

## 7 国語における IRT 分析及び学力特性値の推定

### 7.1 本調査の国語に関する IRT 分析上の留意点

この章では全国的な学力調査における多段階 IRT モデルの導入可能性について国語データにもとづき検討する。その際、注意すべき点として、

- 1) 問題の出題形式とその採点手続との区別を明確にすること、
- 2) 国語の記述問題に限らず数学の証明問題や理科のパフォーマンス・アセスメントなどへの適用も視野に入れること、
- 3) 今回の国語の調査デザインが BIB デザインではないこと、
- 4) 国語の性質上、局所独立の仮定が必ずしも満足されていないこと、

の4つが指摘できる。

まず、第1点であるが、出題形式がいかなるものであれ、その結果が正誤パターンで表現されていれば基本的には通常の IRT モデルが適用可能であることを考えれば分かりやすい。解答プロセスを評価することでより高度で複雑な学力 (higher complex achievement) をみようとする場合、従来の採点手続きでも部分点を与えることで、その問題に対する得点を数段階に分類してきた。これは評価の観点ごとに重みをつけて結果として部分点を付与する手続きとみなすことができる。それに対して、多段階 IRT モデルはそれと同じ情報を利用して、部分点ではなくて尺度値をもとめるもう1つの採点手続きとみなすことができる。また多段階 IRT モデルを導入することで、従来の部分得点では見積もることのできなかつた項目特性や尺度値の推定誤差など、様々な統計量を手にすることができるようになり、学力調査にふさわしい多彩な分析が実現できる。

次に、出題形式と採点手続きの分離を明確にすることで、国語の記述式に限らず、数学の証明問題や理科の実験などのいわゆるパフォーマンス・アセスメントへの適用の可能性が広がる点である。採点基準やルーブリックなどの整備と採点者の訓練は必須であることは従来型の採点であっても同様であるが、問題形式のバリエーションを増やすことで児童・生徒の所産としての、いわば「能動的な学力」の測定を重複テスト分冊法においても実現できる。

ただし、第3の点として、今回のテストデザインが不釣り合い型不完備ブロック計画 (BIBD) になっていないことには注意が必要である。本来ならすべての問題が生徒集団に対して偏りなく配置されているべきところが、今回は、実施スケジュールの関係上、図 7.1.1 のように、ある分冊がある学校に集中して配布されるような分冊配置デザインを採用した。なお、図 7.1.1 はどの学校の生徒がどの分冊を受けた

かをわかりやすいように、学校と分冊の順序を入れ替えている。この図からも明らかなように、たとえば学校 16 から 20 は分冊の K2, K3, K5, K6, K4B, K7B をまったく受けていない。逆に学校 2 から 5 は分冊の K2, K3, K5, K6 を受けているがそれ以外はまったく受けていないデザインになっていることがわかる。ただし等化デザインの観点からいえば共通項目・共通受検者を含むデザインとなっているので最終結果を大きくゆがめることにはならないと判断できる。

		分冊									
		K2	K3	K5	K6	K4	K7	K4B	K7B	K1	K8
学 校	1										
	2	■	■	■	■						
	3	■	■	■	■						
	4	■	■	■	■						
	5	■	■	■	■						
	6										
	7										
	8										
	9										
	10										
	11										
	12										
	13										
	14										
	15										
	16										
	17										
	18										
	19										
	20										

■ : 生徒がいる分冊 □ : 生徒がいない分冊

図 7.1.1 分冊と実施対象校の配置

最後の局所独立の仮定についてもIRTモデルの前提から言えば大切な仮定ではあるため、そこから現実のデータがどれだけずれているのかなどは慎重な判断が必要である。局所独立の仮定を満たしていない典型的な状況は、ある問題を解くのにその前の問題の解答が必要な場合などが指摘できる。そのため、具体的には、各問題を独立に解ける、いわゆる小問積み上げ方式がこの仮定をみたすのに最適な方式とされる。ところが、国語の問題では題材文を与えて、それに対応する小問を出題することが多いため、厳密にこの仮定を満たすことは困難である。ただし、実用的には大問に含まれる小問を必ずセットにして分析するなどの配慮をすればクリアできる課題であることは経験的に知られていることでもある。そのため本調査においては各ブロックを大問形式で作成し、いわばブロック単位で分冊を構成するなどの工夫を行っている。

なお、比較の対象としては本調査における数学を利用する。これは過年度調査研究でも実証的に考察したように教科の性質上IRTモデルに良く適合し、本年度調査結果においても同様に分冊ごとの信頼性係数がほとんどの分冊で0.9を越えていることなど、重複テスト分冊法の適用例として非常に好ましい結果を得ているため、それとの比較で国語における適用例の特徴を明らかにしやすいからである。

分冊	K1	K2	K3	K4	K4B	K5	K6	K7	K7B	K8
平均	0.000	-0.038	-0.100	-0.085	-0.126	-0.005	-0.034	-0.173	-0.119	-0.080
標準偏差	1	1.037	1.111	1.197	0.869	0.927	1.122	0.927	0.932	1.166
信頼性係数	0.606	0.511	0.664	0.76	0.683	0.573	0.666	0.562	0.509	0.723
受検者数	274	257	255	170	62	246	244	162	60	224
変換後平均	0.063	0.032	-0.020	-0.008	-0.042	0.059	0.035	-0.081	-0.036	-0.003
変換後SD	0.835	0.866	0.928	1.000	0.726	0.774	0.937	0.774	0.779	0.974

表 7.1.1 国語の分冊に関する基礎統計量

分冊	S1	S2	S3	S4	S5	S6	S7	S8
平均	0	0.093	0.126	-0.068	0.034	0.113	0.198	-0.069
標準偏差	1	1.118	1.055	0.992	1.088	0.993	1.121	0.983
信頼性係数	0.906	0.898	0.903	0.907	0.918	0.903	0.905	0.904
受検者数	256	257	252	251	243	238	233	256
変換後平均	-0.048	0.035	0.065	-0.108	-0.017	0.053	0.129	-0.109
変換後SD	0.892	0.997	0.941	0.885	0.971	0.886	1.000	0.877

表 7.1.2 数学の分冊に関する基礎統計量

表 7.1.1 及び表 7.1.2 にそれぞれ国語と数学の各分冊に関する様々な統計量が示されている。両者を比較して顕著なのは、すでに述べたが、国語の信頼性に比べて数学の信頼性がおしなべて高いこと（図 7.1.2）、また、実施スケジュールの関係上、分冊ごとの人数が数学に比べて国語の方がはるかに不揃い（図 7.1.3）であることがわかる。

数学については昨年度調査研究で実施した問題が過半を占めたため、項目母数の値がすでにわかっている項目がほとんどであり、また、項目母数の数値が芳しくないものは入れ替えるなどの対応をし、事前に分冊ごとの信頼性などを見積もることができた。その結果、見込み通りに、いずれの分冊もほぼ所期の信頼性を得ることができた。これは項目非公開でかつ IRT モデルを適用できることによるメリットを十分に生かし切れた成果である。それに対して国語はリーディング・リテラシー型の問題を含めるなどの調査研究グループにとっては初めての経験であったこと、当然のことながら本年度初めての出題であったため数学のように項目母数の値がまったく情報としてないことの二つの要因が重なり、数学ほどには高い信頼性係数の推定値が得られていない。経年比較などの実現を想定したとき、いかに事前にそのような項目母数の値を確定し、そのもとでテストを構成する必要があるかなどの重要性がこのことからわかる。

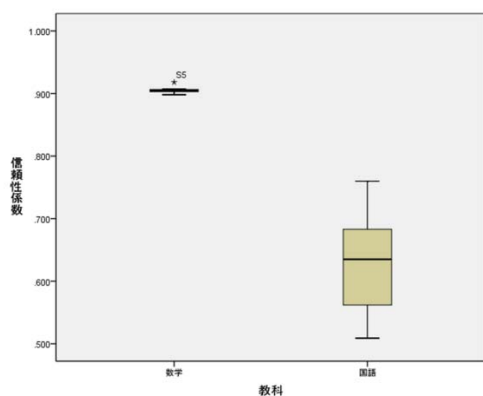


図 7.1.2 分冊ごとの信頼性係数の比較

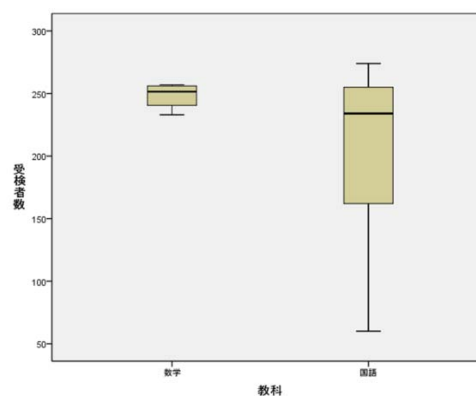


図 7.1.3 分冊ごとの受検者数の比較

同様に昨年度調査で、分冊を配布前に順番ごとに繰り返しセットし、それを対象クラスの数に応じて、いわば輪切り状態にして配布したにも関わらず、分冊番号が後ろになるほど配布枚数が減少するという現象が見られた。そのため本年度は数学においてはセットの際の準備に工夫を凝らし、そのような現象が起こらないようにした。その成果が数学には如実に表れたということができる。しかしながら、国語では先に述べたようにそのようなデザインをとらなかった上、実施スケジュール上、途中で問題の一部を変更するなどの対応をしたため、数学のように分冊間でほぼ同数の受検者数を確保するという望ましい結果を得ることができなかった。今後、国語の実施において改善すべき重要な点の1つである。

次に図 7.1.4 と図 7.1.5 は分冊ごとに受検者集団の学力特性の平均と標準偏差が異なるというモデル（多母集団 IRT モデル）によって推定された母集団ごとの平均と標準偏差を箱ひげ図で表したものである。なお、比較しやすいように、学力特性を表す尺度が間隔尺度の性質（原点と単位が任意に設定できる）を利用して、標準偏差に関してはそれぞれの教科内で最大の標準偏差をとる母集団の標準偏

差が1，すべての母集団平均の平均がゼロとなるように調整していることに注意が必要である。調整前の平均と標準偏差は表 7.1.1 と 7.1.2 の該当欄，変換後の値はそれぞれの表の下の2段に示されているものである。

ここでも事前に項目統計量がわかっていて，その情報を手がかりにテストの構成がやりやすかった数学において，平均の散らばりが国語よりも広く，逆に SD の値が国語に対して高い値のところでは狭い範囲になっていることがわかる。数学の項目母数が分冊間で等化済みであること，ならびにいずれの教科でも BIB デザインになっていないため受検者集団の差が平均に反映されていること，分冊の信頼性の高さを考え合わせると，国語に比して数学の方がより測定精度の高い調査となっているとすることができる。

それに対して国語の方は数学に対して相対的に分冊ごとの SD の広がり具合が大きい一方，平均が数学に比してあまり大きく散らばっていない。もし，完全な BIB デザインをとっているならば平均が狭い範囲でしか広がっていないことは偏りのないサンプリング結果が得られたことを意味するが，再三繰り返しているように今回の調査は BIB デザインではない。そのためむしろ集団ごとの平均が数学のように広い範囲で散らばって当然と予想できるにもかかわらず，このようになったということは，分冊の信頼性係数の推定値の低さを考え合わせれば，数学ほどの測定精度は得られていないということが妥当な結論となる。

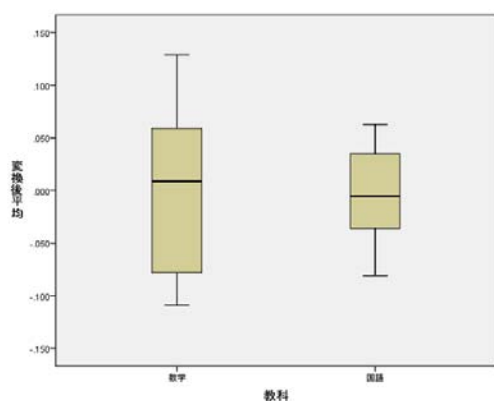


図 7.1.4 分冊ごとの平均の散らばりの比較

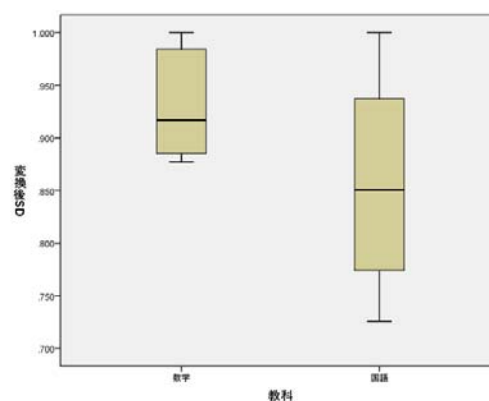


図 7.1.5 分冊ごとの SD の散らばりの比較

## 7.2 多値 IRT モデルの実際

資料 3.2 に公開されているブロック 10 の項目 4 は表 3.3.3.2 に示したように採点基準では正解が 2 段階に分類されている。すなわち、解答を満たすべき条件として、

A) 本文において「控えめな敬語」は「よろしいでしょうか」ではなく「よろしかったでしょうか」である、ということについて正しく理解して解答している。

B) 図のデータから必要な根拠を正しく抜き出して解答している。（「変な言い方だ」と思っている人が多い。「ていねいだ」と思っている人が若い人に少ない。など）

の 2 つを設け、正答としては、

類型 1：条件を全て満たしているもの。

類型 2：A について触れずに、B を満たしているもの。

のような 2 段階の採点となっている。2 値データの場合にはこの 2 つの段階はいずれも正解とみなされる。その結果を項目反応曲線の形で表したのが図 7.2.1 のグラフである。図 7.2.1 は 2 値データから得られたものであり、横軸に学力特性値  $\theta$ 、縦軸に正答確率がとられている。右上がりの曲線が正答確率が  $\theta$  の増大とともに上昇していく様子を、また右下がりの曲線が誤答率が下がっていく様子を表している。

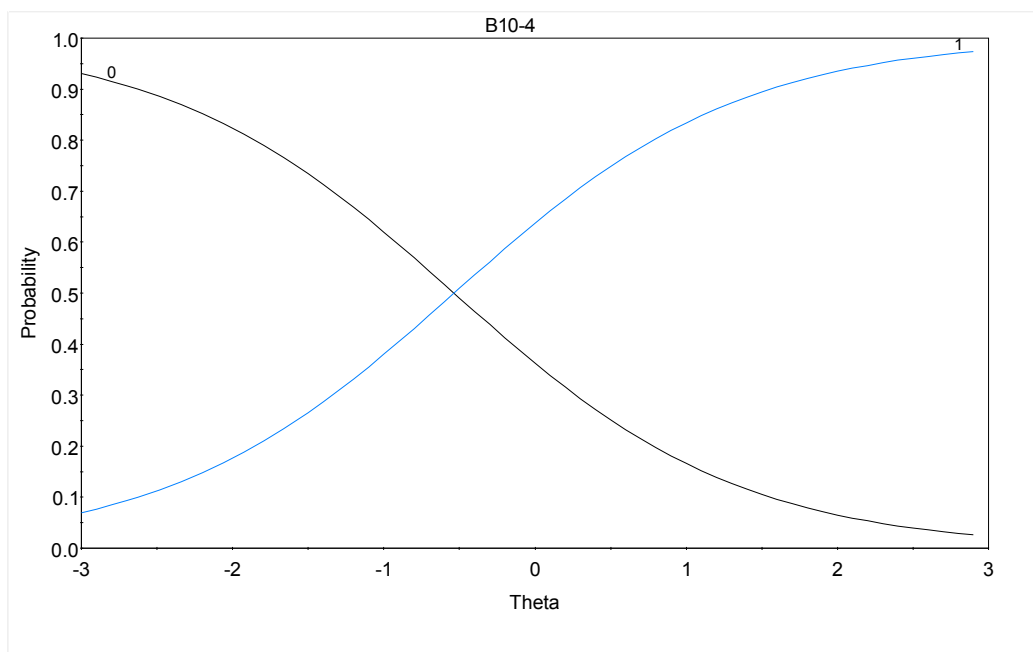


図 7.2.1 2 値反応データの場合の項目特性曲線 (B10-4)

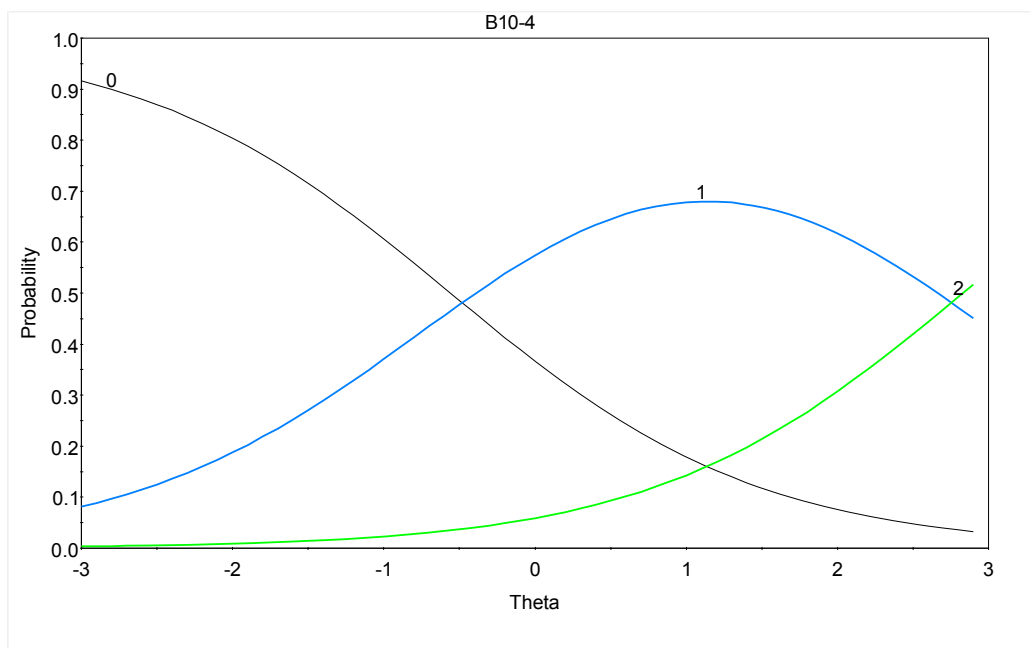


図 7.2.2 多値反応データの場合の項目特性曲線 (B10-4)

一方、図 7.2.2 では、まずこの問いに誤答する確率を右下がりの曲線（曲線の番号 0）で表現し、最初の正答カテゴリー（類型）に採点される確率が山なりの IRCCC（項目反応カテゴリー曲線：第 2 章 2 節参照：曲線の番号 1）として表現され、その次のカテゴリー（類型）に採点される確率が右上がりの IRCCC（曲線の番号 2）として表されている。たとえば  $\theta$  が 1 の付近では誤答する確率とカテゴリー 2 で正答する確率が 0.15 程度でほぼ等しいが、値自体は低い、しかし、そのあたりでカテゴリー 1 で正答する確率は 0.7 程度となっている。このことから、平均から 1 標準偏差程度離れた学力特性値をもつ生徒はこの問題ではカテゴリー 1 の解答をする可能性が一番高いことを示している。また、多値の場合の二つの IRCCC の和が 2 値の場合の項目特性曲線となっていることもわかる。誤答確率をトレースする曲線は両者で一致していることもこのモデルの特徴である。

さらにそれぞれの場合の項目情報曲線を描いたのが次の図 7.2.3 及び図 7.2.4 である。2 値の場合は正答確率が 0.5 となる  $\theta$  で項目情報が最大となっている一方、多値の場合は二つのカテゴリーに応じてその曲線がなだらかに続き、2 つめのカテゴリーに対応する情報量が換算される形で  $\theta=2$  のあたりから若干の増加が見られる。言い換えれば多値型の採点を行うことで 1 つの項目で精度を保ちながら 2 値型でカバーできるよりも広い範囲の学力特性値を扱うことが可能となるといえよう。

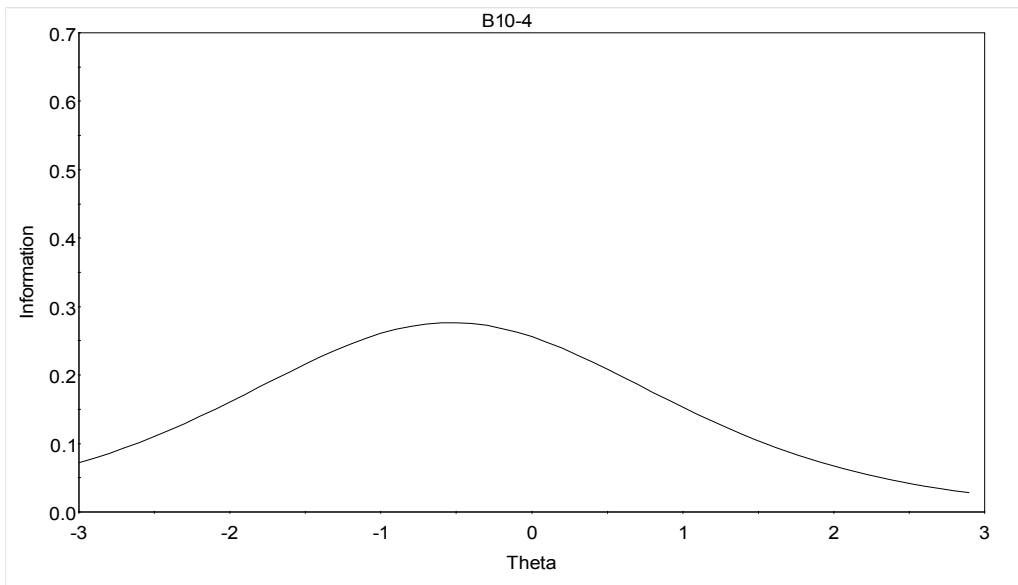


図 7.2.3 2 値データの場合の項目情報曲線 (B10-4)

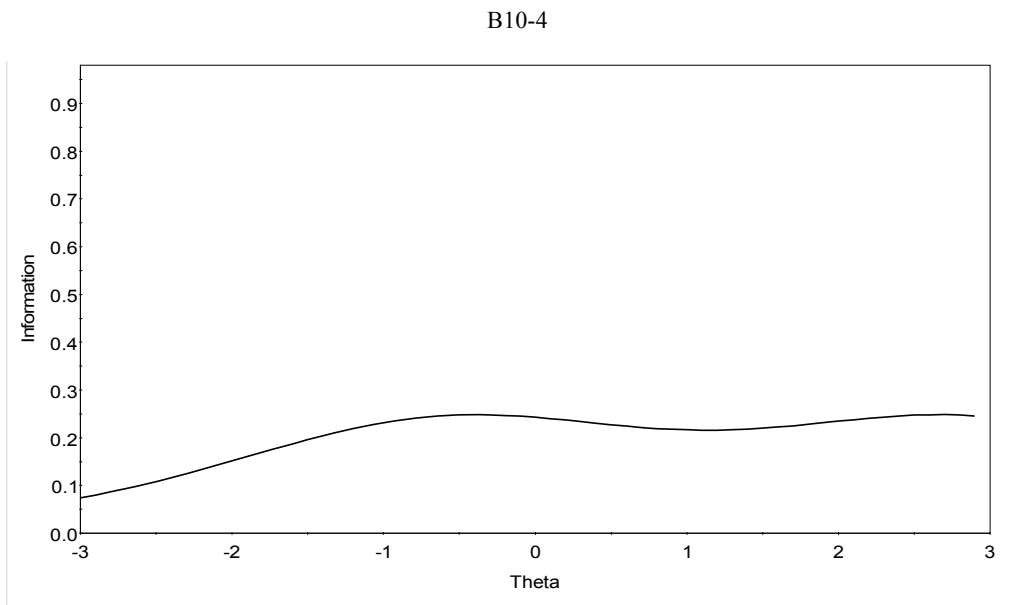


図 7.2.4 多値データの場合の項目情報曲線 (B10-4)

### 7.3 項目母数の推定結果

国語に関する IRT 分析の結果は資料 2.7.1 から資料 2.9 に掲載したとおりである。国語の場合は項目母数の推定の際に、生徒の解答に対して正誤情報のみの 2 値反応パターンからの場合と採点ルーブリックに基づく多段階の多値反応パターンによる場合の 2 通りの推定値をえた。ただし多値反応パターンであっても想定した基準に該当する解答が実際には存在しなかったり、きわめて少数であったりしたものは結果的に段階を統合するなどした。その結果、実施前には 5 段階を想定していた B1-3, B7-2, B12-4 は 4 段階へ、B8-4, B12-4B は 3 段階へ、B1-4, B2-2 は 2 段階へ変更した。そのため多段階モデルでの困難度母数の推定数は 4 段階へ変更となった項目については 3 個、3 段階へ変更した項目は 2 個、2 段階への変更をおこなった項目については正誤反応となるため 1 個である。ルーブリックを設けての多段階での採点には常にこの問題が本質的に発生するため、問題を事前に何らかの形で試行しデータにもとづき実証的に調整しなければならないことに注意が必要である。

図 7.3.1 には 2 値反応パターンから得られた困難度と識別力の推定値の散布図が示されている。識別力が 0.2 以下のものはブロック 7 の項目 2 及び項目 4 である。項目分析図をみてもこの 2 つの項目は国語の学力を識別しているとは言いがたく、改良する余地がある。同じことは多値パターンを使って推定した結果からも指摘できる(図 7.3.2)。なお図 7.3.2 ではわかりやすいように最初の位置母数のみを使って表示している。ここでも項目 7-2 はかなり異常な値を示していることが明確である。

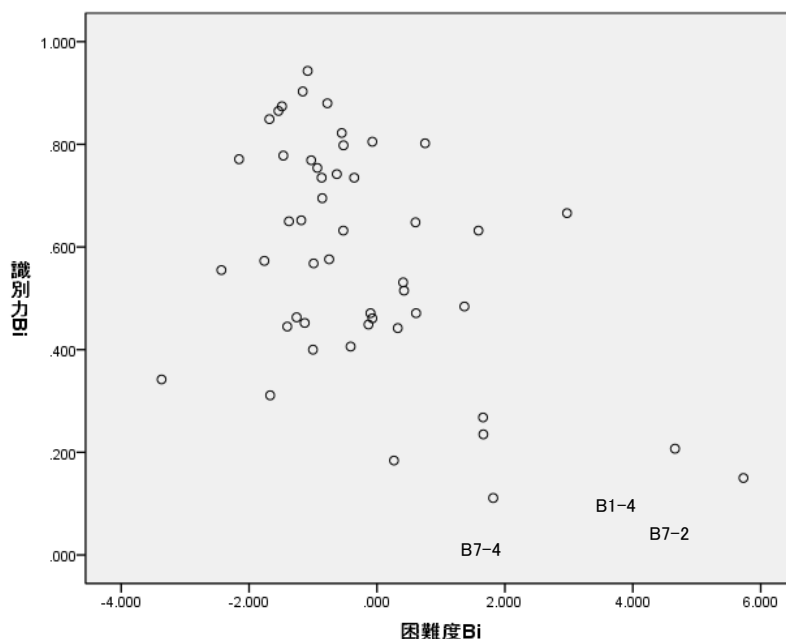


図 7.3.1 困難度と識別力の散布図 (2 値)



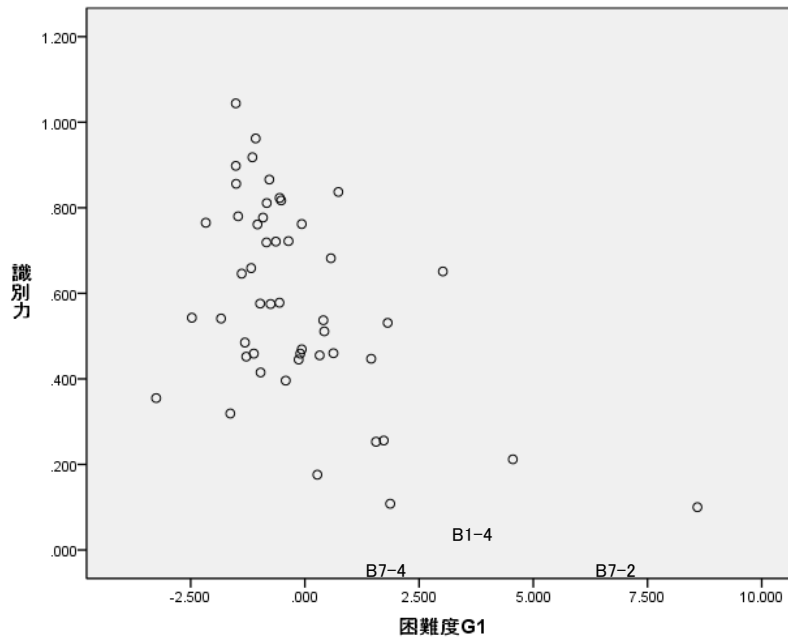


図 7.3.2 困難度と識別力の散布図 (多値)

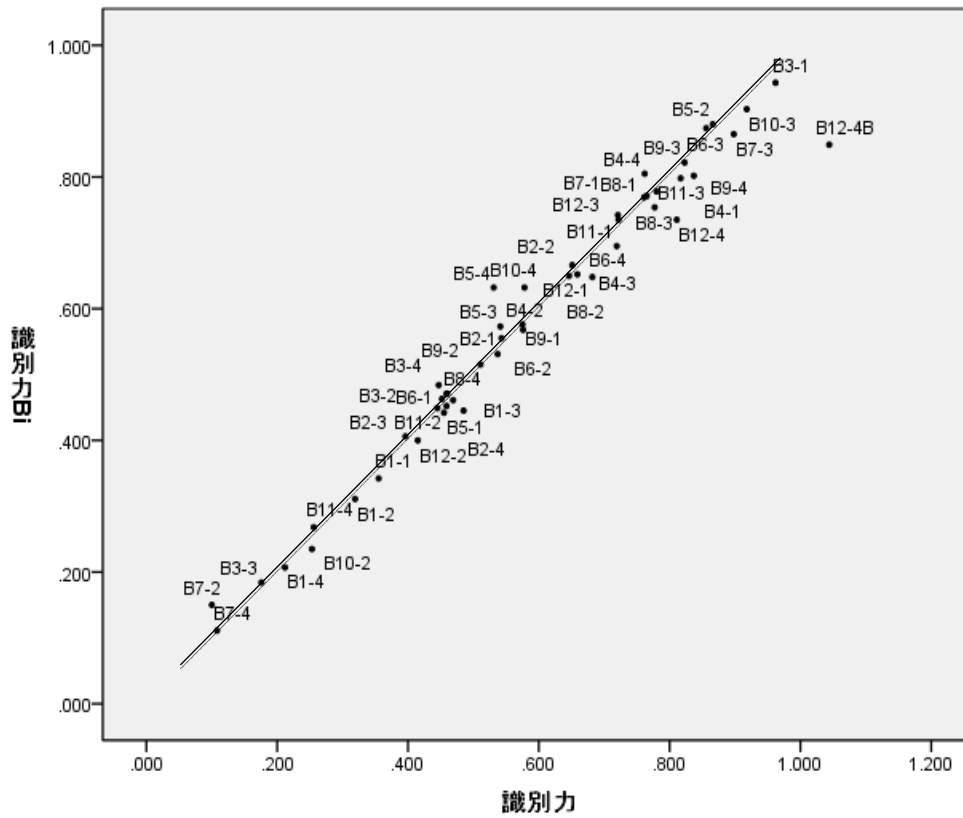


図 7.3.3 多値モデルと2値モデルにおける識別力の散布図

次に多値 IRT モデルと 2 値 IRT モデルのもとで得られた項目母数の比較を行ったのが、図 7.3.3～図 7.3.5 である。図 7.3.3 では横軸に多値 IRT モデルで求めた識別力母数を、縦軸には 2 値 IRT モデルで求めた識別力母数をとっている。項目 B12-4B など除いてほとんど直線上に乗っていることがわかる。

また図 7.3.4 には同じく横軸に多値 IRT モデルからえられた位置母数と縦軸に 2 値 IRT モデルから得られた困難度母数をとった散布図が描かれている。比較しやすいように多値 IRT モデルから得られた位置母数は最初のものだけを使った。位置母数が複数あるものは図 7.3.5 に分けて示した。

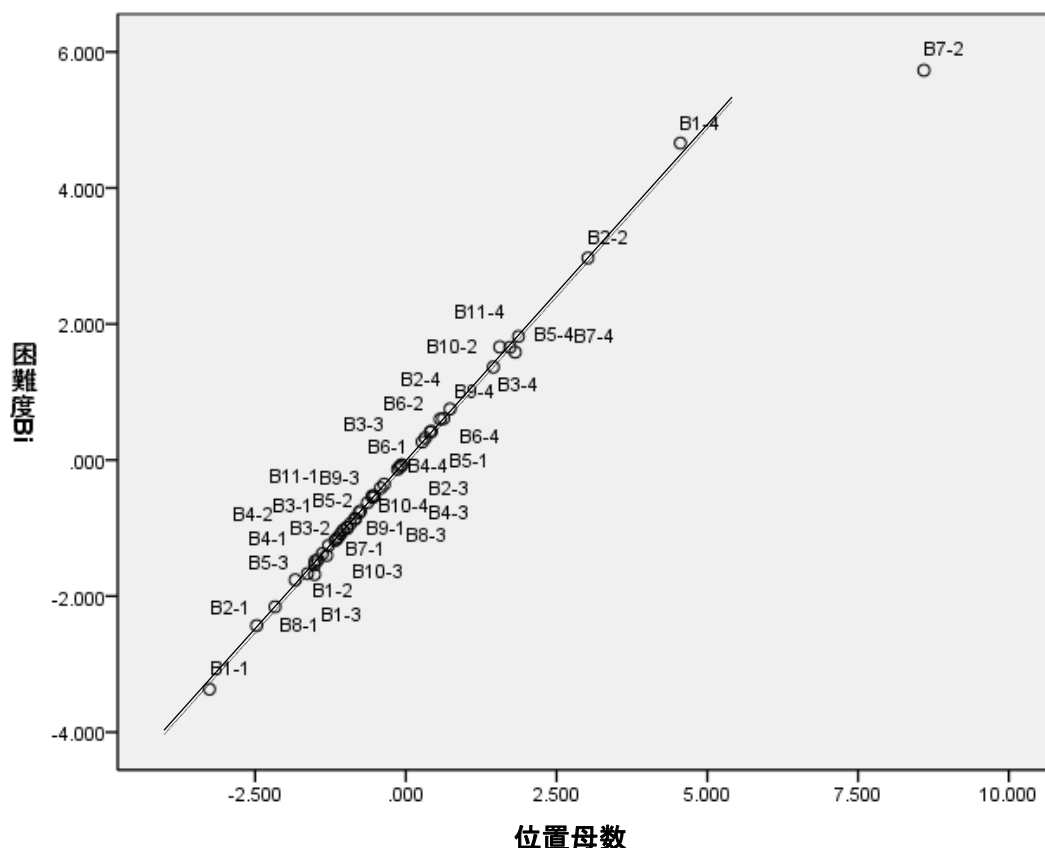


図 7.3.4 多値モデルと 2 値モデルにおける位置母数の散布図

最初の位置母数に関しては項目 B7-2 以外はほぼ直線上に分布していることがわかる。また、異常値を示す項目 B7-2 を除く、B1-3, B4-4, B12-4 の位置母数と困難度の関係を示した図 7.3.5 を観察すると位置母数 1 に関しては B1-3, B12-4, B4-4 の順で図 7.3.4 と同様に大きくなっているが、位置母数 3 に関しては B1-3 のものが 2.0 を越えている。その理由としては、この問題が「理由を 45 文字以上 60 以内で書きなさい」というタイプであったため、採点基準をすべて満たすのにはかなりのハードルが高くなったことが考えられる。このような部分点間の差をどう見積もるかについても事前のデータが得られていれば合理的な配分ができ、ひいてはより精度の高い調査が実現できるのである。

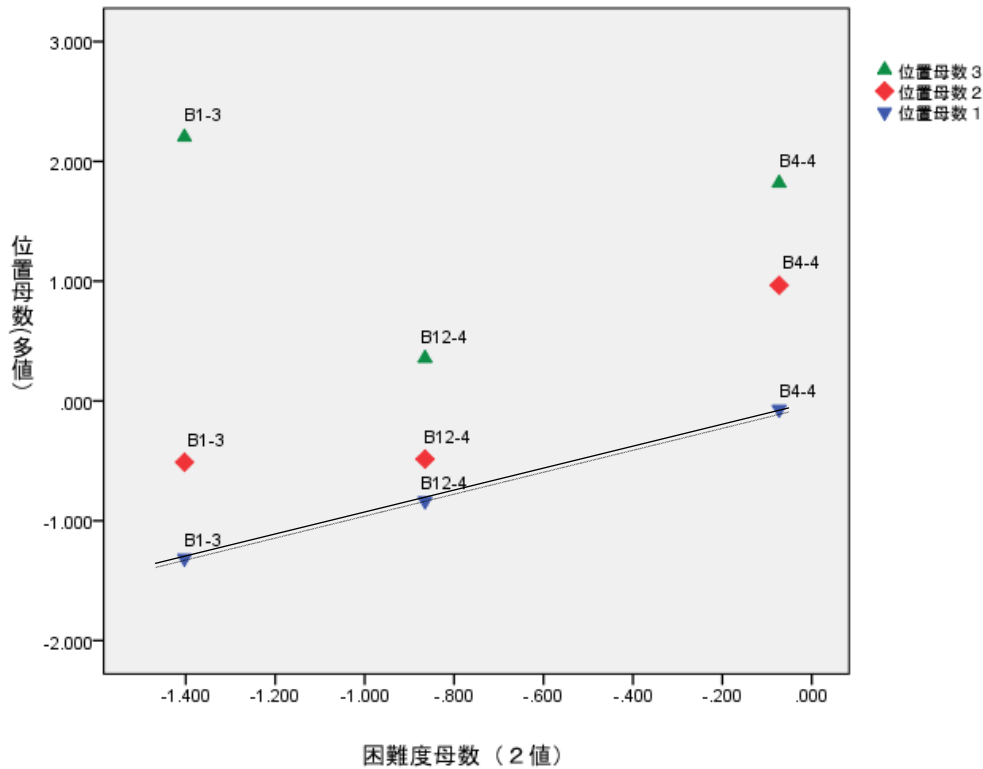


図 7.3 5 困難度母数と位置母数の関係

#### 7.4 推定結果の利用例

次に項目特性曲線の形状から判断して改良可能な項目の例を示す。項目 B7-1 から B7-4 がそれに該当する。改良した問題を再度実施してその振る舞いを実証的に確認する予定のため問題自体は非公開としたが、その項目特性曲線を項目ごとに描くと図 7.4.1 及び図 7.4.2 のようになった。なお、図は左から順に B7-1、B7-2 (図 7.4.1)、B7-3、B7-4 (図 7.4.2) である。

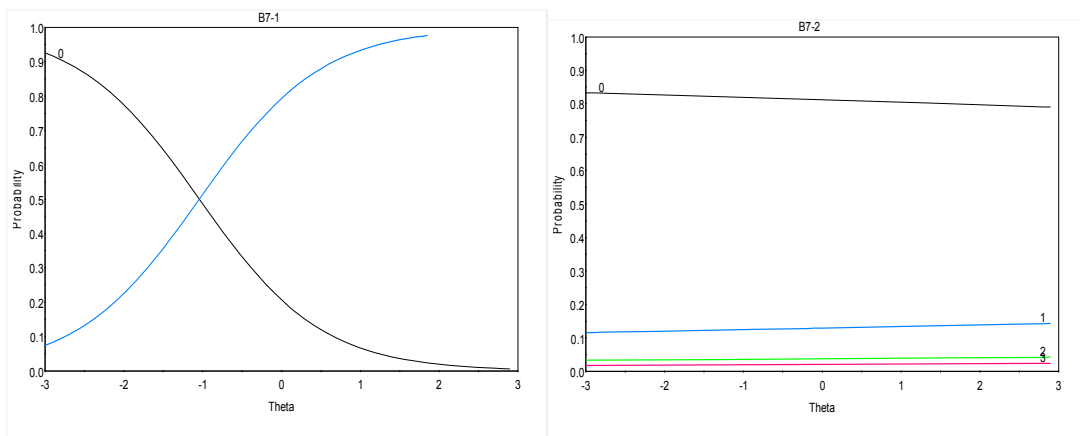


図 7.4.1 B7-1 と B7-2 の項目特性曲線

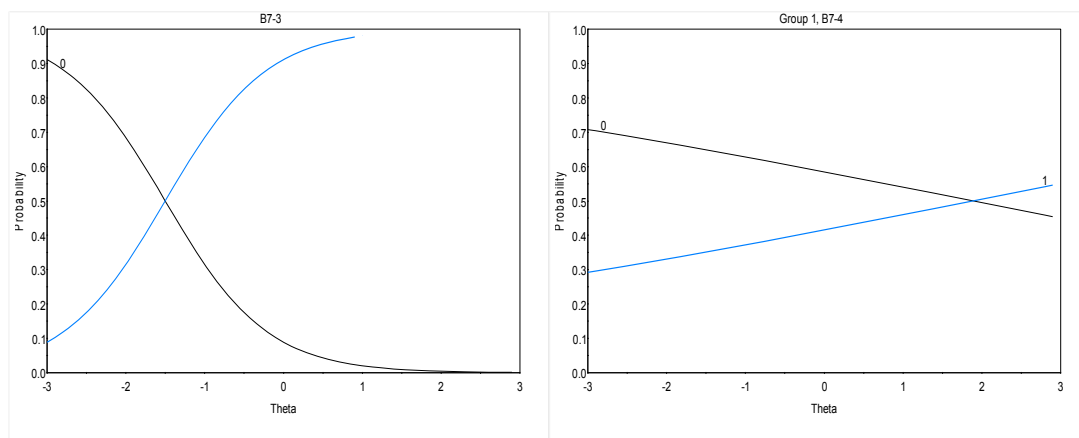


図 7.4.2 B7-3 と B7-4 の項目特性曲線

項目 B7-1 ならびに B7-3 は特に問題となるような形状はしていなく、学力特性値が上昇するにつれて、正答確率も上昇し、逆に誤答確率は減少している。ところが、項目 B7-2 はすべての IRCCC ともほとんど変化がなく、かつ誤答確率を示す番号 0 の曲線のみが突出して上の位置にある。これは項目 B7-2 がなんらかの原因があって非常に難しい問題となり、その結果、ほとんどの生徒が答えられなく、識別力が落ちてしまったと考えることができる。同様のことは B7-4 にも当てはまる上、B7-4 では多枝選択形式において受検者を不必要に混乱させるという理由からなるべく使わないようにする「何々でない」ものを選ばせる形式をあえて使ったことも大きな理由である。実際、項目特性曲線は正答確率が 0.3 から 0.6 の間に入っていて識別力はそれほど大きくない。おそらく、「否定」で問われたために予想通り生徒を混乱させてしまったことや、わかりにくさが原因となって生徒に考えさせる動機を喚起できずに、結果として当て推量を引き起こしてしまったなどの可能性が指摘できる。可能ならばその点を改良した問題にし、項目特性がいかに改善されるかの実証的なデータを収集することとしたい。

## 7.5 小改訂の影響

ここでは B12-4 及び B12-4B の結果を利用して、同じ大問の下のある小問を、形式的にはそれとほぼ同等の小問に入れ替えた場合の多値 IRT モデルにおける項目特性の動きを分析する。B12 では資料 3.3 に示すように題材文で「土用の丑の日」の売り上げが伸びる品物を示した後、その内容をどれだけの確に理解しているかをみるためにそれと同様の品物として B12-4「チョコレート」及び B12-4B「和菓子」の 2 つのパターンの小問を作成した。これを実施して得られた結果を図にしたのが次の図 7.5.1 及び図 7.5.2 である。

B12-4 のチョコレートの問題の方が和菓子の問題に比べて難易度が高くなり、かつ  $\theta$  の広い範囲にわたって情報量が高いことがわかる。逆に和菓子の問題は難易度が低いが、その分、低い  $\theta$  の部分で情報量が大きくなっていることがわかる。対象とする集団の学力の測定精度をどのあたりで一番高くするのかといったようなこともこれらの結果から判断できる。このように同じ構造をもった小問であってもカバーできる範囲がことなることを実証的に示せることが IRT モデルを用いることの利点の 1 つである。

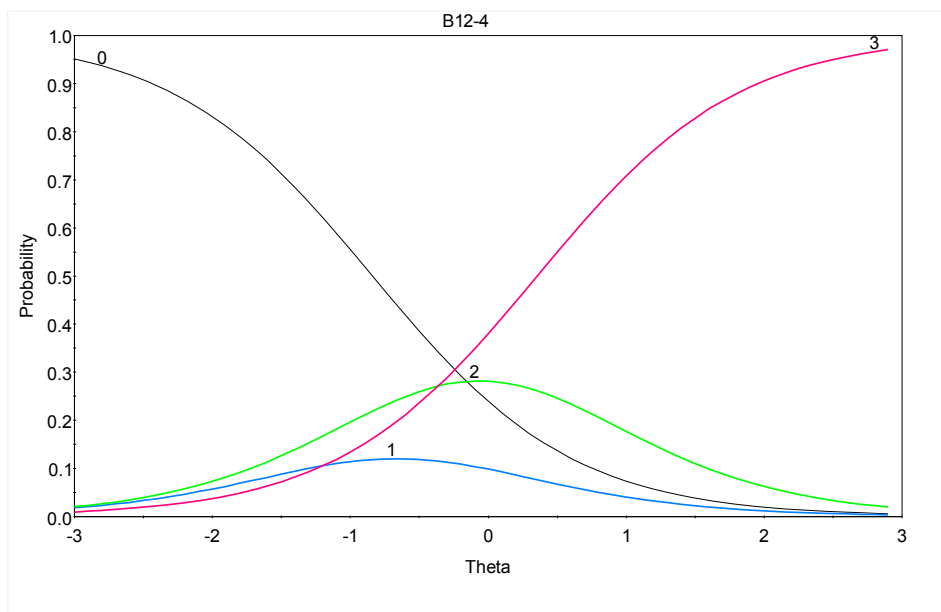


図 7.5.1 B12-4 「チョコレート」の IRCCC

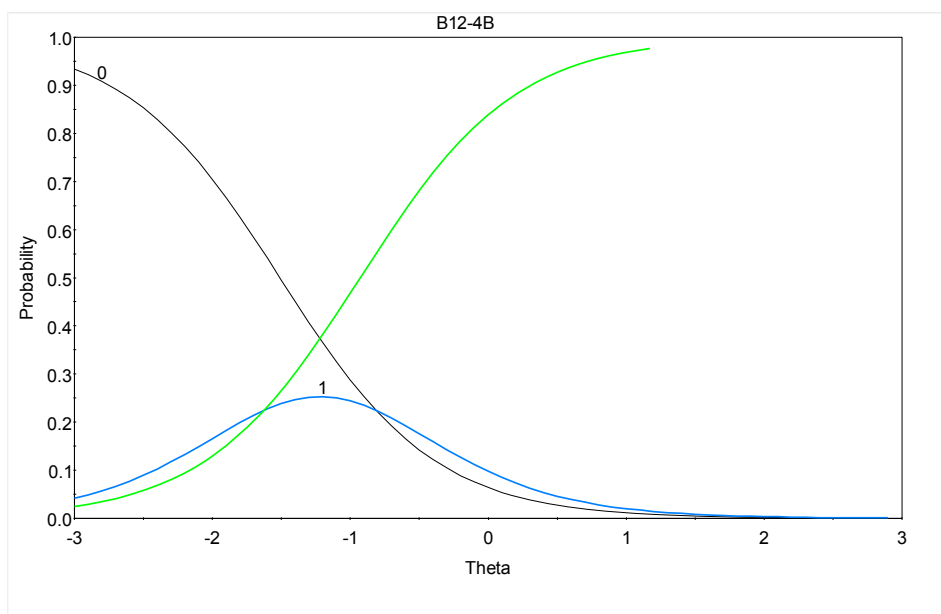


図 7.5.2 B12-4B 「和菓子」の IRCCC

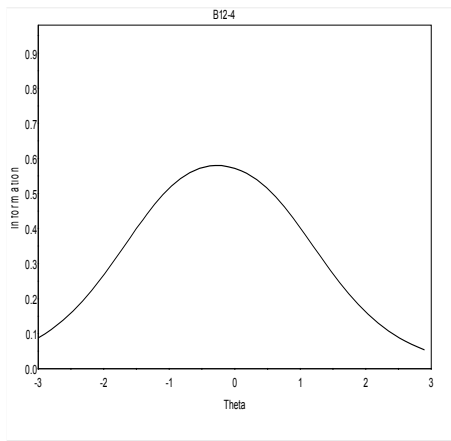


図 7.5.3 B12-4 の項目情報量曲線

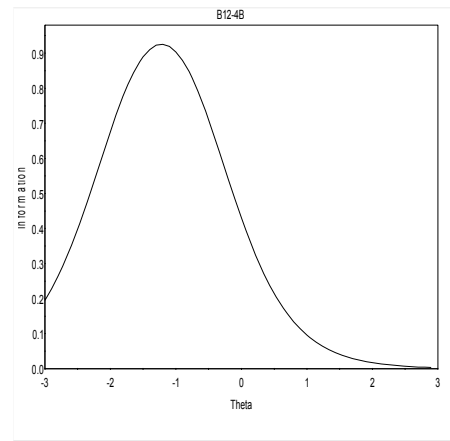


図 7.5.4 B12-4B の項目情報曲線

## 7.6 テスト情報量からの検討

本章の最後にこれらの問題をすべて使った場合のテスト情報曲線を図 7.6.1 またテスト特性曲線を図 7.6.2 に示す。テスト情報量は  $\theta$  が -1 付近で最高になり、点線で表された標準誤差が一番低くなっていることがわかる。調査の場合、 $\theta$  が高い部分でももう少し情報量が高い方が学力の形成要因などを分析する際には好ましいと考えられる。そのため問題を入れ替えるなどして  $\theta$  の高い位置での測定精度の確保に工夫が必要であろう。

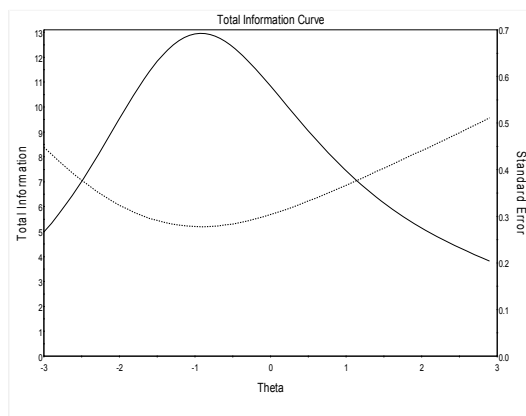


図 7.6.1 テスト情報曲線

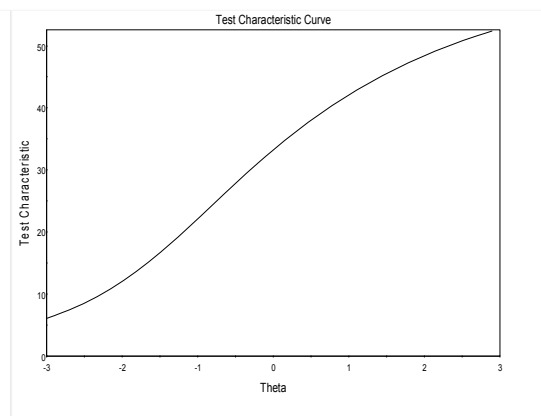


図 7.6.2 テスト特性曲線

## 8 数学と国語の相関について

本章では、従来の一冊子一斉方式のテストよりも幅広い領域が調査できる重複テスト分冊法の利点を活かし、実施したすべての国語問題の特徴を数学の学力特性をはじめとする外的な属性情報を用いて分析した結果を分析例の1つとして示す。これらの分析方法は、外的属性を質問紙調査等で得られた背景情報に置き換えることで、学力の形成要因などの有益な情報を得る手段となるものである。

### 8.1 多次元尺度法による国語のブロックの特徴

国語の問題で使用された文章を類似度に基づいて分類するために、心理学系の学部生及び大学院生9名を対象に、問題冊子の分類実験を実施した。参加者には、分類刺激としてB01～12の問題冊子を1部ずつ渡し、B01とB12の題材となっている文章（以下、文章と略記）を基準として、残りの問題冊子で使用された文章を、類似度の観点からB01のグループとB12のグループのいずれかに分類するよう教示した。その際、いずれにも分類できない冊子があれば新たにグループを設けてよいとし、小問を考慮せずに文章のみに基づいて分類するよう求めた。なお、文章を分類する目的から、同一の文章が使用されたB12とB12Bについては、B12のみを分類刺激として用いた。結果は表8.1.1に示すとおりである。ブロック1とブロック2の文章が同じ分類と見なされた頻度は0、ブロック3とは9名のうち5名が同じグループに分類したことがわかる。他のブロック対の数値も同様に解釈できる。

この結果を多次元尺度法(Multidimensional Scaling; MDS)により分析した。具体的にはSPSS/ALSCALを用い、上で得られた頻度データを類似度とみなし、尺度水準を順序尺度と指定して計算を実行した。その結果得られた座標行列及びそれから描かれた文章の類似度マップを図8.1.1に示す。またディスプレイの刺激座標による説明率は次元数が2でRSQ=.988ときわめて良好な適合度であった。図8.1.1において距離が近いものほど文章としての特徴に類似性があることを表している。

次元1は横方向に大きく2つに分かれており、B01, 02, 03, 04からなるグループと、残りのB05～B12からなるグループが見られた。さらに、B05～B12のブロックは縦方向に3グループに分かれ、B05, 07, 09からなるグループ、B06と08からなるグループ、B10, 11, 12からなるグループが見られた。このことからわかるように題材となる文章は大きく文学的文章と説明的文章の2つの群に分類され、その度合いも他の外的情報による区分に比べかなり大きな要因となっている。リーディング・リテラシーの定義からもこの特徴は重要である。

表 8.1.1 国語のブロックの分類結果 (頻度)

国語	B01	B02	B03	B04	B05	B06	B07	B08	B09	B10	B11
B02	0										
B03	5	5									
B04	7	7	2								
B05	9	9	9	8							
B06	9	9	9	8	3						
B07	9	9	9	8	1	4					
B08	9	9	9	8	5	4	4				
B09	9	9	9	8	0	3	1	5			
B10	9	9	9	8	6	5	7	5	6		
B11	9	9	9	8	6	5	7	5	6	3	
B12	9	9	9	8	6	5	5	3	6	1	3

表 8.1.2 SPSS/ALSCAL による座標行列

ブロック	次元1	次元2
B01	2.092	-0.489
B02	2.092	-0.489
B03	2.088	-0.488
B04	1.191	-0.280
B05	-1.033	-0.270
B06	-0.971	0.176
B07	-1.024	-0.160
B08	-0.959	0.231
B09	-1.033	-0.267
B10	-0.800	0.714
B11	-0.800	0.714
B12	-0.843	0.607



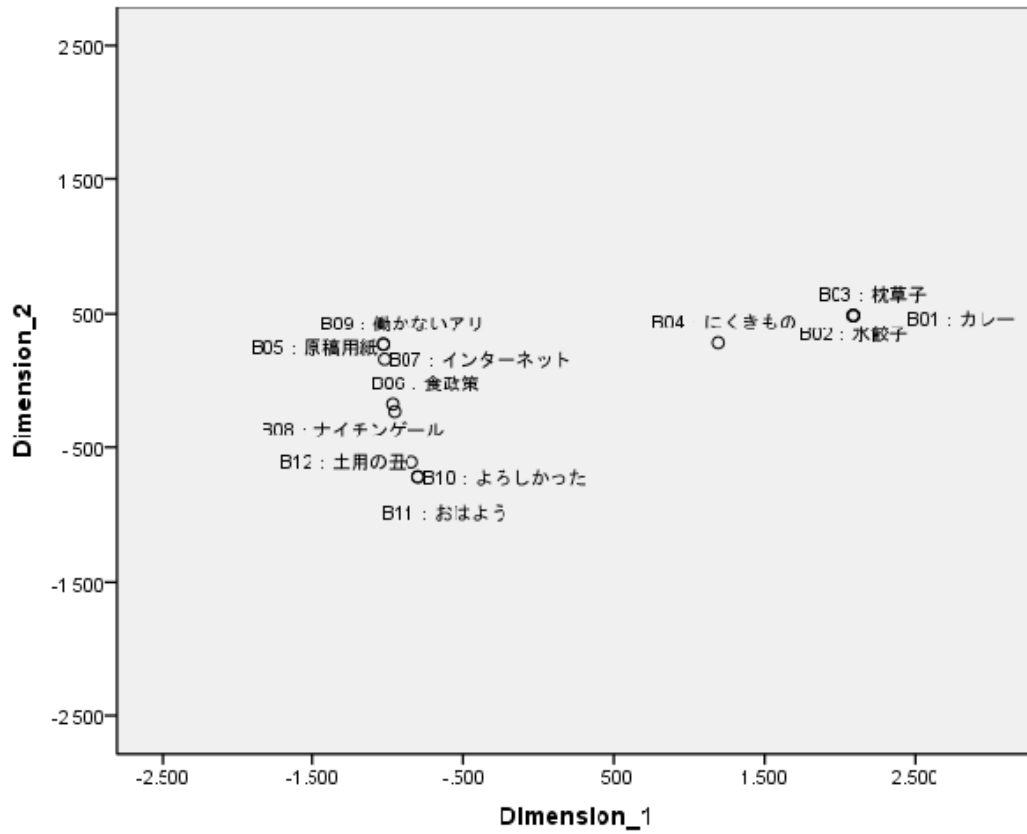


図 8.1.1 国語問題で使用された文章の類似性マップ

各文章の類似度を、文章の特徴に基づいて解釈するために、図 8.1.1 の座標行列を独立変数とし、文章の特徴（説明的文章、文学的文章、図の有無）及び、「国語のブロックごと正答数と数学 $\theta$ との相関係数」を従属変数として重回帰分析を行い、外的情報の方向をベクトルとしてマップの中に埋め込んだ（原理については章末補注を参照）。図 8.1.2 より、説明的文章と文学的文章の矢印が横方向に互いに逆向きに伸びていることから、文学的文章の矢印に近い B01~04 は文学的文章のグループ、説明的文章の矢印に近い B05~12 は説明的文章のグループと解釈できる。一方、縦方向には図の有無の矢印があることから、矢印の先端に最も近い B10~12 は図を含む文章のグループと解釈できる。また、B10~12 は、「国語のブロックごと正答数と数学 $\theta$ との相関係数」の矢印にも近いことから、図を含む文章を読み取る力が、数学の問題を解決する力と関連していると考えられる。

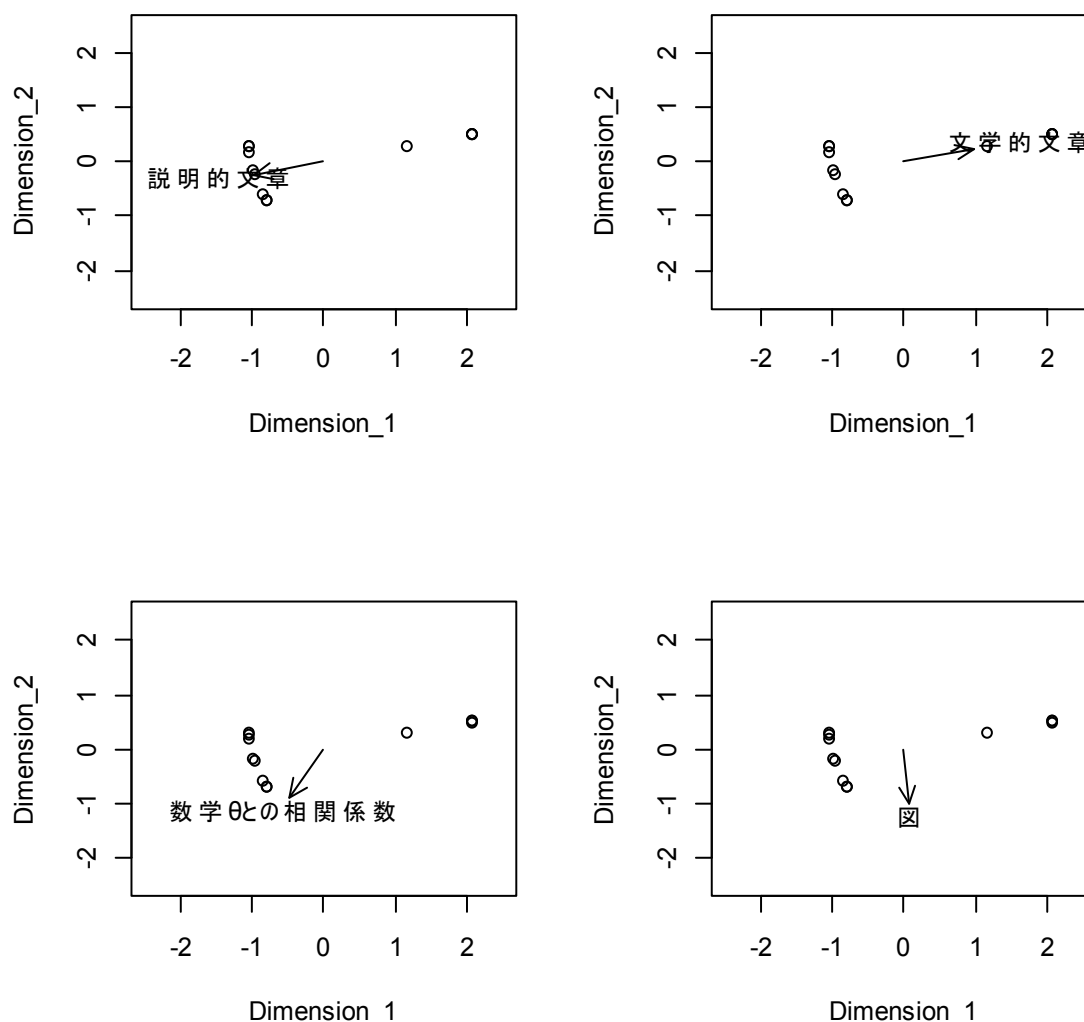


図 8.1.2 文章の特徴及び数学 $\theta$ との相関係数をマップに埋め込んだ結果

次に、各ブロックの文章と、国語問題の定義との関連をみるために、図 8.1.1 の座標行列を独立変数とし、国語問題の定義（情報の取出し、解釈、情報の収集・統合、判断）を従属変数として重回帰

分析を行い、外的情報の方向を表すベクトルの埋込を行った（章末補注参照）。図 8.1.3 より、「情報の取出し」と説明的文章の B05～12 が、「解釈」と文学的文章 B01～04 が関係している。さらに、説明的文章の中でも、B05, 07, 09 は「情報の収集・統合」に関わり、B10, 11, 12 は「判断」に関わる問題であるといえる。

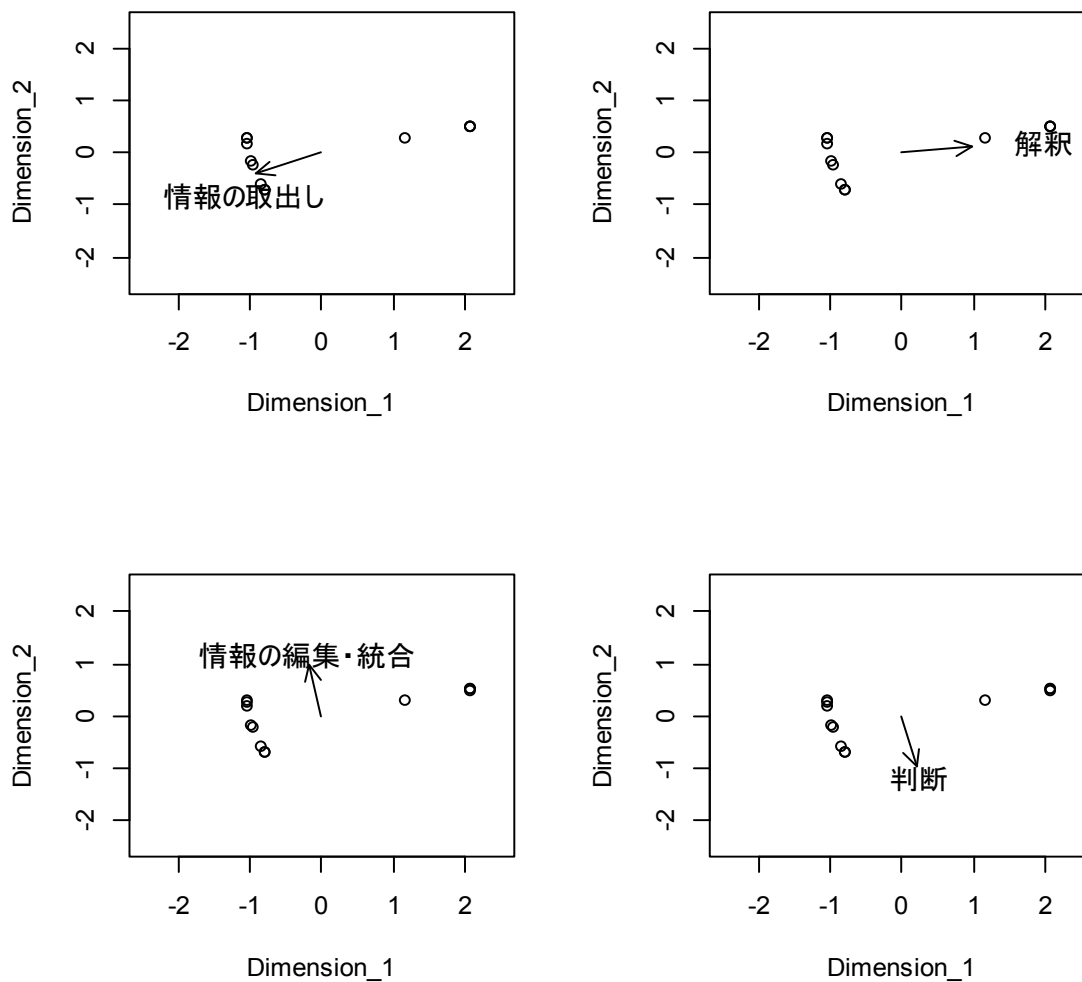


図 8.1.3 国語問題の定義を従属変数とした重回帰分析の結果

## 8.2 数学と国語との相関

本調査研究では、重複テスト分冊法を導入して、多岐に渡る国語能力の特徴を明らかにすることが目的の1つとして掲げられている。そこで、本節では、国語問題の題材や項目の質という観点と、各項目と数学問題から算出された数学の学力特性値  $\theta$  との関係を見ることで、数学能力と関係する国語能力を測定する項目の特徴について検討する。これをさらに一般化すれば、最終的には、学力とその形成要因などの指標との分析への展開に資するものとなることを目指していることに他ならない。

また、第2章3節で述べたように、2003年のPISA調査で示された我が国の特徴として、「熟考・評価」が指導されてこなかった点と「非連続型テキスト」があまり取り入れられてこなかった点が挙げられる。本調査では、「熟考・評価」の側面の項目や、「非連続型テキスト」を題材としたブロックを使用している。本節によって、これらの項目や題材に関連する項目が数学能力とどのような関係があるかという点についても、示唆を与えることができるだろう。

数学の学力特性値  $\theta$  と国語の学力特性値  $\theta$  の相関係数を算出したところ、0.570 と比較的高い値が得られた。これは、本調査研究で作成された国語問題の多くが、リーディング・リテラシーの測定を目的として作成されたものであり、かつ非連続型テキストを題材として扱っていたことが関係しているといえよう。リーディング・リテラシーには、テキスト全体の論理的な構造を把握しながら、テキストの内容を解釈するような能力が含まれる。このようなテキスト全体を俯瞰し、論理構造を把握する能力は、数学にも必要な能力であるといえる。例えば、数学の公式を学習する場合に、公式が導出された論理的な道筋を把握することができることは重要であり、証明問題に回答する場合には、論理的に解を導出し、解法を論理的に証明する力が求められる。

表 8.2.1 に数学  $\theta$  と国語の分冊ごとの正答数との相関係数を示した。受検者数は 60~274 名、項目数は全て 12 項目である。各分冊の相関係数は 0.531 から 0.634 を示しており、比較的高かった。

表 8.2.1 数学  $\theta$  と分冊ごとの国語の正答数との相関係数

分冊	ブロック	$R$	$n$
分冊 1	1,8,9	0.534	274
分冊 2	2,7,10	0.551	257
分冊 3	3,6,11	0.531	255
分冊 4	4,5,12	0.623	170
分冊 4B	4,5,12B	0.586	62
分冊 5	1,5,10	0.536	246
分冊 6	2,6,9	0.589	244
分冊 7	3,7,12	0.580	162
分冊 7B	3,7,12B	0.634	60
分冊 8	4,8,11	0.608	224

表 8.2.2 に国語のブロックごとの正答数と数学  $\theta$  との相関係数を示した。ブロックごとの受検者数は、ブロック 12B のみ 122 名と少ないものの、その他のブロックは 330 名から 520 名程度であった。プロ

ックごとに数学 $\theta$ との相関係数を算出すると、その値は0.217～0.605であり、0.2から0.3程度の低い値を示したブロックが複数みられた。各ブロックの項目数は4つと少なかったことから、値の低さは納得できるだろう。また、相関係数がブロックによってばらついていることは注目すべき点である。数学 $\theta$ との相関の高かったブロックと低かったブロックは、それぞれどのような特徴があったのだろうか。

表 8.2.2 数学 $\theta$ と国語のブロックごとの正答数との相関係数

文章の分類			$r$	$N$
B01	文学的文章(小説)	シズコさん (カレー)	0.217	520
B02	文学的文章(小説)	シズコさん (水餃子)	0.307	501
B03	文学的文章(古典)	枕草子	0.319	477
B04	文学的文章(古典)	枕草子 (にくきもの)	0.516	456
B05	説明的文章(連続型テキスト)	原稿用紙	0.476	478
B06	説明的文章(連続型テキスト)	食政策	0.451	499
B07	説明的文章(連続型テキスト)	インターネット	0.324	479
B08	説明的文章(非連続型テキスト含む)	ナイチンゲール	0.402	498
B09	説明的文章(連続型テキスト)	働かないアリ	0.520	518
B10	説明的文章(非連続型テキスト含む)	よろしかったでしょうか	0.484	503
B11	説明的文章(非連続型テキスト含む)	おはようこんばんは	0.463	479
B12	説明的文章(非連続型テキスト含む)	土用の丑の日	0.567	332
B12B	説明的文章(非連続型テキスト含む)	土用の丑の日	0.605	122

まず各ブロックで使われていた題材の違いというものがあるだろう。本調査で作成した問題では12の題材が使われており、それらは大きく4種類に分類することができる。文学的文章にあたる小説の一部抜粋がB01, B02に、古典の一部抜粋がB03, B04に使われている。さらに、説明的文章の抜粋がB05, B06, B07, B09に使われている。これらの連続型テキストに、非連続型テキストと呼ばれる図表やグラフが含まれている説明的文章が使われているのがB08, B10, B11, B12, B12Bである。

このような題材の質という観点からみると、文学的文章、特に小説を題材にして作成したブロックは数学 $\theta$ との相関が低いことがわかる。それに対し、説明的文章やマルチテキストは全体的に数学 $\theta$ と中程度の相関がある。このような題材の質による数学 $\theta$ との相関係数の違いを調べるために、文学的文章(小説)、文学的文章(古典)、説明的文章(連続型テキストのみ)、説明的文章(非連続型テキスト含む)の4つを群にとり、数学 $\theta$ との相関係数について、一要因の分散分析を行った。その結果、群間の差は有意であった( $F(3, 48)=4.827, p<.005$ )。多重比較(Bonferroni法)を行った結果、文学的文章(小説)に比べて、説明的文章(連続型テキスト)及び説明的文章(非連続型テキスト)の数学 $\theta$ との相関係数が5%水準で高く、文学的文章(古典)は10%水準で高い傾向にあった。以上のように、問題に使われる題材の質によって、数学能力との相関が異なってくることを示された。

特に、第2章第3節で述べたように、我が国の国語教育の特徴の1つとして、従来、図やグラフなどの非連続型テキストは数学教育の問題とされることが多かったことが挙げられる。このような背景も関係し、非連続型テキストは数学 $\theta$ との相関が高かったのであろう。

以上のような、国語の問題の特徴に関する考察をさらに進めるためには、各ブロックの小問を使い、それぞれの題材に対してどのような項目が作成され、どの項目が数学θとの相関関係が高かったのかを見ていくことが必要である。

表 8.2.3 に数学θと国語の項目ごとの正答数との相関係数を示した。受検者数はブロック 12B の項目 4 のみ 122 名と少ないが、その他の項目は 330 名から 520 名程度であった。値は 0.011~0.427 と小さかった。こういった数学θとの相関係数の低さに影響を与えている要因として、まず、天井効果及び床効果などが影響していたことが考えられる。例えば、相関係数が 0.011 と最も低いブロック 7 の項目 2 (以下、B7\_02 と略記) は、正答率が 0.19 とかなり低い。データにある程度のばらつきがないと相関関係はみられにくいいため、B7\_02 のように正答率が過度に低いことが影響して、相関関係が見いだせないという場合もある。

また、数学θとの相関の低かった項目と高かった項目を見ていくことで、項目の質という観点から考察することができるだろう。正答率が偏っていた B7\_02 の次に数学θとの相関が低かったのは、0.090 であった B1\_02、0.098 であった B1\_01、B3\_03 である。例として、ブロック 1 の小問 1 を挙げると

問一  で囲んだ部分には、「私」と「母」についての話が描かれています。この部分から読み取れる「母」の人物像について述べてものとして、最も適切なものを次の中から一つ選びなさい。

巻き寿司は、ほうれん草と玉子、しいたけ、ピンクのデンプが入っていた。やたら太かった。酢でしばったふきんで時々包丁をふき、はじっこを手伝う私にくれた。私はそんな時、母さん私が好きなのだろうかと思った、手伝うごほうびのつもりだったのだろうか。  
少なくとも料理をしている時私をどなったりした事はなく、まるで気の合ったチームのようだった。

- 1 巻き寿司の準備を手伝ったときに「私」にはじっこ部分をくれたように、料理を手伝ったときはいつもほめてくれて、「私」の機嫌をとろうとする「母」。
- 2 巻き寿司の準備を手伝ったときに「私」にはじっこ部分しかくれなかったように、料理を手伝ったぐらいでは、たいしたごほうびはくれない厳しい「母」。
- 3 巻き寿司の準備を手伝ったときに「私」にはじっこ部分をくれたように、料理をしているときは「私」に優しく「私」を仲間として認めてくれる「母」。
- 4 巻き寿司の準備を手伝ったときに「私」にごほうびとしてはじっこ部分をくれたが、それ以外のときは、ごほうびなど私たちきょうだいにはほとんどくれたことがない厳しい「母」。

という項目であった。この項目では、著者である「私」の表現から、「私」が「母」に対して抱いていた人物像を読み取るものであった。また、B1\_02 は、

問二 この文章から、「私」の家の食事はどのようなものだったと考えられますか。次の中から最も適切な一つを選びなさい。

- 1 料理上手な「母」が、いろいろな食材を使って、子供たちが喜ぶような料理を毎日作ってくれた。
- 2 料理上手な「母」が、いろいろな工夫をして、料理を作ってくれた。
- 3 料理上手な「母」が、いつもふつうとは違う材料で料理を作ってくれた。
- 4 料理上手な「母」が、限られた食材で見た目に豪華な料理を作ってくれた。

という項目であった。この項目では、文章全体の展開及び作者である「私」の表現から、「私」からみた「母」の人物像を読み取ることが求めていた。このように、ブロック全体を通して相関係数が低い項目が多かった B01 や B02 は、文学的文章である物語文であった。項目内容も場面の展開や表現の仕方から筆者や登場人物の心情を読み取ることが求めたものを中心であり、このような文章の表現や描写から登場人物の心情や考え方を読み取るような問題の場合、数学能力との相関が低いことが示されたと言えよう。

数学との相関が低かった項目は、文学的文章を使ったブロックに含まれていることが多かったのに対し、ブロック内の項目の多くが数学 0 と相関の高かったブロック 10, 11, 12 などは、全てグラフなどの図表を用いたマルチテキスト問題である。特に数学 0 との相関の高かった項目内容の例として、相関係数が 0.427 と高かった B12\_03 と 0.374 であった B12\_4 がある。B12\_03 は、

問三 本文及び図 3 から言えることを、次の中から一つ選びなさい。

- 1 うなぎは年間を通してどの県庁所在地でも一人当たりの年間支出金額が多い。
- 2 うなぎの収穫量が多いところが、うなぎの一人当たりの年間支出金額も多いとは限らない。
- 3 鹿児島県と愛知県とうなぎの収穫高の合計は、全国の 80% を占める。
- 4 うなぎの値段は、奈良市の方が横浜市よりも高い

という問題であり、B12\_04 は、

問四 特定の日に消費が増える品物としてはチョコレートがあります。チョコレートに関する次の図 4 と図 5 からはどのようなことが言えますか。本文を参考にして書きなさい。

という問題であった。注目すべきは、両者が共に図表の読み取り能力が求められる問題になっていた点である。このような図表等の非連続型テキストの読解を要するリーディング・リテラシー的な能力

を測定する項目は、数学能力との関係が高いとされている。B12\_03 の問いにある「図」はうなぎの都道府県庁所在地別の年間支出額の棒グラフ、B12\_04 の問いにある「図」には、月別及び日別のチョコレート の支出割合が折れ線グラフで表現されている。よって、これらの問いに答えるためには、まず、棒グラフや折れ線グラフを読めることが必要であり、すなわち、数学能力が関係してくる。また、グラフを理解できたからと言ってこの問題に正答できるわけではない。グラフから得られた情報を過不足なくまとめ、問いにあわせて編集して回答を作り出すことも必要である。このような「情報の編集・統合」能力も、数学で求められる能力であり、このような問いの特徴が、数学 $\theta$ との相関に影響を及ぼしていたと考えられる。B12\_03 や B12\_04 と同様に、図の読み取り（図からの情報の取り出し）と、図から得た情報の編集・統合が求められる B10\_04 も、数学 $\theta$ との相関が高い。これらの項目ごとの数学 $\theta$ との相関をみていくことで、非連続型テキストの読み取り及び情報の編集・統合が必要な項目と数学能力との関係が指摘できる。このような、項目の目的や内容と数学 $\theta$ との関係について考察を深めるために、表 8.2.4 では国語問題の定義（情報の取出し、解釈、情報の収集・統合、判断）に基づき、各項目の数学 $\theta$ との相関係数の平均値を算出した（表 8.2.4）。

表 8.2.3 小問ごとの数学 $\theta$ との相関係数

ブロック	文章の質	N	項目			
			1	2	3	4
ブロック 1	文学的文章(小説)	520	.098	.090	.173	.114
ブロック 2	文学的文章(小説)	501	.165	.122	.188	.226
ブロック 3	文学的文章(古典)	477	.338	.181	.098	.163
ブロック 4	文学的文章(古典)	456	.281	.327	.298	.404
ブロック 5	説明的文章(連続型テキスト)	478	.233	.350	.235	.330
ブロック 6	説明的文章(連続型テキスト)	499	.265	.285	.293	.270
ブロック 7	説明的文章(連続型テキスト)	479	.360	.011	.305	.039
ブロック 8	説明的文章(非連続型テキスト含む)	498	.240	.239	.292	.214
ブロック 9	説明的文章(連続型テキスト)	518	.281	.254	.397	.386
ブロック 10	説明的文章(非連続型テキスト含む)	503	.222	.223	.355	.345
ブロック 11	説明的文章(非連続型テキスト含む)	479	.378	.224	.385	.104
ブロック 12	説明的文章(非連続型テキスト含む)	332	.354	.265	.427	.374
ブロック 12B	説明的文章(非連続型テキスト含む)	122				.370

表 8.2.3 より、文学的文章のうち、小説を題材にした項目は指導事項の内容に関係なく、全体的に数学 $\theta$ との相関係数が低く、数学能力とあまり関係していないことがわかる。それに対し、古典を題材にした項目は、部分的に相関係数が高く、数学能力と関係していた項目が含まれていたことがわかる。また、説明的文章は、全体的に相関係数が高めである。さらに、表 8.2.4 を概観すると、特に説明的文章（連続型テキスト）の「情報の取り出し」及び「判断」、説明的文章（非連続型テキスト含む）の「情報の編集・統合」「判断」が数学 $\theta$ との相関係数の平均が高めであることがわかる。



このとき、数学 $\theta$ との相関係数の平均としては0.20～0.25程度であるものの、文学的文章（古典）の「情報の編集・統合」「判断」、説明的文章（連続型テキスト）の「情報の編集・統合」のSDが大きい点には注意すべきである。文学的文章（古典）の「情報の編集・統合」にはB3\_04とB4\_02、「判断」にはB3\_03とB4\_04が含まれる。両者ともにブロック3の項目（B3\_04, B3\_03）は数学 $\theta$ との相関係数が低いのに対し、ブロック4（B4\_02, B4\_04）は相関係数が高い（表8.2.3）。ブロックによって、数学と関係が異なっているため、SDが大きかったといえる。このブロック間の相関係数の違いには設問の質が関係している可能性がある。伝統的な国語の問題とされる文学的文章を題材にした場合でも、設問を工夫することによって、国語に特化した能力だけでなく、数学能力と関係するような「情報の編集」や「判断」能力も測定することが可能であることを示した結果といえよう。両者は共に、PISA型リーディング・リテラシーの「熟考・評価」に値するものであり、我が国であまり重視されてこなかった分野であることから、今後更なる検討を行う必要があるといえよう。

また、同様にSDが大きかった説明的文章（連続型テキスト）の「情報の編集・統合」にはB5\_01, B5\_02, B5\_03, B6\_02, B7\_02, B7\_04, B9\_3の7項目が含まれる。表8.2.3をみると、B7\_02とB7\_04が特に相関係数が低い。先に述べたように、B7\_02は正答率がかなり低い記述問題で、資料2.7.1～2.7.3で記載されているようにIRT分析の結果からも適切な項目であったとは言い難い。さらにB7\_04は問題文中で、一般に避けるべきとされる「不適切なものを選べ」という表現を使用している。以上のようことを考慮し、この2項目を外すと相関係数の平均は0.300と高くなり、SDも0.072と小さくなる。IRT分析の結果及び質的な検討から不適切な項目を外すことで、説明的文章（連続型テキスト）の「情報の編集・統合」と数学 $\theta$ との相関係数の平均が高めであることが示された。

以上のSDの大きい項目についての考察を踏まえ、数学能力と相関の高い国語能力について考察をする。第一に、相対的に、文学的文章を用いた場合よりも、説明的文章を用いた場合の方が、国語の項目の正答数と数学能力との関係性がみられることが示された。文学的文章には、一般的に登場人物の発言や情景から登場人物や筆者の心情を読み取る項目が設定されることが多く、先にも述べたように実際に本調査でも心情把握に関する項目が多かった。しかしながら、このような心情把握は数学で使われることが少ない。それに対し、説明的文章ではテキストに書かれた内容を正確に理解することを求める項目が設定されることが多いが、この能力は数学でも必要とされる能力である。例えば、数学の文章題を解くときには、問題文に書かれている内容を正確に理解することが必要である。以上のように、テキストの質によって、項目で測定する能力が異なっていたため、数学能力との関係にも差がみられたのだろう。

第二に、本調査の国語の定義に従ってみると、特に「情報の編集・統合」「判断」の項目と、数学能力との相関が高かった。「情報の編集・統合」とは情報を過不足無くまとめたり、目的に合わせて情報を編集する力のことを指し、「判断」とは、テキストから得た情報を理解し、他の事例に適応したりする力のことを指す。数学でも、例えば証明問題を解く時には、情報をコンパクトにまとめるような「情報の編集・統合」する力が必要である。また、文章題を読み、過去に解いた問題と同様のパターンであるかを判断し、過去の類似した問題の解法を適用させる場合には「判断」する力が必要となる。このように、テキストの理解に留まらず、テキストを活用する力は、国語に限らず、様々な教科でも必要とされる能力といえるだろう。これらの「情報の編集・統合」「判断」はPISA型リーディング・リテラシーの「熟考・評価」に値するが、本調査ではこの「熟考・評価」の能力が国語に留まらず、多様な方面での能力でも重要であるということ、数学能力との比較で示すことができたといえよう。

表 8.2.4 文章の質別にみた本調査の国語の定義に含まれる項目と数学  $\theta$  との相関係数の平均 $\pm$ SD

	本調査における国語の定義			
	情報の取り出し	解釈	情報の編集・統合	判断
文学的文章 (小説)	0.165	0.094 $\pm$ 0.006	0.177 $\pm$ 0.043	0.114
文学的文章 (古典)		0.275 $\pm$ 0.067	0.245 $\pm$ 0.116	0.251 $\pm$ 0.216
説明的文章 (連続型テキスト)	0.300 $\pm$ 0.042	0.280 $\pm$ 0.036	0.221 $\pm$ 0.147	0.329 $\pm$ 0.058
説明的文章 (非連続型テキスト含)	0.283 $\pm$ 0.077	0.296 $\pm$ 0.092	0.305 $\pm$ 0.080	0.301 $\pm$ 0.132

表 8.2.5 文章の質別の本調査の国語の定義に含まれる項目数

	本調査における国語の定義			
	情報の取り出し	解釈	情報の編集・統合	判断
文学的文章 (小説)	1	2	4	1
文学的文章 (古典)		4	2	2
説明的文章 (連続型テキスト)	4	2	7	3
説明的文章 (非連続型テキスト含)	5	4	4	4

続いて、数学との相関係数とブロックごとの信頼性係数との相関関係を示す。図 8.2.1 より、信頼性係数と数学との相関係数の間に右上がりの関係があることがわかる。先に述べたように B01 や B02, B07 は相関係数が小さかったが、これらのブロックは、ブロックごとの信頼性係数も 0.231, 0.340, 0.162 と低かった (表 5.2.2 参照)。一方、相関係数の高かった B04, B09, B12, B12B は、信頼性係数も 0.514, 0.497, 0.464, 0.440 と他のブロックと比べて比較的高い値を示していた。すなわち、ブロックとしてまとまっているものほど、数学との相関も高かった。このことから、両者ともに高いブロックは、多様な国語能力のうち、特に数学と関係するような能力に特化した項目を多く含んでいたブロックであるといえよう。

第 5 章 2 節でも記載した通り、一般に、数学に比べ、国語能力は信頼性が低いという特徴がある。その理由として、国語の能力が多岐に渡ることがあると考えられる。表 5.2.2 及び表 8.2.2 より、本調査で作成したブロックのうち、数学能力との相関が高く、つまり、数学能力にも必要とされるような

国語能力に特化した場合、ブロックの信頼性も高いことがわかった。さらに、そのようなブロックは説明的文章、特に非連続型テキストを含むようなものが多いことが示された。

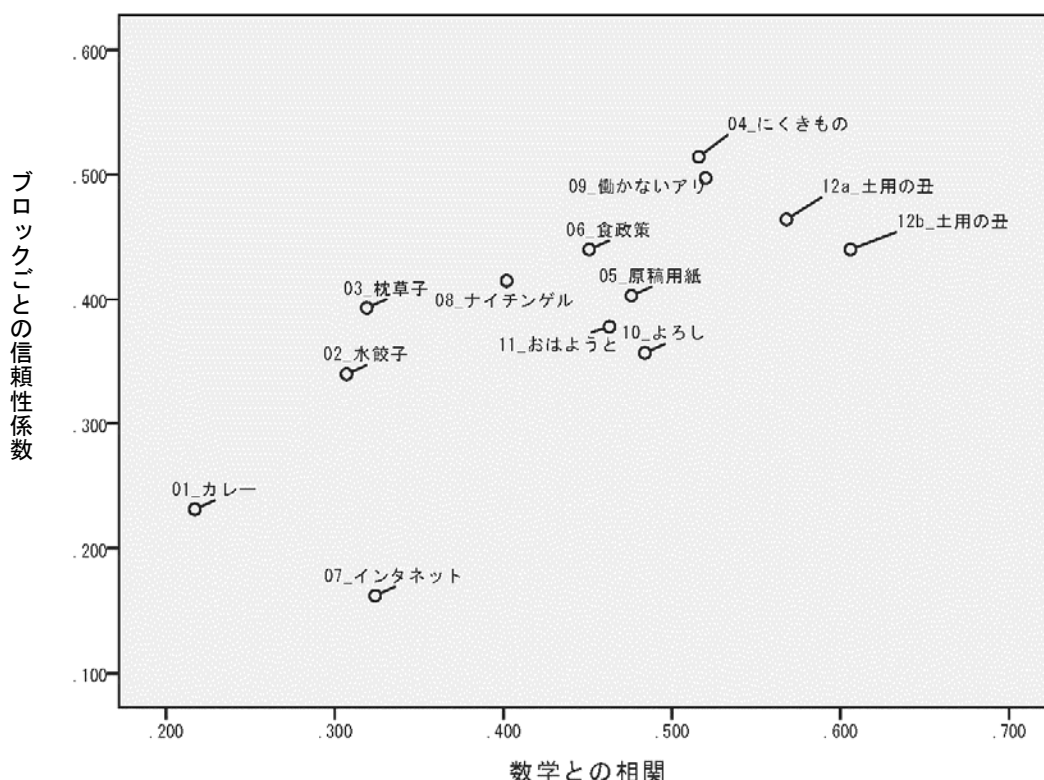


図 8.2.1 ブロックごとの信頼性係数と数学  $\theta$  との相関係数の散布図

### 8.3 文学的文章及び説明的文章の正答率と記述問題の無答率、誤答率との関係

第2章3節で述べたように、PISAの2003年の調査より明らかになった我が国の生徒の特徴の1つとして、無答率の高さが挙げられている。生徒の設問に対する反応が誤答である場合、その解答には、生徒の知識の定着状況や、生徒がどのような点で誤った理解をしているのかを知るためのヒントが隠されていることが多い。誤答分析を行うことで、生徒がつまづいている箇所を明らかにできれば、テスト後にその箇所の理解を促すためのフォローアップを行うことができる。このように生徒の誤答には生徒と教師の両者にとって有益な情報が含まれている。しかしながら、無答の場合には、「なぜ間違えたのか」「なぜ正答を得られなかったのか」という教育的に重要な問題に関する情報を得ることが難しい。そのため、無答率の高さは問題視されてきた。この無答率の高さに関する情報を得ることが本節の目的である。

第8章1節より、文学的文章と説明的文章では文章の質が異なっている可能性が示唆された。そこで、両者の選択問題の正答率と、記述問題の無答率との関係を調べることにする。具体的には、文学的文章（選択問題）の正答率の高低と、説明的文章（選択問題）の正答率の高低の2つの要因を群にとり、記述問題の無答率、誤答率を比較する。

文学的文章の選択問題の正答率の平均 0.69 (SD=0.31) を基準に、文学的文章（選択問題）の高低群分けを行った。その結果、低群 1158 名、高群 796 名となった。同様に、説明的文章の選択問題の正答率の平均 0.64 (SD=0.25) を基準に、説明的文章（選択問題）の高低群分けを行った。その結果、低群は 751 名、高群は 1203 名であった。

文学的文章の選択問題の成績及び、説明的文章の選択問題の成績が、記述問題の無答率及び誤答率とどのような関係にあるかを調べるため、文学的文章（選択問題）の高低及び説明的文章（選択問題）の高低の二要因を群にとり、記述問題の無答率及び誤答率について、分散分析を行った（表 8.2.6）。

その結果、文学的文章（選択問題。以下、選択）高低の主効果は無答率について有意であったが ( $F(1, 1950)=31.443, p<.01$ )、誤答率については有意でなかった ( $F(1, 1950)=0.534, n.s.$ )。説明的文章（選択）高低の主効果は無答率 ( $F(1, 1950)=94.890, p<.01$ )、誤答率 ( $F(1, 1950)=8.747, p<.01$ ) の両方で有意であった。文学的文章(選択)と説明的文章(選択)の交互作用も同様に、無答率 ( $F(1, 1950)=15.102, p<.01$ )、誤答率 ( $F(1, 1950)=8.747, p<.01$ ) の両方で有意であった。

そこで、多重比較 (Bonferroni 法) を行ったところ、無答率については、説明的文章（選択）の低群で文学的文章（選択）高低の単純主効果が有意であった ( $F(1, 1950)=33.348, p<.01$ )。誤答率については、説明的文章低群では文学的文章（選択）高低の単純主効果は有意傾向 ( $F(1, 1950)=3.695, p=.06$ )、説明的文章高群では有意であった ( $F(1, 1950)=16.467, p<.01$ )。さらに、文学的文章（選択）高低群別に、説明的文章（選択）高低の単純主効果をみたところ、無答率については、文学的文章（選択）の高低群ともに、説明的文章（選択）高低の単純主効果が有意であり（低群： $F(1, 1950)=137.569, p<.01$ ；高群： $F(1, 1950)=12.936, p<.01$ ）、誤答率については、文学的文章（選択）高群でのみ、有意であった ( $F(1, 1950)=17.698, p<.01$ )。

自由度が大きい中での分散分析の結果であるため、その解釈は慎重にする必要があるが、そのうえで、以上をまとめると、1) 文学的文章の多肢選択問題の正答率には関係なく、説明的文章の多肢選択問題の正答率が高いと無答率が低く、2) 説明的文章の多肢選択問題の正答率が低かった群では、文学的文章の正答率が高いと無答率が低いことが指摘できる。つまり、全体的に説明的文章の多肢選択問題の正答率が高ければ、記述問題での無答が少ないものの、正答率が低い場合に限れば、文学的文章の多肢選択問題の出来と無答率との関係が示唆されたといえるであろう。この差が統計的有意差だけではなくて、実質的な差であると仮に解釈するとすれば、従来からの文学的文章を教材とする国語教育の、目立たないが重要な教育上の成果を示すものといえるであろう。

表 8.2.6 文学的文章（選択）の高低×説明的文章（選択）の高低ごとの記述問題の無答率及び誤答率の平均と SD

文学的文章（選択問題） 説明的文章（選択問題）	低群 (n=1158)				高群 (n=796)			
	低群 (n=556)		高群 (n=602)		低群 (n=195)		高群 (n=601)	
	平均	SD	平均	SD	平均	SD	平均	SD
記述問題の無答率	0.26	0.37	0.08	0.22	0.13	0.26	0.05	0.19
記述問題の誤答率	0.50	0.33	0.51	0.28	0.55	0.27	0.44	0.27

分散分析の結果より示された説明的文章の正答率と記述問題の無答率との関係を、より鋭敏に捉えることを目的とし、信頼性の高かった分冊を取り上げ、文学的文章の記述問題の無答の有無と、説明的文章の成績との関係を検討する。第5章1節では、本調査で使用した分冊のうち、特に分冊4と分冊8の信頼性が高いことが示された。両者に共通して含まれていたブロックがB04である。また、資料2.7.1

より、ブロック4の記述問題B4\_4は、他の記述式問題に比べ、識別力母数が0.805と高く、難易度も-0.071と0に近く、IRT分析の結果から、識別力及び難易度が適当な記述問題であったと解釈できる。

そこで、記述問題と選択問題の関係をより鋭敏に捉えるために、分冊8に焦点を当て、B4\_4と他の問題との関係を検討する。分冊8に含まれる3ブロックにはB04, B08, B11の3つのブロックが含まれるが、そのうち、適切な記述問題と判断されるB4\_4について、無答者に0、正答の有無に関係なく回答者に1を与え、説明的文章を題材とした2つのブロック（B08とB11）の合計得点との相関係数を算出した。その結果、相関係数は0.331であり、B4\_4と説明的文章の得点と弱い正の相関関係にあることがわかった。この結果は、文学的文章であるB4の記述問題の無答者は回答者に比べて、説明的文章の得点が低いことを示したものである。これは、先に述べた説明的文章の正答率の高い者ほど、無答を避けるという結果と一致するものである。

第2章3節でも述べたように、我が国の国語教育では、無答率の高さが問題視されてきた。本節より、この記述問題の無答率の高さに、説明的文章の多肢選択問題の正答率が関係していること、さらにこの説明的文章の多肢選択問題の正答率の低い者にとっては、文学的文章の多肢選択式問題の正答率が関係していることが示唆された。この結果は、以下の3つの点において、意義があるといえよう。

第一に、説明的文章に関する能力の重要性を示した点がある。従来我が国の国語教育では、連続型テキストが中心的に扱われており、非連続型テキストはあまり扱われてこなかったという背景がある。本調査では、説明的文章として、非連続型テキストを積極的に導入している。この非連続型テキストを含んだ説明的文章を題材とした問題の正答率が、無答率の低さに影響を与えていることを示せたことは、近年の国語教育における非連続型テキストを含めた多様なテキストを導入しようという流れの有効性を示す結果であるといえる。

第二に、説明的文章を題材とした問題の正答率の重要性を指摘した反面、これまで行われてきた文学的文章を題材とした国語能力の重要性も指摘したことがある。本調査結果を概観すると、説明的文章の読解力が、文学的文章の読解力よりも無答率に影響を及ぼしているようにみとれる。しかしながら、詳細に検討してみると、説明的文章の読解力が低い場合に、文学的文章の読解力がある方が、無答を避け、誤答であっても回答をするという傾向がみられた。教育的意義から考えると、たとえ正答でなくても誤答を恐れずに回答することは非常に重要であるとされる。このような姿勢に文学的文章の正答率が影響していることを示したことは、従来国語教育で実践されてきた文学的文章を使った国語教育の重要性を指摘するよい材料となるだろう。

第三に、本調査で作成した記述問題は、全て本調査の定義のうち「情報の編集・統合」「判断」、すなわち PISA 調査におけるリーディング・リテラシーの3側面うち「熟考・評価」に含まれる項目である。これは、我が国ではあまり重点的に教育されてこなかった側面であり、この項目への不慣れが無答率の高さに影響していると考えられてきたことは、第2章3節で述べた通りである。このような側面に関連する項目の無答率に、影響を与える要因について示唆を与えられたということは、「熟考・評価」の側面を強化しようという近年の国語教育の流れに合致するものといえよう。

補注：外的情報の埋込（super-impose）の原理とRプログラム例

多次元尺度法によってもとめられた刺激が布置されている  $p$ -空間にある方向をもつベクトル  $\beta$  を埋め込むことを考える。いま、刺激  $i$  の座標ベクトルを  $x_i = (x_{i1}, x_{i2})$  とする（簡単のため空間の次元数は  $p=2$  とした）。このベクトルからベクトル  $\beta_i = (\beta_{i1}, \beta_{i2})$  に落とした正射影の長さを  $\hat{y}_i$  とおくと、そのベクトルは

$$\hat{y}_i \beta = \frac{\langle x_i, \beta \rangle}{\langle \beta, \beta \rangle} \beta$$

と表現できる。ここで、 $\langle x_i, \beta \rangle$  はベクトル  $x_i$  とベクトル  $\beta$  の内積を表す。いま、ベクトル  $\beta$  の長さを 1 とすると  $\langle \beta, \beta \rangle = 1$  であるから、 $\hat{y}_i$  は

$$\hat{y}_i = \langle x_i, \beta \rangle$$

と表すことができる。次に刺激  $i$  に与えられた外的情報の値を  $y_i$  とおき、この  $\hat{y}_i$  との誤差の二乗和が最小となるような  $\hat{\beta}$  を求める問題：

$$\text{minimize } \sum_{i=1}^m (y_i - \hat{y}_i)^2 = \sum_{i=1}^m \{y_i - \langle x_i, \beta \rangle\}^2$$

を考えると、これは刺激の座標行列  $X$  を独立変数、外的情報のベクトル  $y$  を従属変数とする重回帰分析と同値の問題となる。したがって、外的情報ごとに  $\hat{\beta}$  を計算し、それを刺激の布置図に埋め込めば求める方向が表現できる。なお、具体的な計算手続きについては次ページの R によるプログラム例を参照のこと。

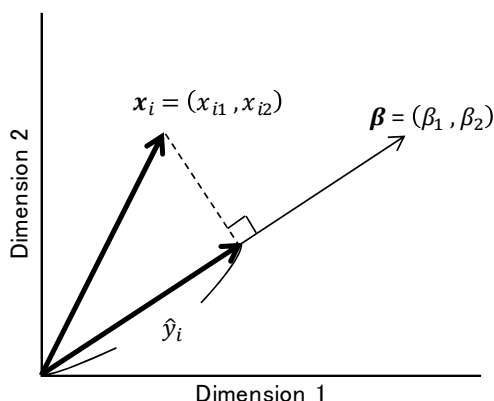


図 外的情報の埋込

## Rによるプログラム例

```

1 #
2 # MDSの結果に外的属性の方向をベクトルで埋め込む方法
3 # 2 0 1 2. 0 3. 2 2
4 # 文部科学省調査研究
5 #
6 #           programmed by SHIBAYAMA, T
7 #           with MIYATA, K. & NAKANO, Y
8 #           at TOHOKU UNIVERSITY
9 #
10
11
12 #大問ラベル
13 kokugo.label <- c("B01", "B02", "B03", "B04", "B05", "B06", "B07", "B08", "B09", "B10", "B11", "B12")
14
15 #座標の読み込み
16 #座標自体はSPSS/ALSCALからの結果を使った
17 x1 <- c(2.0921, 2.0921, 2.0875, 1.1911, -1.0326, -0.9712, -1.024, -0.9591, -1.0325, -0.8001, -0.8001, -0.8433)
18 x2 <- c(-0.4885, -0.4885, -0.4877, -0.2797, -0.2695, 0.1762, -0.1603, 0.2307, -0.2669, 0.7135, 0.7136, 0.6073)
19
20 #座標の方向に意味がないので第2軸を逆転
21 # 因子分析は変数間の角度が問題なのに対して、MDSは刺激間の距離が問題
22 x2 <- -x2
23
24 #データフレームによる座標行列の作成
25 x <- data.frame(TAIMON=kokugo.label, DIM1=x1, DIM2=x2)
26 x
27
28 #ラベル付きプロット
29 #参考URL http://www1.doshisha.ac.jp/~mjn/R/06.html
30 win.graph()
31 plot(x[, 2], x[, 3], type="n", xlim=c(-2, 2), ylim=c(-2, 2), xlab="dimension 1", ylab="dimension 2")
32 text(x[, 2], x[, 3], x[, 1])
33
34 #ラベルなしプロット
35 #win.graph()
36 #plot(x[, 2], x[, 3], xlim=c(-2, 2), ylim=c(-2, 2), xlab="dimension 1", ylab="dimension 2")
37
38
39 #外的属性のベクトルを埋め込む
40
41 # 文学的文章
42
43 y <- c(1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0)
44 y <- y - mean(y) #中心化
45
46 # 外的属性の示す方向を求める
47 # 回帰分析に帰着できる
48 # 中心化しているため切片は0を式中に明示
49 # 回帰係数のみ取り出している
50
51 beta <- coefficients(lm(y ~ 0 + x$DIM1 + x$DIM2))
52
53 #ベクトルbetaの長さを1に標準化する
54
55 beta <- beta /sqrt(sum(beta*beta))
56 arrows(0, 0, beta[1], beta[2])
57 text(beta[1], beta[2], "文学的文章")
58
59 # 以下、外的属性を変えながら同じ作業を繰り返す
60
61 # 説明的文章
62 y <- c(0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1)
63 y <- y - mean(y)
64 beta <- coefficients(lm(y ~ 0 + x$DIM1 + x$DIM2))
65 beta <- beta /sqrt(sum(beta*beta))
66 arrows(0, 0, beta[1], beta[2])
67 text(beta[1], beta[2], "説明的文章")
68
69 # 図
70 y <- c(0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1)
71 y <- y - mean(y)
72 beta <- coefficients(lm(y ~ 0 + x$DIM1 + x$DIM2))
73 beta <- beta /sqrt(sum(beta*beta))
74 arrows(0, 0, beta[1], beta[2])
75 text(beta[1], beta[2], "図")
76
77 #横書き
78 y <- c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0)
79 y <- y - mean(y)
80 beta <- coefficients(lm(y ~ 0 + x$DIM1 + x$DIM2))
81 beta <- beta /sqrt(sum(beta*beta))
82 arrows(0, 0, beta[1], beta[2])
83 text(beta[1], beta[2], "横書き")
84
85

```

## 9 教師質問紙の分析

実施校の担当教員におこなったアンケート結果である。選択式の質問と自由記述の部分から構成されている。

### 9.1 選択式質問

回答比率が50%を超えたものには網掛けをした。回答数は、65件。教員の担当教科の内訳は、下記の通りである。

担当科目	人数
国語	18
数学	10
社会	19
理科	5
英語	6
音楽	1
美術	4
技術・家庭	1
保健体育	2

**問1** 学習指導要領全体をカバーしたテストが仮にあったとして、それをすべての生徒に一斉に解いてもらうと何時間もかかり、生徒には大きな負担がかかります。そこで、この調査では、各教科数種類の問題冊子を生徒一人ひとりに割り当てる方式を取ることで、生徒の負担を軽くしながら、全体で学習指導要領の広範囲の問題を出題できるように設計されています。こういった調査方式について、どのように思われますか？ あてはまるものすべてに○をつけてください。

	回答比率
(1) 幅広い範囲についての学力・学習状況がつかめるので有益な方法と思う。	21%
(2) 事前に新しい調査方式に関する十分な説明がなされれば実施可能と思う。	45%
(3) 生徒がそれぞれ異なる問題を解いているので、結果を返すときの指導に工夫が必要と思う。	67%
(4) 調査用紙の配布方法など従前の方法に比べ多少煩雑さを感じる。	20%



**問2** 学力調査の結果を平素の学級の指導に役立てるために、必要と思われる情報は次のうちどれですか？ あてはまるものすべてに○をつけてください。

	回答比率
(1) 平均点（学校全体及び学級別）	41%
(2) 大問別，観点別，内容別の正答率（学校全体及び学級別）	77%
(3) 生徒一人ひとりの正誤状況	73%
(4) 学校全体としての経年変化	32%
(5) 生徒個人ごとの経年変化	42%

**問3** 学力調査の方法として、国際調査（例：PISA，TIMSS）のように全体傾向を測るためにあえて個人には成績結果をフィードバックしない方法もあります。このような調査方法について、どのようにお考えですか？ あてはまるものに○をつけてください。

	回答比率
(1) 目的によってはあってもよいと考える	38%
(2) 生徒指導の観点からは，個人への結果のフィードバックは必ず必要である。	23%
(3) あってもよいと考えるが，参考としては個人の成績表は必要である。	41%
(4) 生徒一人ひとりが異なる問題を解いているのであれば，個人成績表は必要ない。	11%

**問4** 上記のような国際調査の報告では以下のような統計用語が使われています。このうち目にされたことのある用語がございましたら○をつけてください。

	回答比率
(1) 平均	95%
(2) 分散	70%
(3) 標準偏差	98%
(4) 中央値	61%
(5) 相関係数	61%
(6) 識別力	14%
(7) 通過率	41%
(8) 信頼性係数	17%
(9) 妥当性	48%
(10) 換算点	24%
(11) パーセンタイル順位	21%
(12) 信頼区間	17%

(13)	測定の標準誤差	24%
(14)	回帰式	15%
(15)	等化	12%

## 9.2 自由回答

(問 1)

- ・ 何を目的にした調査なのかによって答がちがうと思います。

(問 2)

- ・ 必要な情報は他にも多くある。

(問 3)

- ・ 目的しだいです。学力調査の目的が何か、お知らせ下さい。学力向上であれば、直接的に学力向上に取り組むことが近道です。

(その他)

- ・ 8種のテストの一番難しい問題が学力的に低い生徒に当たった場合、その生徒にとって実施している45分間は苦しみだけの時間となる。また、その結果をフィードバックすれば、もっと自信をなくす。実施している時間は、それぞれ、国と数1時間ずつ教科で時数カウントしたが、生徒によっては、大人の都合でつくられたムダな時間になってしまう。

## 参考文献

- 足立幸子. (2005). マルチリテラシーズ. 月刊国語教育研究 395, 46-51.
- Andersen, E. B. (1972). The numerical solution of a set of conditional estimation equations. *The Journal of the Royal Statistical Society, series B*, 34(1), 42–54.
- 荒井克弘・倉元直樹 (編) (2008). 全国学力調査—日米比較研究—. 金子書房
- Baker, F. B., & Kim, S. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Marcel Dekker.
- Birnbaum, A. (1968). Some latent trait models and their uses in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 395-479). Reading, MA: Addison-Wesley.
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46(4), 443–459.
- Bock, R. D., & Lieberman, M. (1970). Fitting a response model for  $n$  dichotomously scored items. *Psychometrika*, 35(2), 179–197.
- Cai, L., Thissen, D., & du Toit, S. (2011). IRTPRO 2.1 for Windows [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.
- Cope, B. & Kalantzis, M. (2000). *Multiliteracies: Literacy learning and the design of social futures*. London: Routledge.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39(1), 1-38.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. London: Lawrence Erlbaum.
- 伏見正則(1989). 乱数. 東京大学出版会.
- Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, 22, 144-149.
- Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, 9(2), 139-164.
- 肥田野直 (編) (1972) . 心理学研究法第 7 巻テスト I. 東京大学出版会.
- Holland, P. W., & Dorans, N. J. (2006) Linking and Equating. In R. L. Brennan (Ed.), *Educational Measurement*. 4th ed. Westport, CT: American Council on Education and Praeger Publishers. pp. 187-220.
- 池田 央 (1994) . 現代テスト理論. 朝倉書店.
- 経済協力開発機構(OECD), 国立教育政策研究所訳. (2004). PISA2003 年調査 評価の枠組み—OECD 生徒の学習到達度調査—. ぎょうせい.
- 経済協力開発機構(OECD), 国立教育政策研究所訳. (2007). PISA2006 年調査 評価の枠組み—OECD 生徒の学習到達度調査—. ぎょうせい.
- 経済協力開発機構(OECD), 国立教育政策研究所訳. (2010). PISA2009 年調査 評価の枠組み—OECD 生徒の学習到達度調査—. 明石書店.
- Kendall, M. G., & Stuart, A. (1979). *The advanced theory of statistics* (4th ed., Vol. 2). New York: Oxford University Press.

- 国立教育政策研究所. (2002). 生きるための知識と技能—OECD 生徒の学習到達度調査(PISA)2000 年調査国際結果報告書—. ぎょうせい.
- 国立教育政策研究所. (2004). 生きるための知識と技能 2—OECD 生徒の学習到達度調査(PISA)2003 年調査国際結果報告書—. ぎょうせい.
- 国立教育政策研究所. (2007). 生きるための知識と技能 3—OECD 生徒の学習到達度調査(PISA)2006 年調査国際結果報告書—. ぎょうせい.
- 国立教育政策研究所. (2010). 生きるための知識と技能 4—OECD 生徒の学習到達度調査(PISA)2009 年調査国際結果報告書—. ぎょうせい.
- 熊谷龍一 (2009). 初学者向けの項目反応理論分析プログラム EasyEstimation シリーズの開発. 日本テスト学会誌, 5(1), 107-118.
- 熊谷龍一・野口裕之 (印刷中). 推定母集団分布を利用した共通受検者法による等化係数の推定. 日本テスト学会誌, 8, 9-18.
- 熊谷龍一・山口大輔・小林万里子・別府正彦・脇田貴文・野口裕之 (2007). 大規模英語学力テストにおける年度間・年度内比較—大学受験生の英語学力の推移—. 日本テスト学会誌, 3, 83-90.
- Lord, F. M. (1952). A theory of test scores. *Psychometric Monograph*, No. 7.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, 14, 139-160.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149-174.
- 文部科学省. (2006). 読解力向上に関する指導資料—PISA 調査(読解力)の結果分析と改善の方向—. 東洋館出版社
- 文部科学省. (2008). 中学校学習指導要領解説 国語編. 東洋館出版社
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16(2), 159-176.
- 村木英治 (2006). 全米学力調査 (NAEP) 概説—テストデザインと統計手法について. 東京大学大学院教育学研究科教育測定・カリキュラム開発講座 2005 年度研究活動報告書, 51-66.
- 村木英治 (2011). 項目反応理論. 朝倉書店.
- Muraki, E., & Bock, R. D. (2003). PARSCALE 4 for Windows: IRT based test scoring and item analysis for graded items and rating scales [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.
- New London Group (1996). A pedagogy of Multiliteracies: Designing social futures. *Harvard Educational Review* 66(1), 60-92.
- Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, 16(1), 1-32.
- 日本テスト学会 (編) (2007). テスト・スタンダード—日本のテストの将来に向けて—. 金子書房.
- 日本テスト学会 (編) (2010). 見直そう, テストを支える基本の技術と教育. 金子書房
- 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の等化方法. 教育心理学研究, 31, 233-238.
- 野口裕之 (1986). 共通被験者の反応パターンを利用した潜在特性尺度等化法. 教育心理学研究, 34,

315-323.

- 野口裕之・熊谷龍一 (2011). 共通受検者デザインにおける Mean & Sigma 法による等化係数推定値の補正. 日本テスト学会誌, **7**, 15-22.
- 野口裕之・熊谷龍一・大隅敦子 (2007). 日本語能力試験における級間共通尺度構成の試み. 日本語教育, **135**, 70-79.
- Olson, U (1979) Maximum likelihood estimation of the polychoric correlation coefficient, *Psychometrika* 44, 4
- 大友賢二 (1996). 項目応答理論入門. 大修館書店.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.
- Reckase, M., D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, 4(3), 207-320.
- Rychen, D. S., & Salganik, L. H. (eds.), (2003). *Key competencies for a successful life and a well-functioning society*. Hogrefe & Huber Publishers. ドミニク・S・ライチェン, ローラ・H・サルガニク編著. 立田慶裕 (監訳) ・今西幸蔵ほか訳. キー・コンピテンシー—国際標準の学力をめざして—. 明石書店
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 34(4), 100-114.
- 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・佐藤喜一・熊谷龍一・佐藤誠子 (2011). 平成 22 年度文部科学省委託研究「学力調査を活用した専門的な課題分析に関する調査研究」研究成果報告書.
- Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, **7**, 201-210.
- 豊田秀樹 (2002). 項目反応理論[入門編]—テストと測定の科学—. 朝倉書店.
- 豊田秀樹 (編著) (2002). 項目反応理論[事例編]—新しい心理テストの構成法—. 朝倉書店.
- 豊田秀樹 (編著) (2005). 項目反応理論[理論編]—テストの数理—. 朝倉書店.
- van der Linden, W. J., & Hambleton, R. K. (Eds.) (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.
- 全国的な学力調査の在り方等の検討に関する専門家会議(2011). 平成 23 年度以降の全国的な学力調査の在り方に関する検討のまとめ. 文部科学省.
- Zimowski, M.F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). BILOG-MG. IL: Scientific Software International.
- Zimowski, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG: Multiple-group BILOG*. Chicago, IL: Scientific Software International, Inc.