7	7	8	7	7	2ーイ
8	8	9	8	8	2ーウエ
9	9	10	9	9	2ーオカ
10	10	11	10	10	2ーキク
11	11	12	11	11	2ーケコ
12	12	13	12	12	2ーサシ
13	13	14	13-14	13	2ースセソ
14	14	15	15	14	3ーアイ
15	15	16	16	15	3ーウエオ
16	16	17	17	16	3ーカキク
17	17	18	18	17	3ーケコサ
18	18	19	19	18	
19	19	20	20	19	
20	20	21	21	20	
21	21	22	22	21	
22	22	23	23-24	22	
23	23	24	25-26	23	
24		25	27-28	24	
25		26	29	25	
26		27	30	26	
27		28	31	27	
28		29	32	28	
29		30	33	29	
30		31	34-35	30	
31		32	36	31	
32			37	32	
33			38	33	
34				34	

4. 項目パラメタの推定

各科目の項目パラメタおよび共通項目を加えた合成テストの項目パラメタを「GETAB」(Baillie, 1980)によって推定した。図2は

その結果を平行箱形図 (Tukey, 1977) によって示したものである。図は, 各科目単独に推定された項目パラメタの分布を比較したものと, 合成テストによって推定された項目パラメタの分布の比較をしたものとを別々に示している。

分布をみると, 項目識別力の中央値は, 物理で大きく, 分布も拡がりを見せており, 他の3科目と異質な傾向を示している。しかし, 共通項目として数学 I を加えた場合, 数学 I の影響で各科目共中央値は大きくなる傾向を示し, 分布も類似した傾向を示すようになる。これは, 数学 I の項目の項目識別力が全般に大きく, このことが影響したものと考えられる。物理と数学 I の項目識別力は他の科目に比べ項目識別力の大きい項目から構成されていることが分かる。

これに対し, 項目困難度の分布をみると, 中央値は科目間の差異は小さく, 分布も類似した傾向を示している。共通項目を加えた場合も同様な傾向を示している。

5. 項目の選別 I

さて, モデルを適用する際には, テストの対象となる項目データがそのモデルを十分説明しているか, モデルへの適合性について検討を加える必要がある。ここでは, 等人数 (500人) になるよう学力 θ を 10 の階級に分けたときのそれぞれの階級の正答率と各階級の中央の値に対応するモデルの期待値との適合度を χ^2 統計量によって検定 ($df = 9, \alpha = 0.005$) する方法をとった。検定の結果有意となった項目のうち特に適合度の悪い項目を分析の対象から外すことにする。検定は, 項目パラメタが各科目から推定された場合と, 合成テストから推定された場合のいずれの場合についても行い, いずれかで適合度の悪かった項目

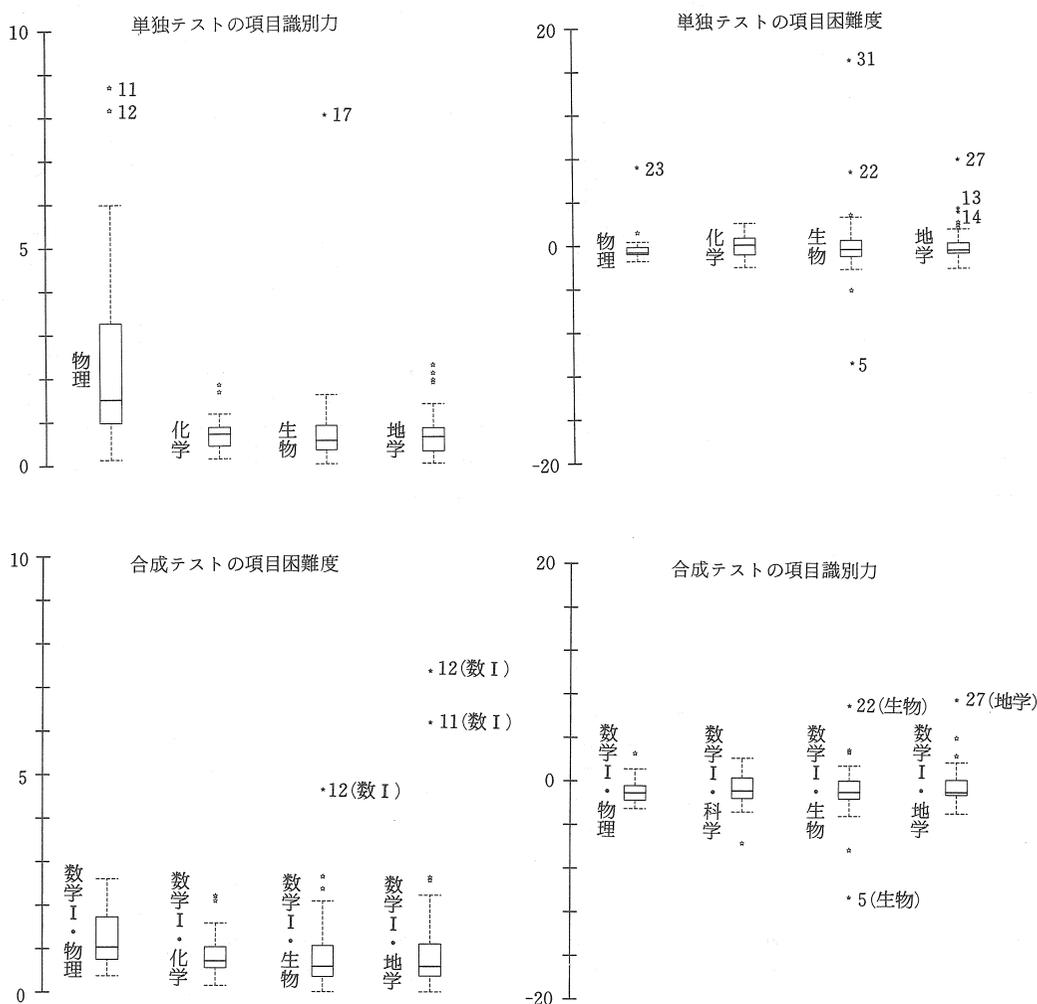


図2. 理科各科目の項目パラメタの分布

を除外する。表1の右側欄は、このようにして除外された項目を除いた項目数と除外された項目番号(下線が引いてある)を示している。なお下線の引いてない項目番号は1次元性の確認の際に除外されたもので、このことについては後述する。

除外された項目は、物理で2, 化学で2, 生物で5, 地学で4, 数学Iで4つであった。表3は、これらの除外された項目について、各階級毎の正答率と対応するモデルの期待値

とを比較したものである。

これらの除外された項目について、昭和62年度共通第1次学力試験に関する意見・評価(大学入試センター, 1987)に基づいて分析する。その結果は以下の通りである。

物理で適合度の悪い項目は、項目22と23(第4問, 問2)であった。これらの項目は、原子核の飛行時間と半減期を関連づける問題で、問題文中の条件から飛行中にガンマ線を放出しない原子核があることに気付かないと

表3. 正答率適合度の悪い項目

(物理)

階級	学 力	項 目 22		項 目 23	
		実 際	モデル	実 際	モデル
1	-2.29~-0.78	0.176	0.102	0.106	0.149
2	-0.78~-0.67	0.170	0.134	0.158	0.154
3	-0.67~-0.54	0.202	0.159	0.118	0.157
4	-0.54~-0.40	0.190	0.195	0.142	0.161
5	-0.40~-0.26	0.214	0.238	0.110	0.165
6	-0.26~-0.09	0.250	0.292	0.136	0.170
7	-0.09~ 0.11	0.288	0.365	0.152	0.176
8	0.11~ 0.39	0.402	0.465	0.210	0.183
9	0.39~ 0.94	0.620	0.631	0.296	0.195
10	0.94~ 4.48	0.974	0.912	0.344	0.264
適 合 度 χ^2		56.29		66.37	

各階級は全て500人

(化学)

階級	学 力	項 目 11		項 目 15	
		実 際	モデル	実 際	モデル
1	-4.28~-1.17	0.098	0.064	0.437	0.296
2	-1.17~-0.79	0.166	0.119	0.384	0.328
3	-0.79~-0.53	0.188	0.169	0.331	0.346
4	-0.53~-0.29	0.200	0.216	0.321	0.360
5	-0.29~-0.08	0.222	0.267	0.288	0.373
6	-0.08~ 0.14	0.281	0.325	0.303	0.386
7	0.14~ 0.40	0.353	0.397	0.345	0.401
8	0.40~ 0.71	0.441	0.483	0.373	0.418
9	0.71~ 1.20	0.662	0.604	0.458	0.442
10	1.20~ 5.00	0.860	0.819	0.625	0.508
適 合 度 χ^2		35.26		79.96	

各階級は全て500人

(生物)

階級	学 力	項 目 17		項 目 20		項 目 25		項 目 27	
		実 際	モデル						
1	-3.55~-1.06	0.000	0.000	0.514	0.381	0.164	0.093	0.126	0.137
2	-1.06~-0.69	0.006	0.011	0.464	0.421	0.116	0.125	0.344	0.354
3	-0.69~-0.53	0.016	0.159	0.396	0.439	0.130	0.142	0.684	0.495
4	-0.53~-0.35	0.660	0.594	0.396	0.450	0.164	0.154	0.490	0.585
5	-0.35~-0.19	1.000	0.923	0.412	0.461	0.120	0.166	0.592	0.671
6	-0.19~ 0.05	1.000	0.993	0.411	0.475	0.128	0.182	0.760	0.763
7	0.05~ 0.32	1.000	1.000	0.462	0.492	0.206	0.204	0.850	0.849
8	0.32~ 0.67	1.000	1.000	0.490	0.513	0.240	0.234	0.926	0.919
9	0.67~ 1.26	1.000	1.000	0.564	0.542	0.292	0.281	0.976	0.966
10	1.26~ 4.65	1.000	1.000	0.692	0.620	0.468	0.434	0.992	0.995
適 合 度 χ^2		72.13		43.64		44.59		49.31	

各階級は全て500人

階級	学 力	項 目 31	
		実 際	モデル
1	-3.55~-1.06	0.482	0.498
2	-1.06~-0.69	0.374	0.500
3	-0.69~-0.53	0.410	0.501
4	-0.53~-0.35	0.360	0.502
5	-0.35~-0.19	0.402	0.503
6	-0.19~ 0.05	0.388	0.504
7	0.05~ 0.32	0.374	0.505
8	0.32~ 0.67	0.374	0.507
9	0.67~ 1.26	0.398	0.508
10	1.26~ 4.65	0.452	0.513
適 合 度 χ^2		118.37	

表3. 正答率適合度の悪い項目 (続)

(地学)

階級	学 力	項 目 1		項 目 6		項 目 7		項 目 25	
		実 際	モデル	実 際	モデル	実 際	モデル	実 際	モデル
1	-2.55~-1.04	0.338	0.238	0.252	0.213	0.256	0.167	0.134	0.499
2	-1.04~-0.78	0.352	0.341	0.480	0.372	0.258	0.199	0.124	0.500
3	-0.78~-0.56	0.427	0.415	0.388	0.468	0.220	0.217	0.124	0.501
4	-0.56~-0.36	0.422	0.482	0.450	0.553	0.232	0.233	0.104	0.502
5	-0.36~-0.18	0.474	0.543	0.644	0.657	0.198	0.254	0.100	0.503
6	-0.18~ 0.03	0.554	0.606	0.747	0.752	0.196	0.278	0.092	0.504
7	0.03~ 0.26	0.658	0.668	0.874	0.836	0.262	0.307	0.102	0.505
8	0.26~ 0.63	0.776	0.748	0.924	0.904	0.300	0.343	0.090	0.507
9	0.63~ 1.29	0.878	0.845	0.966	0.957	0.424	0.398	0.112	0.508
10	1.29~ 4.97	0.984	0.954	0.992	0.992	0.606	0.542	0.124	0.513
適 合 度 χ^2		33.14		36.86		61.27		1538.22	

各階級は全て500人

正解できない、応用力を試す良い問題とされている。しかし、原子物理分野は高等学校の物理の最後にあつて現役生は十分に学習できていない恐れがあるので共通第 1 次学力試験の問題としては不相当であるとの意見のある問題でもある。

表 3 をみると、項目 22 については、学力が中位以下の層 ($\theta < 0.11$) で正答率が 0.2 と低く、これらの層はあて推量によって解答している傾向が強いことが分かる。この項目は、学力の高い層 ($\theta > 0.11$) と中位以下の層 ($\theta < 0.11$) とを識別する項目である。一方、項目 23 は、全ての層にわたって正答率が 0.1 ~ 0.3 の範囲に集まっており、この項目は測定しようとしている学力差を識別しない。

化学では、モデルとの適合度が悪い項目として、項目 11 (第 2 問, 問 3 b), 項目 15 (第 3 問, 問 1 d) の 2 つが検出された。項目 11 は、液体の蒸気圧およびその蒸気を含む混合気体の全圧と成分気体の分圧に関する問題であり、深い理解力を問う良問とされている。しかし、この問題は、前問の結果に関連づけられていることから、このような形式の問題は避けてほしいとの要望も出されている問題で

もある。表 3 によると、項目 11 は、学力が中位以下の層 ($\theta < 0.14$) で正答率が低く、この層とそれ以上との層 ($\theta > 0.14$) とを識別する項目であることが分かる。学力が中位以下の層に対しては、前問の不出来が大きく影響したものと思われる。

項目 15 は、発熱反応について判断を求める問題である。出題形式としては工夫され好感がもてるとの評価もあるが受験者には程度が高すぎるとの意見もある問題である。分析の結果をみると、ほとんどの層で正答率が 0.3 から 0.4 の間で変動し問題の難しさを反映しており、それがモデルとの適合度を悪くしたものと考えられる。この項目は、学力差を識別しない。

生物では、項目 17, 20 (第 3 問, 問 1 問 3), 25, 27 (第 4 問, 問 3, 問 4), 31 (第 5 問, 問 4) がモデルとの適合度が悪かった。項目 17, 20 を含む第 3 問は、実験を題材にして、グラフの読み取りからホルモン同士の関係やフィードバック機構について考えさせる問題で良問との評を得ている。

これらの項目について、批判的な意見は出されていないが、項目 17 は、学力が低い層 (

$\theta < -0.53$) とそれより高い層 ($\theta > -0.53$) との間の正答率に極端な差がついており、学力の低い層を識別する項目である。これに対し、項目20は、学力の全域に渡って正答率が0.4～0.6の間で変動しており、学力差を識別しない。

項目25,27を含む第4問は、細胞の分裂に伴うDNA量の変化、染色体数とDNA量との関係等について正しく理解されているかどうかを問う基本的な問題であるとの評を得ている。また、項目25は、難解であったとの評も得ている。分析でも、最上位階級(1.26～4.65)を除くほぼ全域に渡って0.1から0.2の低い正答率が表れており、この見解を支持している。この項目は、最上位階級だけを識別する。また、項目27は、学力の低い層 ($\theta < -0.53$) で正答率が変動し、この層における学力差を識別しない。これに対し、項目31は、細胞間の識別についてよく理解しているかどうかを問うた問題であるが、分析結果をみると、項目31は全域で正答率が0.3～0.4の間を変動し、すべての層について識別しない項目であることが分かる。

地学では、項目1,6(第1問, a A群, c B群), 7(第2問, 問1)がモデルとの適合度が悪かった。項目1,6を含む第1問は、恒星と宇宙に関する基本的事項を総合的に扱い、系統的理解力を問う良問とされているが、分析の結果、学力の低い層 ($\theta < -0.36$) の正答率の変動がモデルとの適合度を悪くしたものであると思われる。

項目7は、太陽の黒点の明るさに関する出題で、基礎・基本を踏まえた応用力や計算力を問うた問題である。学力の高い層 ($\theta > 0.63$) を除き、それ以下の層 ($\theta < 0.63$) では計算力の不足のせいか、正答率が0.1から0.2と低く、あて推量で解答していると思われる。この項目は、学力の高い層 ($\theta > 0.63$)

を識別する項目である。

また、項目25(第6問, 問3)は、地質調査結果に関し、科学的探究の方法・探究の過程を盛り込み、再調査するという新しいアイデアのもとに、よく工夫された良問との評を得ている。分析の結果をみると、全ての層で正答率は0.1前後で変動しており、これがモデルとの適合度を悪くしている。この問題の識別力は低い。

数学Iについては、各合成テスト毎にモデルの適合度について検定を行い、その中で特に適合の悪かった項目だけを取り上げた。それらの項目番号は、項目11,13(第2問), 15,16(第3問)であった。その中で、項目13は全ての合成テストにおいてモデルとの適合が悪いものとして検出された。また、項目15は「数学I—生物」および「数学I—地学」の合成テストにおいて検出されたが、残りの項目11,16は「数学I—生物」の合成テストのみで検出された。

これらの項目は、いずれも内容的には良問との評を得ている。項目11は、生物受験者の中で学力が低い層 ($\theta < -1.11$) を識別し、また項目15,16は生物あるいは地学受験者の中で学力が高い層 ($\theta > 0.71$) を識別する項目である。これに対し、項目13はいずれの科目の受験者の場合も正答率が階級によって大きく上下に変動し、学力の高い層 ($\theta > 0.77$ 前後) より、中位の層 (θ が $-0.04 \sim 0.19$) の方が正答率が高くなるといった現象を示しており、これがモデルとの適合度を悪くしたものである。

6. 1次元性の検討と項目の選別II

項目反応理論の大きな特徴である1次元性の仮定の確認については、この理論を適用する際の必要不可欠の条件である。この仮定を

確認する方法として、一般には因子分析を利用する機会が多いが、ここでは、項目パラメタの性質を利用して、次のような手続きによって確認する。

- ① 単独テスト（例えば物理のテスト）と合成テスト（例えば数学 I—物理からなる合成テスト）の 2 つのテストを考えてみる。
- ② 今、単独テスト（項目数 n ）が 1 次元の学力を測定していると仮定する。この仮定の下で、分析の対象となる受験者によって、単独テストの項目パラメタ $a_i, b_i, (i=1, 2, \dots, n)$ を推定する。
- ③ 次に、合成テスト（項目数 $m, m > n$ ）も 1 次元の学力を測定していると仮定して、同一受験者を利用して、項目パラメタ $a_i^*, b_i^*, (i=1, 2, \dots, m)$ を推定する。
- ④ もし、2 つのテストが同一の学力を測定しているのであるならば、2 つのテストに共通な項目の項目パラメタは同じ値をもつはずである。すなわち、 $a_i = a_i^*, b_i = b_i^*$ の関係が成り立つはずである。言い換えれば、2 つのテストの項目パラメタをプロットしたとき、各点 (a_i, a_i^*) および (b_i, b_i^*) は傾斜 1 で切片 0 の直線上 ($Y=X$) に散布する。
- ⑤ しかし、もし項目パラメタの推定値が $Y=X$ の直線上から外れるならば、その項目に正答するか否かは、テストに含まれる項目に依存することになり、すべての項目が一つの学力で説明されるという仮定と矛盾が生ずることになる。すなわち、その項目は 1 次元ではなく、他の学力も測っていることになる。

上記の手続きによって、それぞれの「理科」の科目について単独テストと共通項目である数学 I を加えたときの合成テストの項目パラ

メタを 2 パラメタ・ロジスティック・モデルによって推定した。そして、その結果を散布図で確認したところ、直線 $Y=X$ から大きく外れる項目がいくつか見つかった。そこで、これらの項目は、1 次元性が低いとみなし、テストを構成する項目群から除外することにした。表 1 の最右欄に示した項目のうち下線がついていない項目番号がそれである。除外された項目数は、物理では 9、化学では 2、生物では 1、地学では 5 つであった。これらの項目の項目識別力は総じて大きく、正答率も高い傾向を示した。

図 3 は、こうして項目を選別した後の項目群について再推定しなおして、 (a_i, a_i^*) および (b_i, b_i^*) をプロットしたものである。項目識別力の相関図をみると、各科目共相関は高く ($r > 0.93$)、直線の傾きは 0.83~0.89 の間で、切片も 0 の近傍にある。また項目困難度の相関図をみると、各科目共相関係数は高く ($r > 0.99$)、ほぼ直線 1 で傾き 0 の直線上にある。

ところで、累積正規曲線を仮定した場合、項目識別力 a_i と項目困難度 b_i は、項目 i に正答する学力と共通学力 θ との相関係数 ρ_i の関数で表されることが知られている (Lord, 1968)。ただし、 γ_i は、項目 i に正答することのできる能力の限界値である。

$$a_i = \frac{\rho_i}{1 - \rho_i^2} \quad (1)$$

$$b_i = \frac{\gamma_i}{\rho_i} \quad (2)$$

今、2 つのテストで推定された同一項目 i の相関係数を ρ_i, ρ_i^* とする。このとき、上式から $\rho_i = \rho_i^*$ であれば、 $a_i = a_i^*, b_i = b_i^*$ が成立し、 $\rho_i > \rho_i^*$ であれば、 $a_i > a_i^*, b_i < b_i^*$ が成立することは明らかである。従って、もし、1 次元性が成り立たないのであるならば、この相関係数 ρ_i と ρ_i^* の間には何らかの

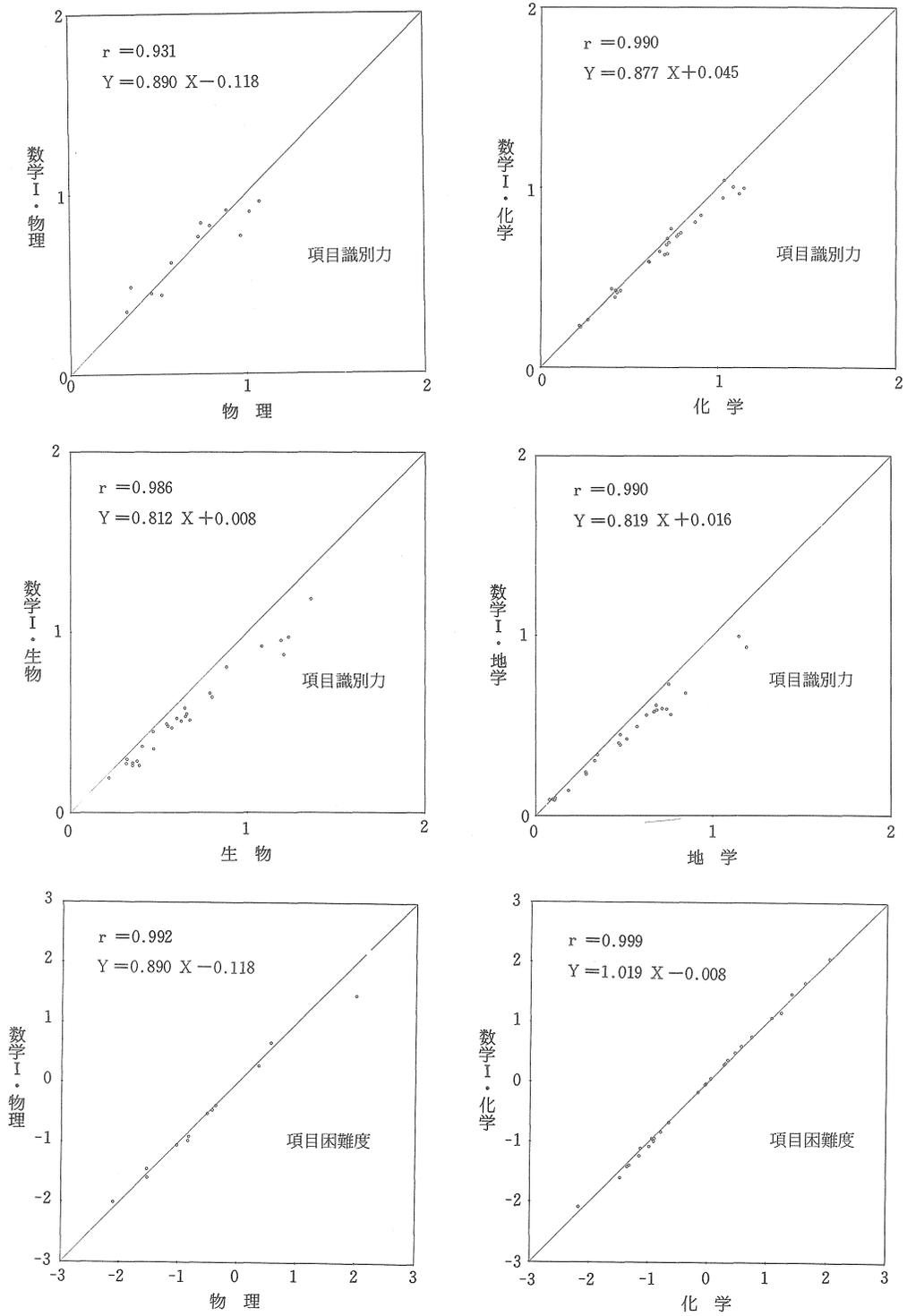


図3. 項目選別後の項目識別力等の相関図

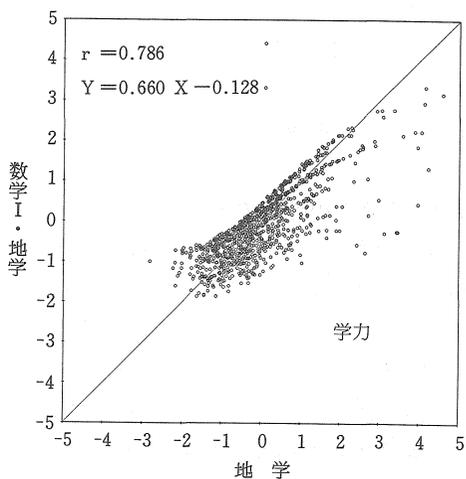
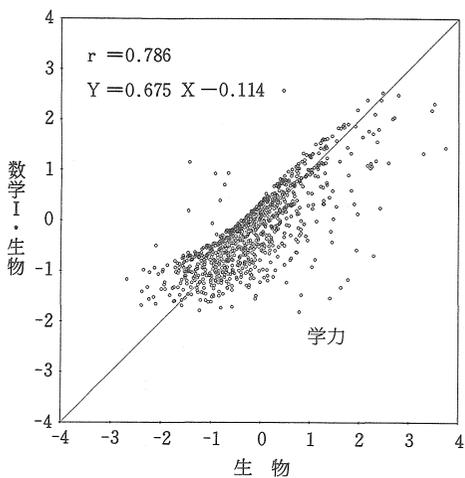
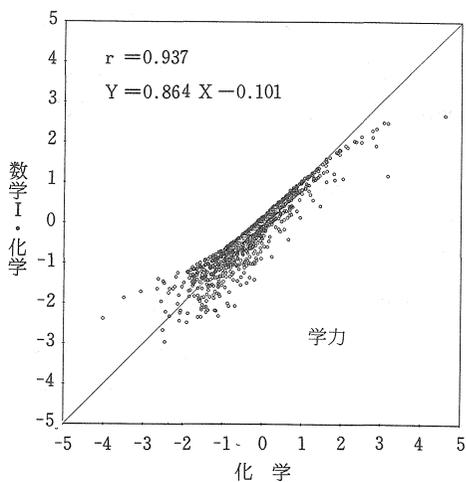
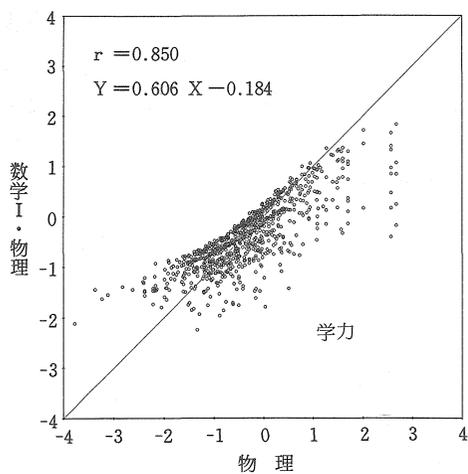
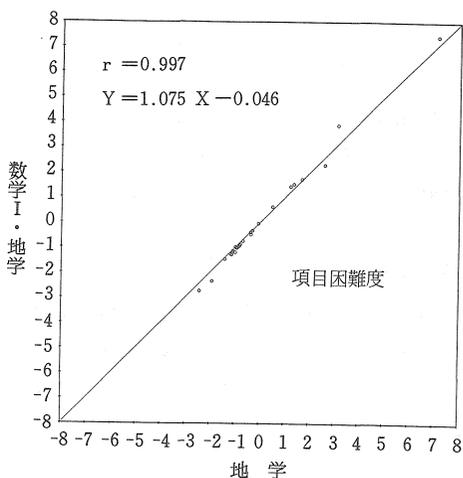
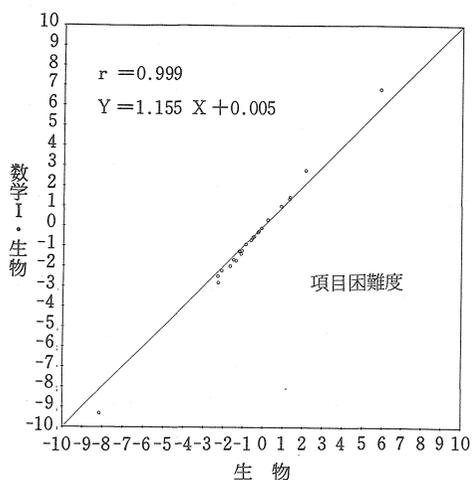


図3. 項目選別後の項目識別力等の相関図 (続)

大小関係が生じ、項目パラメタに影響するはずである。このことは、共通項目の項目パラメタをプロットしたとき、項目識別力については直線 $Y = X$ より下に、また、項目困難度については直線 $Y = X$ より上に散布することを意味する。

そこで、相関係数 ρ_i の近似として因子分析の第1因子負荷量を考えたとき、2つのテストで得られた共通項目の第1因子負荷量を比較してみると、単独テストの場合に比べ合成テストの方が因子負荷量は僅かに小さくなる傾向が見出された。そのため、項目識別力の散布は $Y = X$ の直線から僅かに外れる傾向を示したものと思われる。しかし、項目識別力は項目困難度と比べ推定の精度は劣る (Bejar, 1980) ことを考慮すれば、共通項目として数学 I を加えた合成テストと理科の1科目からなる単独テストとはほぼ同一の学力を測定していると考えても差支えないと考えられる。

なお相関図には、項目選別後の2つのテストの項目パラメタによって推定された受験者 (1,000人) の学力 θ についても併せて載せてある。この散布状況を見ると、左上側面の散布が押しつぶされた形状を示しているが、これは完全な1次元性を保持していないために生じたものと推測される。

7. テストの等化

7.1 テストの等化の方法

さて、前節までで、項目反応理論を適用する際の1次元性の仮定に関して確認を行ったわけであるが、本節より1次元性の確認が得られた理科の単独テストと数学 I を加えた合成テストを利用して、テストの等化 (Hambleton & Swaminathan, 1985) を行うことを考える。テストの等化とは、各テストの結果

を共通の尺度上の値に変換することを意味する。

そこで、単独テストと合成テストを利用して、図4に示すような等化のためのテスト・デザインを考える。これは共通項目として数学 I を利用し、これを基にして理科の各科目を等化しようとするものである。等化の対象となる合成テストの項目は先の2回の選別で外した項目を除いた項目であり、表1に示すように、「数学 I—物理」は25項目、「数学 I—化学」は40項目、「数学 I—生物」は40項目、「数学 I—地学」は39項目から成っている。

合成テスト j が項目 i を含むとき、その合成テストを用いて推定された項目パラメタの値を a_{ij} , b_{ij} , また、同様にして推定される受験者の学力の値を θ_j とすると、これらは、等化後の共通尺度上でのパラメタの値 a_i , b_i ,

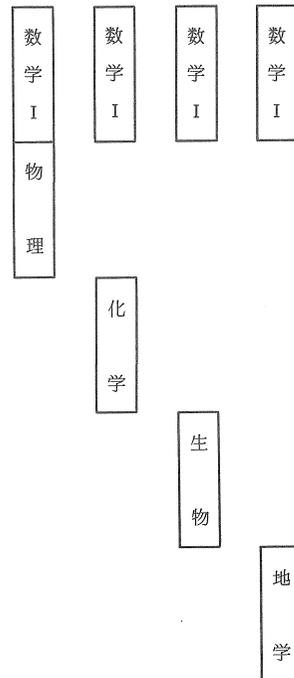


図4. 等化のためのテストデザイン

θ とすると、次の関係をもつと想定される。

$$a_i = a_{ij} / r_j \quad (3)$$

$$b_i = r_j b_{ij} + q_j \quad (4)$$

$$\theta = r_j \theta_j + q_j \quad (5)$$

ただし、 q_j と r_j は等化のための係数であり、実際には、 a_{ij} と b_{ij} の推定値が与えられた時に、それらを用いて、(3)と(4)の関係を近似的に満たす様な等化係数と共通尺度上での項目パラメタの値を推定しなければならない。ここでは、前川・鈴木(1988)が Q_2 と呼ぶ、等化前の項目特性曲線 $P_{ij}(\theta_j)$ と等化後の項目特性曲線 $P_i(\theta)$ との2乗誤差を最小にする方法を用いた。

このとき、学力 θ_j が平均0、分散1の分布を仮定すると(5)式から、合成テストを受験した者の学力分布は、共通尺度上において平均 $E(\theta)$ が、

$$E(\theta) = q_j \quad (6)$$

で、分散 $V(\theta)$ が、

$$V(\theta) = r_j^2 \quad (7)$$

で与えられることが分かる。また、テストの難易度の指標となるテスト特性曲線は、

$$T(\theta) = \sum P_i(\theta) \quad (8)$$

として、テストに含まれる項目特性曲線の和として与えられる。

7.2 単独テスト間の1次元性の確認

ところで、テストの等化を行う際の条件として、対象となるテスト間で測定している学力が同一であること(Lord, 1980, Hambleton & Swaminathan, 1985)があげられている。先の1次元性の確認は、単独テストおよび合成テストへのモデルの適用可能性を確認するためのものであって、科目間での1次元性の確認は行っていない。そこで、この点について、先の各テスト毎に確認した方法を利用し、次のような手続きで確認する。

① 今、全ての単独テスト(数学I、物理、

化学、生物、地学)が1次元であるという仮定の下で、「数学I-物理」、「数学I-化学」、「数学I-生物」、「数学I-地学」の合成テストの項目パラメタを利用してテストの等化を行う。そのとき、等化によって得られた数学I、物理、化学等の項目パラメタをそれぞれ a_i 、 b_i ($i=1,2,\dots,n$)とする。

② 次に、各単独テスト毎に1次元性を仮定して、それぞれの単独テストの項目パラメタを推定する。その値を a_{ij} 、 b_{ij} (j =数、物、化、生、地、 $i=1,2,\dots,n$)とする。

③ もし、①の仮定が正しいのであれば、項目特性曲線の性質により、 a_i と a_{ij} および b_i と b_{ij} は線型変換で一致するはずである。すなわち、 (a_i, a_{ij}) と (b_i, b_{ij}) の相関係数は全ての i に関して1となるはずである。

以上の手続きにより、等化後の項目パラメタとすでに求めた各科目テストの項目パラメタの相関係数を算出した。図5は、それらの項目パラメタの相関図を示したものである。その結果をみると、項目識別力の相関係数は数学Iを除き0.93以上で、項目困難度の相関係数は0.94以上であり、かなり高い相関を得ていることが分かる。数学Iの項目識別力の相関係数が低かったのは、単独テストにおいて項目7が外れ値として大きな値で推定されたためである。なお、これらの相関係数は先の図3に示した結果と一致しているが、これは、「理科」の全ての科目に共通項目として数学Iを利用したためで、もし、一部で共通の項目を利用(例えば、過年度の2科目選択における「物理-化学」や「物理-生物」のような組み合わせ)する場合は、異なった相関係数が得られるはずである。

以上から、各単独テストは同一の学力を測定していると考え、等化を行うための条件は

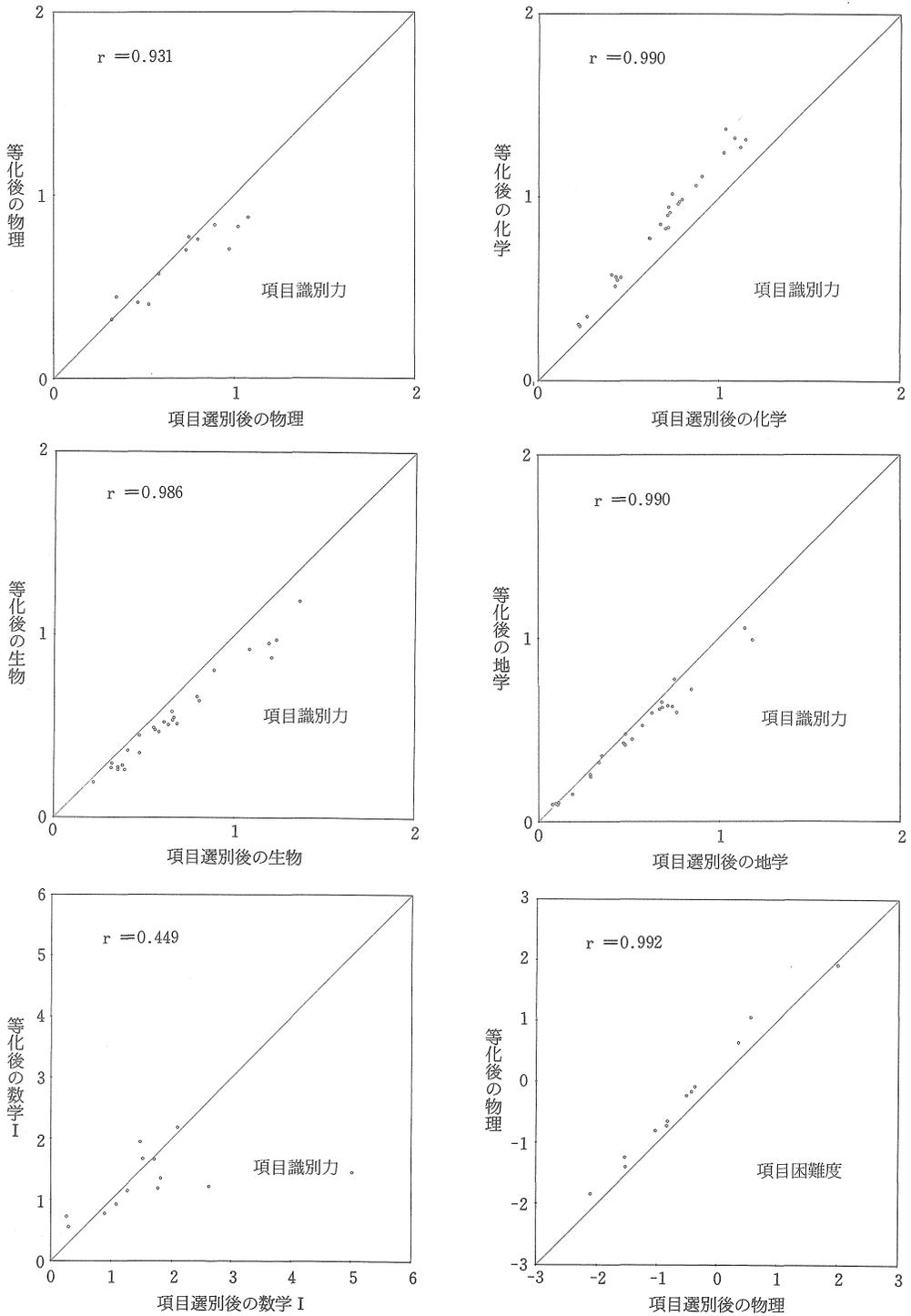


図5. 等化のための1次元性の確認

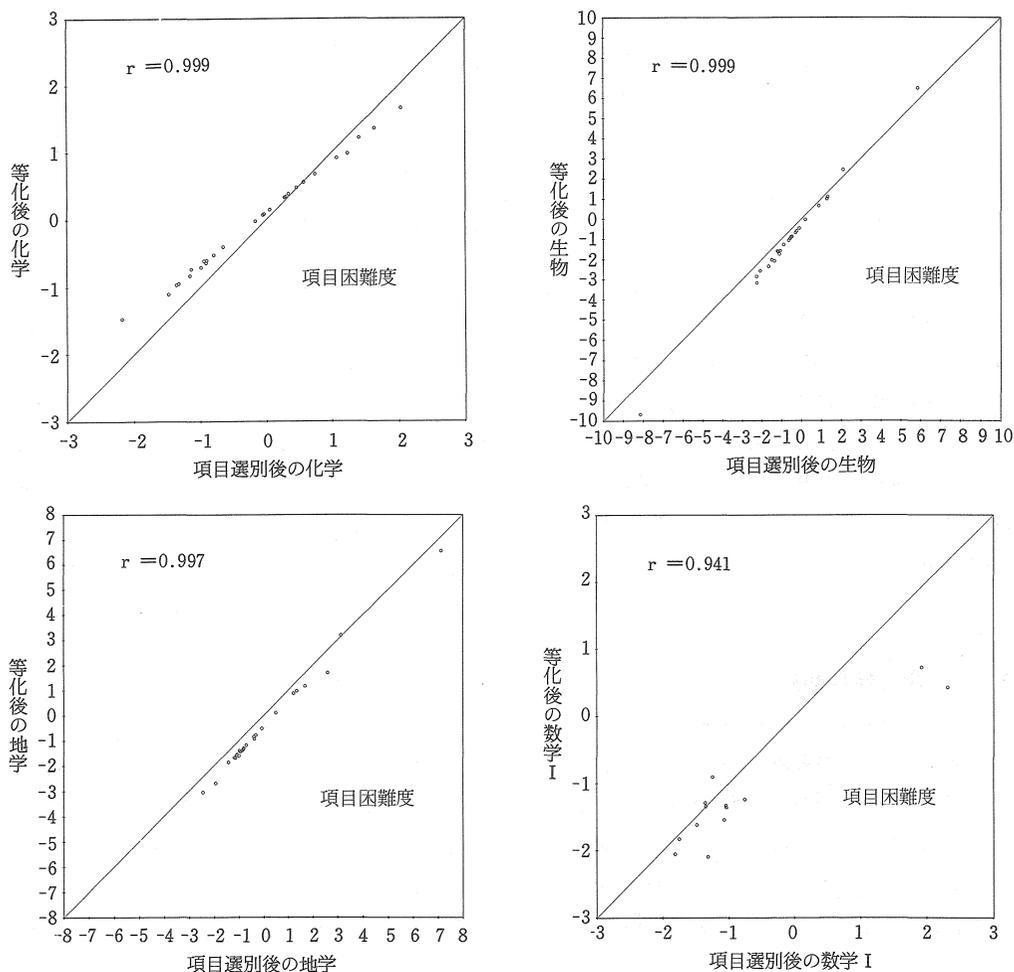


図5. 等化のための 1 次元性の確認 (続)

満足していると考える。

7.3 等化後の各科目のテスト特性

7.3.1 合成テストに関する項目パラメータと学力の分布

合成テストを利用してテストの等化を行い、その結果として、共通尺度の上での各項目パラメータの値を得ることができた。表 4 は、共通尺度上でのそれらの項目パラメータと学力の分布の要約値を示したものである。

等化後の合成テストの項目識別力の平均値

をみると、「数学 I—物理」，「数学 I—化学」で大きく、「数学 I—生物」，「数学 I—地学」で小さい値を示している。また、項目困難度の平均値をみると、「数学 I—化学」で大きく、「数学 I—生物」で小さな値を示している。このことから、「数学 I—化学」は、やや難しい項目から構成され、それらの項目の識別力の高いテストであることが分かる。逆に、「数学 I—生物」は、やや易しい項目からなり、項目の識別力の低いテストであることが分かる。

表4. 合成テストに関する項目パラメタと学力の分布

合成テスト	項目識別力		項目困難度		学 力	
	平 均	標準偏差	平 均	標準偏差	平 均	標準偏差
数学 I—物理	1.069	0.640	-0.776	1.039	0.339	1.089
数学 I—化学	1.057	0.532	-0.408	0.993	0.101	0.757
数学 I—生物	0.844	0.608	-1.125	2.129	-0.346	1.001
数学 I—地学	0.809	0.635	-0.758	1.704	-0.467	0.944

表4の右欄の学力は、テストの等化によって調整された後の各科目を選択した集団毎の学力の分布の要約値を示したものである。この平均値をみると、物理を受験した集団は、他の科目を選択した集団より学力が高いことを示している。次いで、化学、生物、地学をそれぞれ選択した集団の順に下がっていく傾向にある。

7.3.2 項目特性曲線

図6は、各科目の項目特性曲線を描いたものである。またその要約結果を表5に示す。

項目特性曲線は、横軸に学力 θ をとり、それに対応する項目 i の正答確率 $P_i(\theta)$ を縦軸にとって表したものである。図中、項目特性曲線で項目識別力が小さいものは、項目番号を

付している。この項目特性曲線と項目パラメタの要約値から次の点を指摘することができる。

- (1) 物理の項目特性曲線をみると、いずれの項目も類似した形状を示しているが、困難度に少しばらつきがある。
- (2) 化学の項目特性曲線をみると、特に項目識別力が類似した項目が多い。項目困難度の平均値は、化学で最も大きい値を示していることを考えると、化学は他の科目に比べ難しい科目であることが分かる。その中で項目識別力の小さい項目が項目12,13,14に見られた。項目12,13,14は第3問に含まれた化学現象について問う問題であるが、選択肢が3つということで偶然性が高いのではないかとの意見も出されている問題で

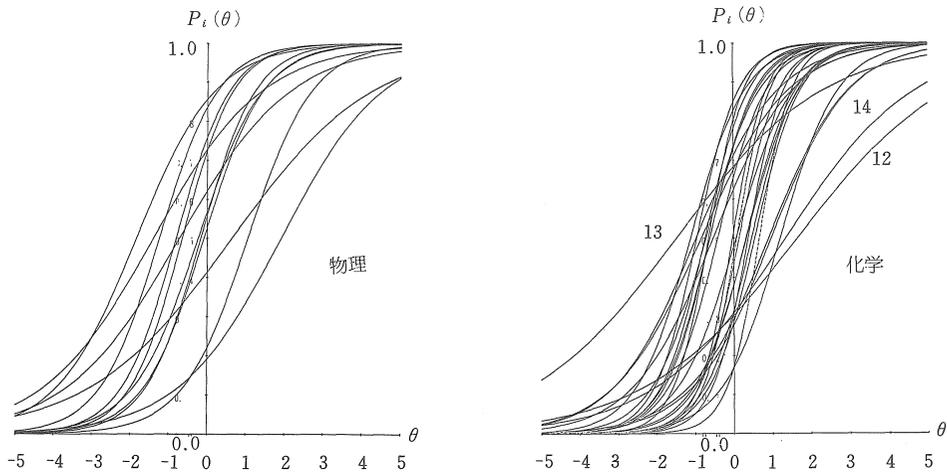


図6. 項目特性曲線

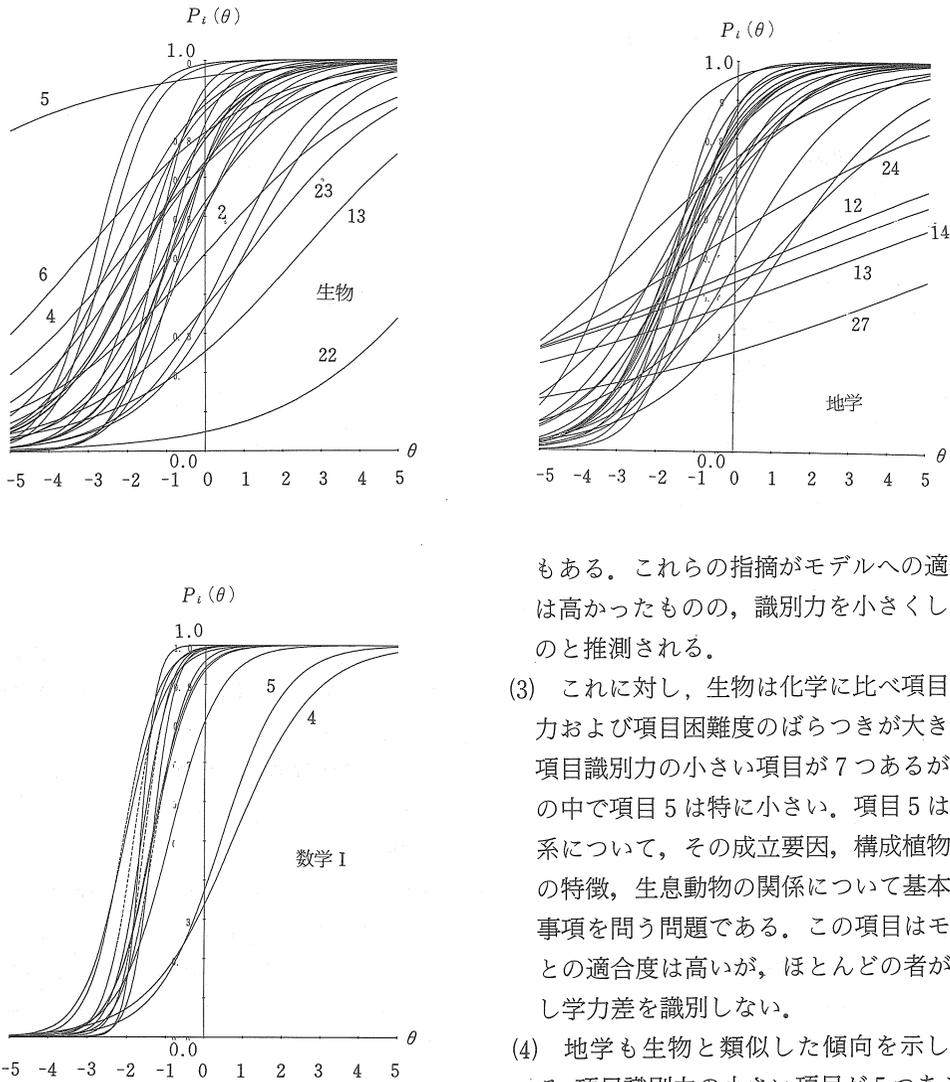


図6. 項目特性曲線 (続)

もある。これらの指摘がモデルへの適合度は高かったものの、識別力を小さくしたものと推測される。

- (3) これに対し、生物は化学に比べ項目識別力および項目困難度のばらつきが大きい。項目識別力の小さい項目が7つあるが、その中で項目5は特に小さい。項目5は植物系について、その成立要因、構成植物群落の特徴、生息動物の関係について基本的な事項を問う問題である。この項目はモデルとの適合度は高いが、ほとんどの者が正答し学力差を識別しない。
- (4) 地学も生物と類似した傾向を示している。項目識別力の小さい項目が5つあるが、項目27は、学力の大きい層でも正答確率が

表5. 単独テストに関する項目パラメタと平均正答率

単独テスト	項目識別力		項目困難度		平均正答率		
	平均	標準偏差	平均	標準偏差	$\theta = -1$	$\theta = 0$	$\theta = 1$
物 理	0.634	0.187	-0.293	1.030	0.360	0.583	0.763
化 学	0.858	0.308	-0.016	0.812	0.247	0.513	0.772
生 物	0.542	0.255	-1.078	2.526	0.532	0.692	0.798
地 学	0.479	0.259	-0.526	1.963	0.527	0.681	0.778
数 学 I	1.470	0.649	-1.222	0.827	0.691	0.869	0.936

低く、学力差を識別しない。項目27は第7問の先頭の問題である。地殻を構成する元素と SiO_4 四面体に関する問題であるが、教科書にはそれに関する記述がほとんどなく、地学の問題としての意義があるかどうか疑問が残るとの意見もある。

- (5) 数学 I も化学と同様に項目識別力および項目困難度が類似した項目が多い。また、要約値から、数学 I は項目識別力の平均値は最も大きく、項目困難度は最も小さい値を示すことから、この科目は、他の科目に比べ学力の小さい層を識別する項目群から構成されていることが分かる。

7.3.3 テスト特性曲線

テストの特性を表す一つの指標としてテスト特性曲線がある。テスト特性曲線は、テストを構成する項目の項目特性曲線の総和をとったものである(式(8)参照)。従って、テスト特性曲線は、学力 θ に対するテストの平均正答率を表したものと考えることができる。このテスト特性曲線は、テストを構成する項目の数や項目パラメタの値によって形状は変化する。また、学力 θ を実際の得点に写像するための非線型変換関数でもある。例えば学力の分布が正規分布していたとしても、このテスト特性曲線による変換によって実際のテスト結果はいろいろな形状に変化する。図7は各科目のテスト特性曲線をまとめて1つの図の中に描いたものである。

図7をみると、 $\theta=0$ では、テスト特性曲線は数学 I で最も大きく、生物、地学、物理、化学の順となっている。すなわち、 $\theta=0$ では化学が最も難しく、次いで物理、地学、生物、数学 I の順に易しくなっていくことが分かる(表5参照)。同様に、 $\theta=-3$ では、数学 I が最も難しく、化学、物理の順に易しくなり、生物と地学では殆ど難しさは同じであること

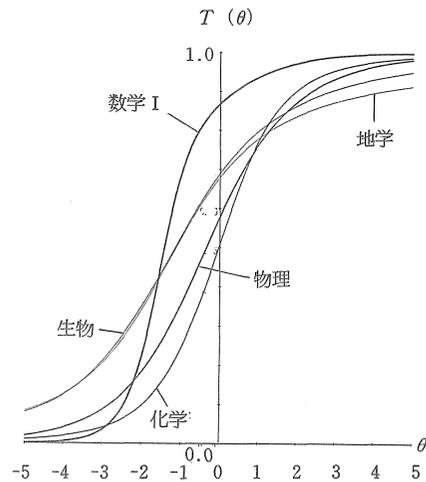


図7. 等化後の数学 I・理科各科目のテスト特性曲線

が分かる。これに対し、 $\theta=3$ では、地学が最も難しく、次いで生物、物理、化学、数学 I の順に易しくなることが示される。このように、テスト特性曲線は、学力 θ の違いによる各テストの難しさを連続的に表すことができる点に特徴がある。

図8は、そのテスト特性曲線を科目毎に示したものである。この図には、学力 θ が各集団毎に表4に示す平均と分散をもつ正規分布をすると仮定したときの得点分布も併せて載せてある。

各科目の得点分布をみると、物理は、少し下に裾を引いた正規分布に近い形状を示し、化学は台形の形状を示している。これに対し、生物、地学は類似した形状を示し、下に長く裾を引いていることが分かる。また、数学 I は、得点の高い層に山があるが、中腹にも小さな山がある形状を示している。

図9は、各科目毎に推定された学力と実際の科目得点(100点満点)との関係を1,000人の受験者について示したものである。図中には、さらに周辺分布である学力分布と得点

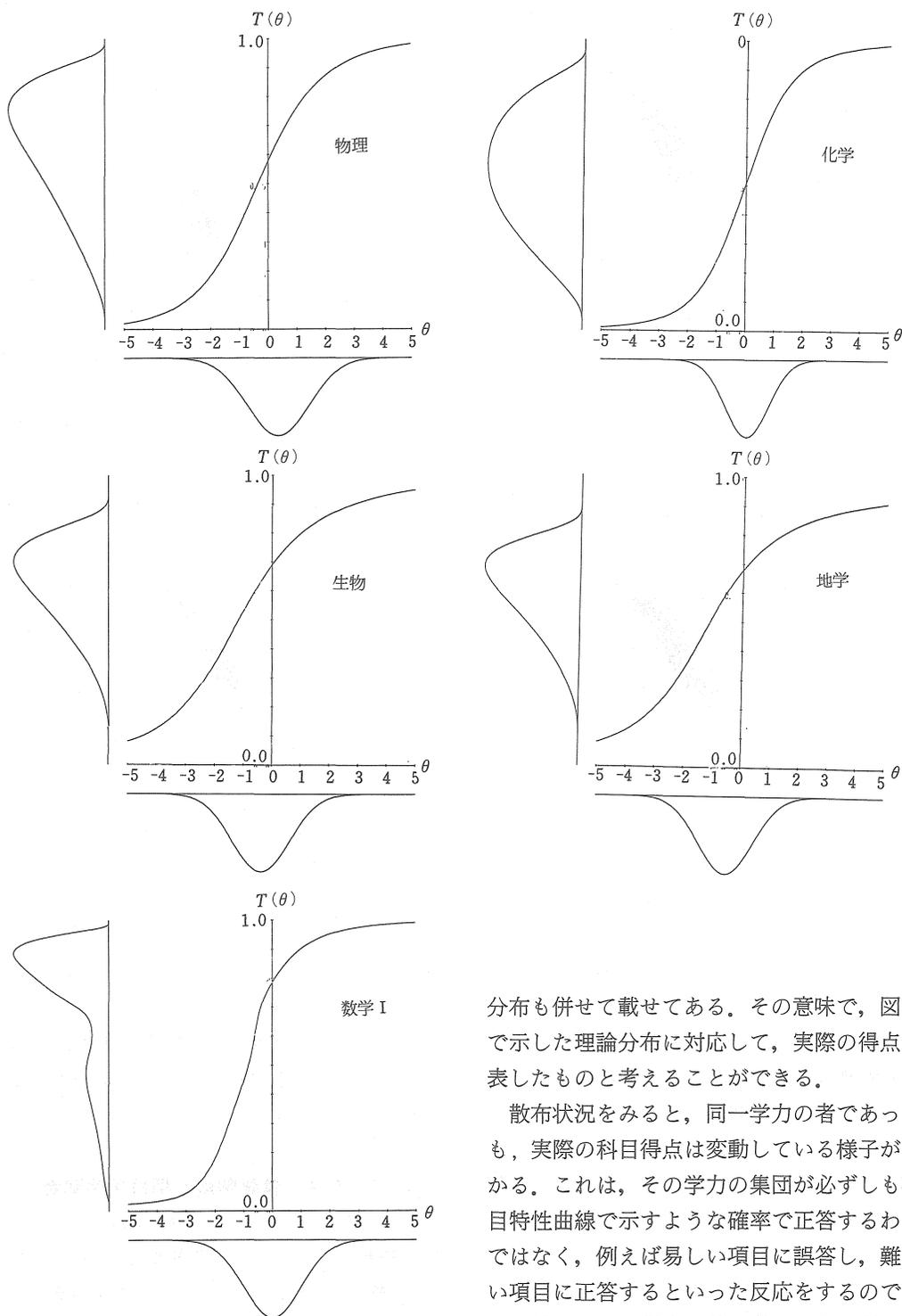


図8. テスト特性曲線

分布も併せて載せてある。その意味で、図8で示した理論分布に対応して、実際の得点で表したものと考えることができる。

散布状況を見ると、同一学力の者であっても、実際の科目得点は変動している様子が分かる。これは、その学力の集団が必ずしも項目特性曲線で示すような確率で正答するわけではなく、例えば易しい項目に誤答し、難しい項目に正答するといった反応をするので、その結果が項目得点の積み重ねである科目得

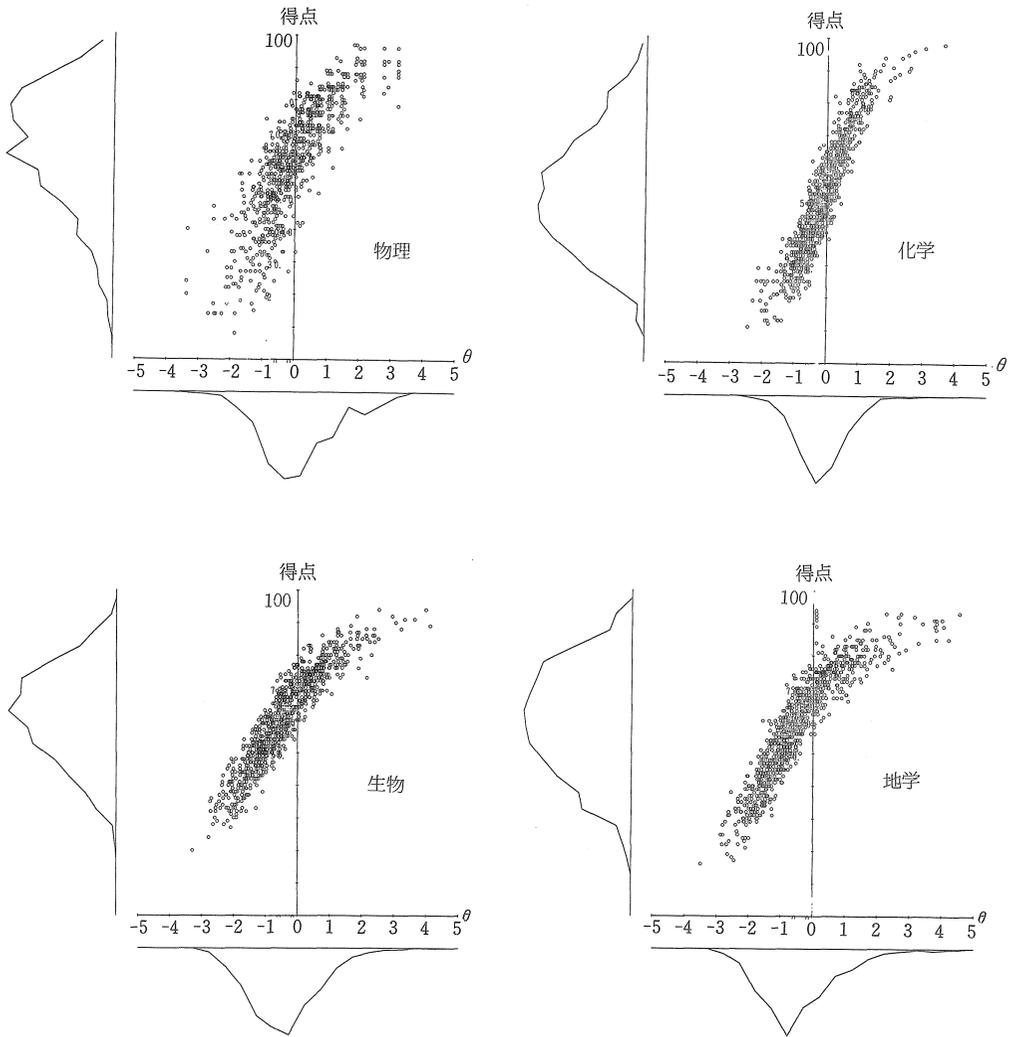


図9. 学力 θ と科目得点との関係

点に影響を与え、このような変動を生じたものと推測される。しかし、散布の状況から、学力と科目得点の間にはおよそ図8のテスト特性曲線で表されるような関係が見てとれる。

各科目の学力分布を比較すると、学力の平均値は、先の表4に示したような傾向を示していることが分かる。また、学力分布と得点分布を対比してみると、概ね各科目共学力の

分布に対応するような得点分布が得られているが、生物、地学の得点は学力に比べやや高く表れている様子が伺える。

7.3.4 テスト情報関数と項目情報関数

テストがどのような学力 θ を識別するかを表す指標としてテスト情報関数 $I(\theta)$ がある。同様に、項目がどのような学力 θ を識別するかを表す指標として項目情報関数 $I_j(\theta)$ が

ある。

先のモデルの適合性の検討に際し、学力差の識別について議論をしたが、これらの議論を要約したものが、この項目情報関数である。従って、項目情報関数は、項目の識別力と密接な関係がある。項目 i の項目情報関数は、次のように定義されている。

$$I_i(\theta) = \frac{P_i'(\theta)^2}{P_i(\theta) \cdot Q_i(\theta)}$$

ここで、 $Q_i(\theta) = 1 - P_i(\theta)$ 、また、 $P_i'(\theta)$ は $P_i(\theta)$ の θ に関する微分である。テスト情報関数は、項目情報関数の総和をとったものである。テストがあらゆる学力層に対し高い情報量（識別力）を必要とするならば、情報量があらゆる層で高くなるような項目を選んでテストを構成すればよいし、逆に狭いある学力層に限って情報量を得たいのであれば、その層で情報量が高くなる項目を選択したテストを構成すればよい。この観点から、分析の対象となった科目のテスト情報関数について比較してみる。図10は、各科目のテスト情報関数を描いたものである。また、図11には各科目毎の項目情報関数を描いてある。

各科目のテスト情報曲線のピークをみると、数学 I が最も高く、次いで化学、生物、地学、物理の順となっている。数学 I のピークは $\theta = -1.60$ で生じており、情報量の高い範囲は狭い。化学は、 $\theta = 0.02$ でピークを生じており、数学 I に比べれば丸いが、他の科目よりは尖っている。生物および地学はピークはほぼ同じ位置 ($\theta = -1.30$) で生じているが、情報量は低くテスト情報関数も類似している。物理は、他の4科目に比べ情報量は最も低く（ピークは $\theta = -0.50$ ）、台形に近い形状を示している。

これらの各科目の情報量が最大になる θ に対応するテスト特性曲線の値をみると、およそ 0.5 から 0.55 の間にある。従って、共通

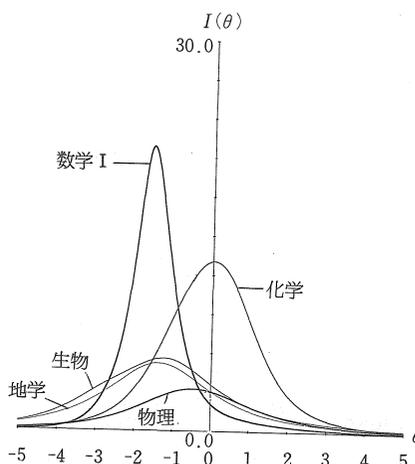


図10. 等化後の数学 I・理科各科目のテスト情報関数

第1次学力試験の「理科」の各科目は50点から55点の得点を得る者について学力を識別する科目であると言えよう。また、情報量が最大になる各科目の θ の大小関係を比較すると、化学はやや高い学力層を識別する傾向が強く、もう少し易しい項目を加えてもいいのではないかと思われる。ただし、物理のピークにおける情報量が低いのは、項目数が少ないために生じたものと推測される。なお、数学 I については、共通第1次学力試験の「数学」は数学 I と数学 II からなっているので、2つの科目を加えた形で分析しない限り、十分な判断を下すことはできない。

7.4 前年度の「理科」各科目のテスト特性

さて、昭和62年度の「理科」の各科目に関してテスト特性を比較したが、この傾向が安定しているかどうか確認するため、前年度のテスト特性を調べてみる。先に述べた手順と同様な手順を踏み、テストの等化を行う。昭和61年度の「理科」は2科目選択制で、大部分の受験者が理科 I を必須科目として選択し

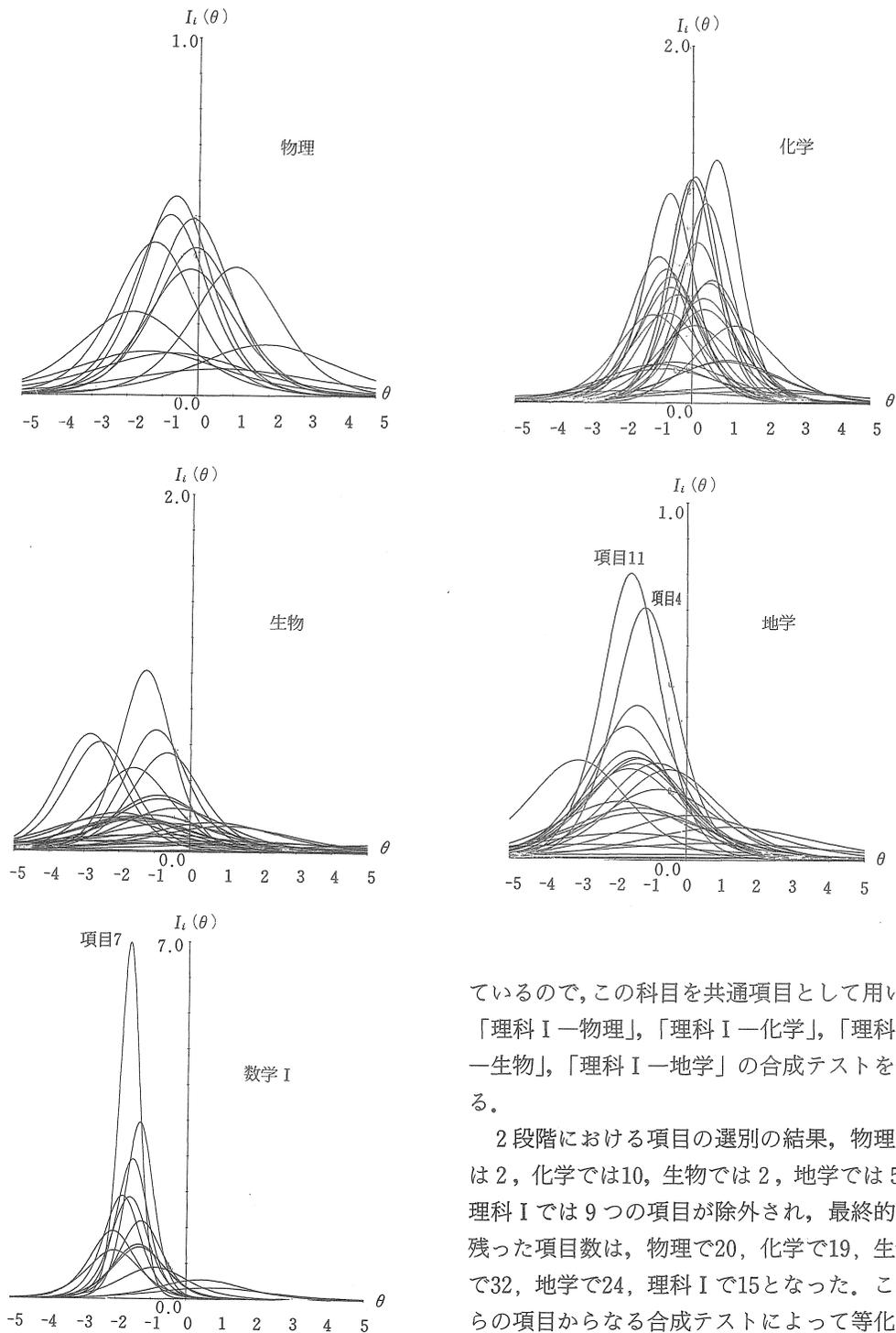


図11. 項目情報関数

ているので、この科目を共通項目として用い、「理科 I - 物理」、「理科 I - 化学」、「理科 I - 生物」、「理科 I - 地学」の合成テストを作る。

2段階における項目の選別の結果、物理では2、化学では10、生物では2、地学では5、理科 I では9つの項目が除外され、最終的に残った項目数は、物理で20、化学で19、生物で32、地学で24、理科 I で15となった。これらの項目からなる合成テストによって等化を行った。表6および表7は、その結果をまとめ

たものである。さらに、図12および図13に、それぞれ各科目のテスト特性曲線およびテスト情報関数をまとめて示している。

以上の表および図をみることにより、次のように要約することができる。

- (1) 学力の平均は物理を選択した集団の学力が最も高く、次いで化学、生物、地学の順となっている(表6)。
- (2) テスト特性曲線をみると、学力 $\theta=0$ で

は、理科Iが最も難しく、次いで地学、化学、物理、生物の順に正答する確率が高くなる傾向にある。 $\theta=-1$ では、物理>化学>理科I>地学>生物の順に正答する確率が高くなり、 $\theta=1$ では、物理I>地学>化学>生物>物理の順に正答する確率が高くなる傾向を示している(表7、図12)

- (3) また、テスト情報関数をみると、生物、物理が高いピークを示しており、化学、地学

表6. 合成テストに関する項目パラメタと学力の分布(昭和61年度)

合成テスト	項目識別力		項目困難度		学 力	
	平均	標準偏差	平均	標準偏差	平均	標準偏差
理科I-物理	0.920	0.467	-0.732	1.010	0.332	1.010
理科I-化学	0.675	0.255	-0.756	1.208	0.157	0.987
理科I-生物	0.785	0.434	-0.974	1.163	-0.363	0.848
理科I-地学	0.611	0.611	-0.955	1.173	-0.525	0.938

表7. 単独テストに関する項目パラメタと平均正答率(昭和61年度)

単独テスト	項目識別力		項目困難度		平均正答率		
	平均	標準偏差	平均	標準偏差	$\theta=-1$	$\theta=0$	$\theta=1$
物 理	1.173	0.434	-0.618	0.655	0.364	0.725	0.913
化 学	0.748	0.240	-0.656	1.093	0.445	0.700	0.852
生 物	0.879	0.470	-1.016	1.073	0.538	0.743	0.865
地 学	0.628	0.294	-0.999	1.061	0.480	0.688	0.836
理 科 I	0.583	0.243	-0.884	1.330	0.458	0.649	0.796

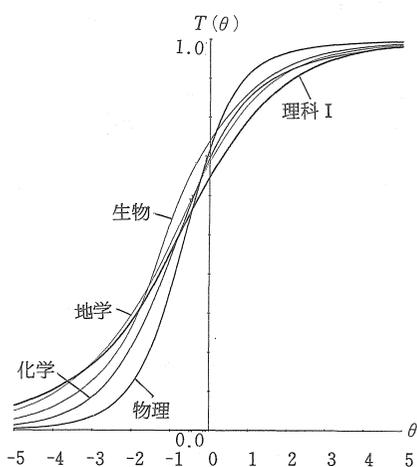


図12. 等化後の理科各科目のテスト特性曲線(昭和61年度)

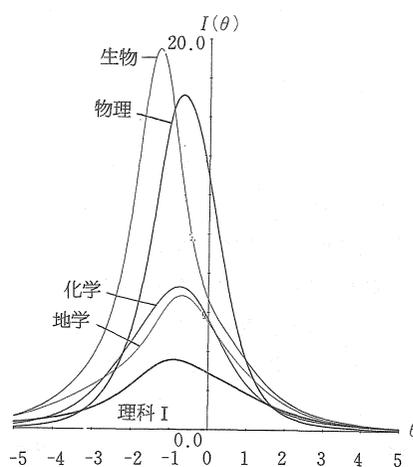


図13. 等化後の理科各科目のテスト情報関数(昭和61年度)

は低い(図13)。共通項目である理科 I は最も低いピークを示し、今年度の共通項目として利用した数学 I とは際立った差異を見せている。また、いずれの曲線のピークも学力 θ が-1近傍に集まっており、同程度の学力を識別するテストであることが分かる。

以上の結果から、今年度との比較により、科目を選択する集団の学力の平均は安定しているが、科目の難易差は実施される年度によって異なる傾向にあることが推測された。

8. 討議とまとめ

8.1 学力について

項目反応理論に基づいて、「理科」の各科目に関するテスト特性の比較分析を試みた。その際、項目反応理論で最も重要な点は、1次元の潜在特性あるいは学力を仮定している点にある。学力は実際には目では見えないものであり、観測できない潜在変数である。我々は、その学力を実際に観測するテストの得点によって推定する。しかし、そのテストが測定している学力が完全に1次元であるとの保証は得られない。この点については、因子分析や本研究で用いた方法等によっておおよその検証を行うことができるが、100%の保証を与えるものではない。実際、当然のことではあるが、「理科」の各科目は本研究で用いた数学 I と「理科」に関する学力以外に、科目独自の学力に関する因子もあるはずである。また、科目内の大問毎にまとめられた項目間では、それぞれ異なった学力に関する因子も含まれているはずである。例えば、石塚・山田(1987)は、化学について、知識、理解、実験、思考力、態度の5つの側面から分析しているが、その結果、いずれの観点も大なり

小なり影響を及ぼしているとの報告もある。

このように、1つの科目内においてさえも、完全な1次元性を保証できないことは、共通第1次学力試験が高等学校の一般的・基礎的な学習の達成度を測定する観点からいえば、数学 I と「理科」の合成テストが完全に1次元を測定していると断定することはできない。本研究は、この点を踏まえ、合成テストを構成する項目間の内部一貫性が高いことを考慮して、あえて「数学 I」と「理科」に関する学力と呼ぶことにした。

分析では、この学力の1次元性を2つの側面から検討を行った。1つは、理論の適合に関する検討であり、他の1つはテストの等化に関する検討であった。その結果、いくつかの項目が外された。除外された項目については、試験問題に関する意見・評価(1987)の中で良問であるとの評価を受けた問題もあるが、多くは何らかの形で批評が加えられていた。その点から言えば、この項目反応理論を利用した項目分析手法は、問題作成の段階における有用な情報を提供するものと思われる。

分析では、さらに「社会」の4科目(倫理・政治・経済、日本史、世界史、地理)について、項目反応理論の適用可能性について検討を加えた。しかし、共通項目として英語あるいは国語を用いて1次元性の確認の結果、共通項目と科目との間に1次元性があると結論を出すに至らなかった。そのため、「社会」の各科目に関するテスト特性の比較を行わないことにした。図14は、英語を共通項目として用いたとき、1次元性を確認するためにプロットした各科目の項目識別力の散布図を示している。図で明らかかなように、傾斜1の直線からかなり外れる位置に散布しており、英語を加えることによって異なる次元の学力を測定している様子が伺える。

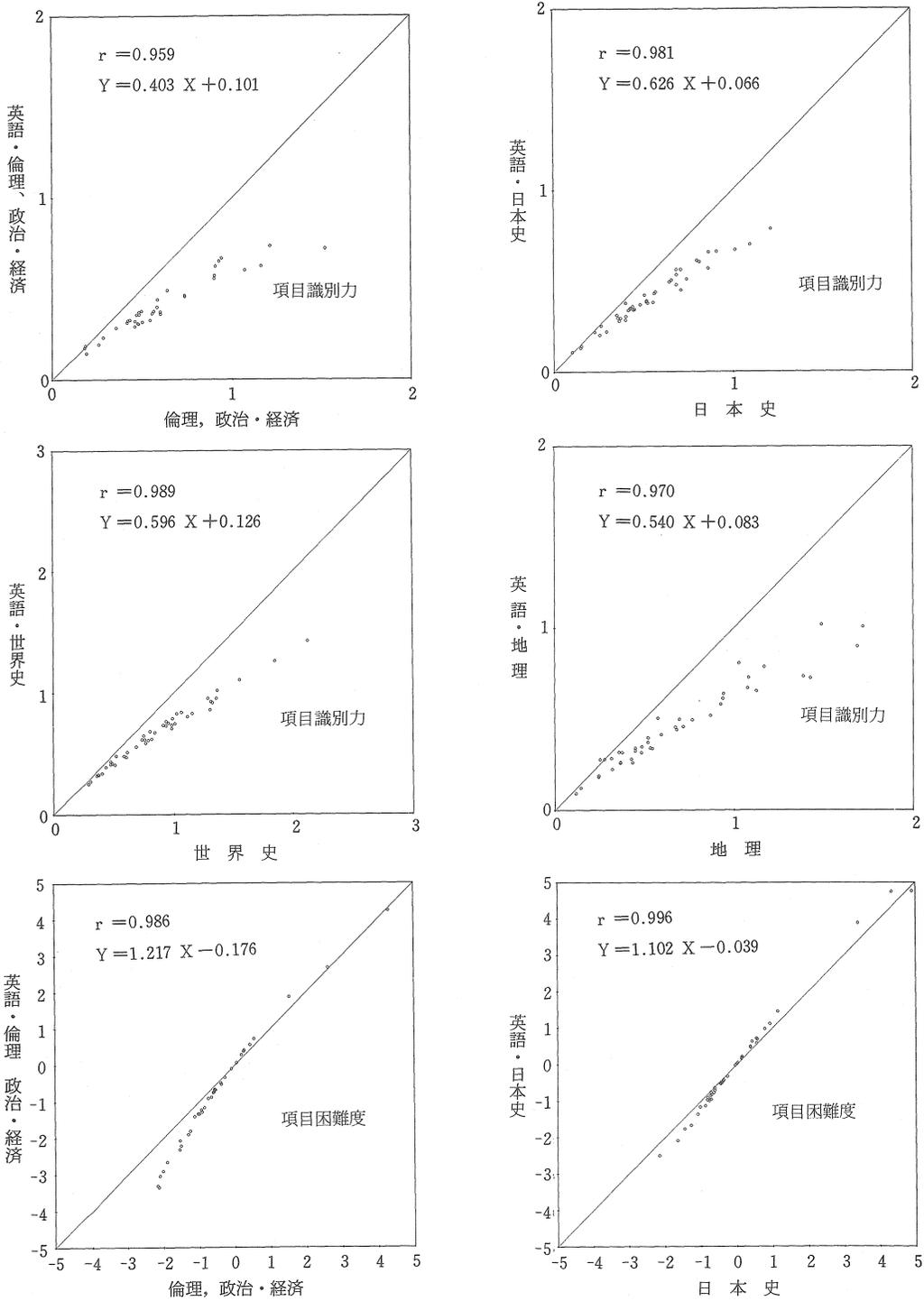


図14. 項目選別後の項目識別力等の相関図

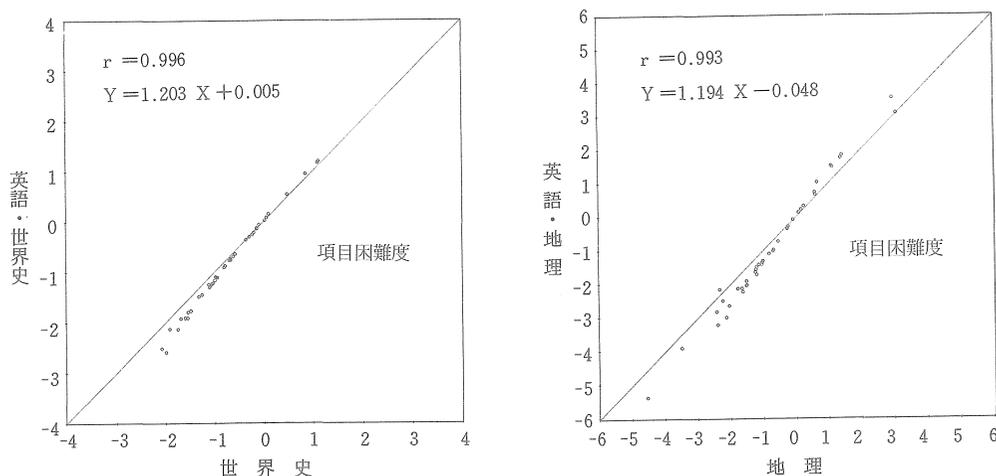


図14. 項目選別後の項目識別力等の相関図 (続)

8.2 テスト特性の比較について

各テストの1次元性を確認しながら、「理科」の各科目に関するテスト特性を比較するため、テストの等化のためのデザインを行い、これに基づいてテストの等化を行った。その結果、同一尺度上で比較可能な項目特性曲線、テスト特性曲線、項目情報関数、テスト情報関数を得ることができた。以下、これらのテスト特性から得られた知見をまとめてみる。

- (1) 各科目の項目特性曲線群は、各科目の特性を反映してか、科目によって傾向が異なっている。数学Iはある学力層を鋭く識別する科目であり、物理、化学もその傾向のある科目であることが分かる。数学Iは、その識別を学力の低い層 ($\theta = -1.60$) で、物理は中位の層 ($\theta = -0.50$) で、また、化学は最も高い層 ($\theta = 0.02$) で行っている。これに対し、生物と地学は数学Iや物理に比べ識別力の小さな項目からなっており、学力差の識別は小さい。特に、識別力が0.3より小さな項目がいくつかあるが、これらの項目が影響したものと思われる。
- (2) テスト特性曲線により学力 θ の尺度上の関数としてテストの難しさを表現でき、ま

た学力の分布を定めることができた。従来、テストの難易については集団全体の要約値で議論されることが多いが、この点きめ細かにテストの難易を比較することが可能である。例えば、今年度の化学は他の「理科」の科目に比べ難しかったと言われているが、テスト特性曲線を見ると、 $\theta < 0.5$ では確かに最も難しいが、 $\theta > 0.5$ では逆に地学が最も難しく、化学が最も易しくなっている。このように、化学が難しい科目であったとしても、学力の違いによってその難しさは異なることに留意する必要がある。

- (3) テスト情報関数や項目情報関数を得ることができ、各科目がどのような学力を識別しようとしているか検討することができた。その結果、情報量が最大になる θ の値に対応するテスト特性曲線の値を調べることにより、「理科」の各科目は、共通第1次学力試験の得点にして50点から55点を得る者の学力を識別するテストであることが分かった。

項目反応理論は、大問主義からなる日本のテストには適していないかもしれないが、テ

ストの1次元性を確認しながらテストの等化を行い、「理科」の各科目のテスト特性をきめ細かに比較することができた。テストの比較を行うといった場面はよくみかける。テストを比較する際には、先に述べた条件が揃ってこそ、正しく判断が下せるのであって、比較を行う際の状況を考慮に入れずに行うべきではないと考える。本研究は、その一つの手続きを提示したつもりである。

謝辞

本研究を遂行するにあたり多くの方から御意見を頂きましたが、特に、大学入試センター研究開発部の石塚助教授には心からお礼申し上げます。また、データの整理に協力していただきました清水恵美子さんには厚く感謝いたします。

参考文献

大学入試センター(1987) 昭和62年度共通第1次学力試験の試験問題に関する意見・評価
石塚智一・山田文康(1987) 共通第1次学力試験の試験問題の分析—昭和57年度「化学I」を中心として— 大学入試センター研究紀要 No.16
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 第17号
清水留三郎(1983) 共通第1次学力試験の「社会」と「理科」の選択科目における差異の統計解析 大学入試センター研究紀要 No.6
塗師 武(1982) Tuckerの方法による共通第1次学力試験の「社会」と「理科」の選択科目間の等化 大学入試センター研究紀要 No. A3
野口裕之(1983) 被験者の推定尺度値を利用した潜在特性尺度等化法 教育心理学研究 第31巻 第

3号 48—53
前川眞一・鈴木規夫(1988) 独立に推定されたテスト項目パラメタの共通尺度上での等化法 大学入試センター研究紀要 No.17
Baillie, R. 1980, *GETAB: A computer program for estimating item and person parameters of the one- and two-parameter logistic model on the PLATO system*. Urbana:University of Illinois, Computer-based Education Research Laboratory.
Bejar, I.I. 1980, *A Procedure for Investigating the Unidimensionality of Achievement Tests Based on Item Parameter Estimates*. Journal of Educational Measurement, Vol 17, No.4 283-296.
Dorans, N.J. & Kingston, N.M. 1985, *The Effects of Violations of Unidimensionality on the Estimation of Item Response Theory Equating of the GRE Verbal Scale*. Journal of Educational Measurement, Vol 22, No.4 249-262.
Haebara, T. 1980, *Equating Logistic Ability Scales by a Weighted Least Squares Method*. Japanese Psychological Research. Vol 22, No.3 144-149.
Hambleton, R.K. & Swaminathan, H. 1985, *Item Response Theory - Principles and Applications*. Kluwer • Nijhoff Publishing.
Lord, F.M. 1980, *Applications of Item Response Theory*. Lawrence Erlbaum Associates, Inc.
Lord, F.M. & Novick, M.R. 1968, *Statistical Theories of Mental Test Scores*. Reading Mass: Addison-Wesley.
Stocking, M.L. & Lord, F.M. 1983, *Developing a Common Metric in Item Response Theory*. Applied Psychological Measurement, Vol 2, No.2. 159-173.
Tukey, J.W. 1977, *Exploratory Data Analysis*. Addison-Wesley.

Norio SUZUKI* Shin-ichi MAYEKAWA** (1988). Comparison of Test Characteristics among Academic Subjects in Natural Science of the Joint First-Stage Achievement Test in the 1987 Academic Year based on Item Response Theory, *Res. Bull. Nat. Cent. Univ. Ent. Exam.*, No. 17, 219-247

Comparison of test characteristics based on Item Response Theory (IRT) was performed using the 1987 academic data of the Natural Science test in the Joint First-Stage Achievement Test (JFSAT). Since applicants were required to take only one of the four academic subjects in Natural Science, namely, Physics, Chemistry, Biology, Geology, IRT equating is needed to transform parameters on the common scale to compare each academic subject.

For this purpose, for each of the four academic subjects, two sets of item parameters were estimated: first one is based on the original items of a academic subject, and the other is based on the compound set of items such as Physics and Mathematics I, and, Chemistry and Mathematics I. Using the compound set item parameters, all the items included in the four academic subjects were placed on a common scale by IRT equating. The unidimensionality of the four academic subjects were tested by comparing the item parameters from the first sets and the ones on the common scale, and some items were deleted as misfitting items.

The test characteristics of the four academic subjects were compared on the common scale. The main findings are;

- (1) The main trend of Item Characteristic Curves (ICC) were different among four academic subjects. Mathematics I discriminates sharply the low ability group (lower than -1.60) and Physics discriminates the middle ability group (near -0.50) and Chemistry discriminates the high ability group (higher than 0.02). Whereas Biology & Geology have lower discriminating power than other subjects.

* Information Processing Section, Research Division, The National Center for University Entrance Examination

** Scholastic Aptitude Section, Research Division, The National Center for University Entrance Examination

- (2) As the result of the comparison of test characteristic curves (TCC), it was found that, although it is said that Chemistry is the most difficult test among the four academic subjects in the Natural Science, the TCC of the Chemistry test has the trend resulting in higher scores for higher ability students and vice versa.
- (3) From the comparison of test information function curves, the ability corresponding to the peak of the information curves were between -1.5 to 0.0 , which translate to the proportion-correct scores from 50% to 55%.

Key Words

item analysis, item response theory (IRT), IRT equating, unidimensionality

