全国規模の学力調査における
マトリックス・サンプリングにもとづく
集団統計量の推定について

平成２４年度文部科学省委託研究
「学力調査を活用した専門的課題分析に関する調査研究」
研究成果報告書

平成２５年３月３０日

国立大学法人東北大学

はしがき

この報告は，平成 24 年度文部科学省企画公募研究「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」の『全国的な学力調査の調査手法における技術的課題に関する調査研究』に応募し，技術審査会等を経て採用された調査研究，「全国規模の学力調査におけるマトリックス・サンプリングにもとづく集団統計量の推定について」の成果をまとめたものである．

全国規模の学力調査のような大規模アセスメント（large-scale assessment）では，偏りのない結果を得るために，その出題内容はいうまでもなく，それを支える体系的な調査技術が必須のものである．たとえば，抽出調査ならば児童・生徒・学校等の母集団からのサンプリングをどのように実施するかの方法論，すなわち，標本調査法が必要というのはすぐに理解できる．しかし，その一方で，大規模アセスメントにおいては，調査を必要とする学習領域や学習状況なども，実は幅広く，児童・生徒・学校等のサンプリングと同様に，調査項目のサンプリングとそれをテスト冊子に組み上げていく方法論が必要であることにも目を向けなければならない．

この調査研究は，これまでの我が国の大規模アセスメントではあまり関心の払われてこなかった上記の方法論に焦点を当て，平成 22 年度，平成 23 年度の 2 年間にわたる調査研究の成果の上に立って，以下の３つの新しい測定技術の導入とそのノウハウの獲得を目的とした．すなわち，

１）テスト仕様の作成の際のマトリックス・サンプリングの適用，

２）分冊の組み上げデザインとしての釣合い型不完備ブロックデザインの採用，

３）事後分析のために必要最小限の調査項目から偏りのない推定値を得る推算値の導入，

である．また，あわせて

４）従来の国語科の問題との違いを鮮明にすることで，リーディング・リテラシーの概念を整理し，

５）リーディング・リテラシーから見た数学の問題の特徴分析も試みたこと

も本調査研究で得られた成果の一つである．

平成 25 年度全国学力・学習状況調査はいわゆる「きめ細かい調査」として，全数調査としての本体調査の他に，オプションとして経年変化分析，保護者に対する調査，教育委員会に対する調査が予定されている．その先には，経年変化分析と本体調査のリンキングを実現することによって，たとえば学校単位での経年変化がトレースできるようにするなどの解決すべき技術的課題が存在しているのである．このように，「指導」も「調査」も大切にしようとする日本型の大規模学力調査の今後のあり方を支える技術論的基盤のひとつとして，本調査研究を位置づけることができるであろう．

研究代表者　柴山　直

謝　辞

ご多忙の中，この調査研究にご協力いただきました宮城県塩竈市・白石市・多賀城市・利府町・大和町・富谷町の各中学校，生徒の皆様，先生方，ならびに各自治体教育委員会の諸氏，また本調査研究にご理解をいただき，厚いご支援をたまわりました宮城県教育庁の皆様に，ここに記して深く感謝申しあげます．皆様のご協力なくしてはこの調査研究を遂行することはできませんでした．

事業概要

事 業 名	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
事業内容	全国的学力調査の調査手法における技術的課題に関する調査研究
委託期間	平成 23 年 8 月 1 日から平成 24 年 3 月 31 日
事業者名	国立大学法人東北大学・大学院教育学研究科長・本郷　一夫
事 業 費	６，７４０千円

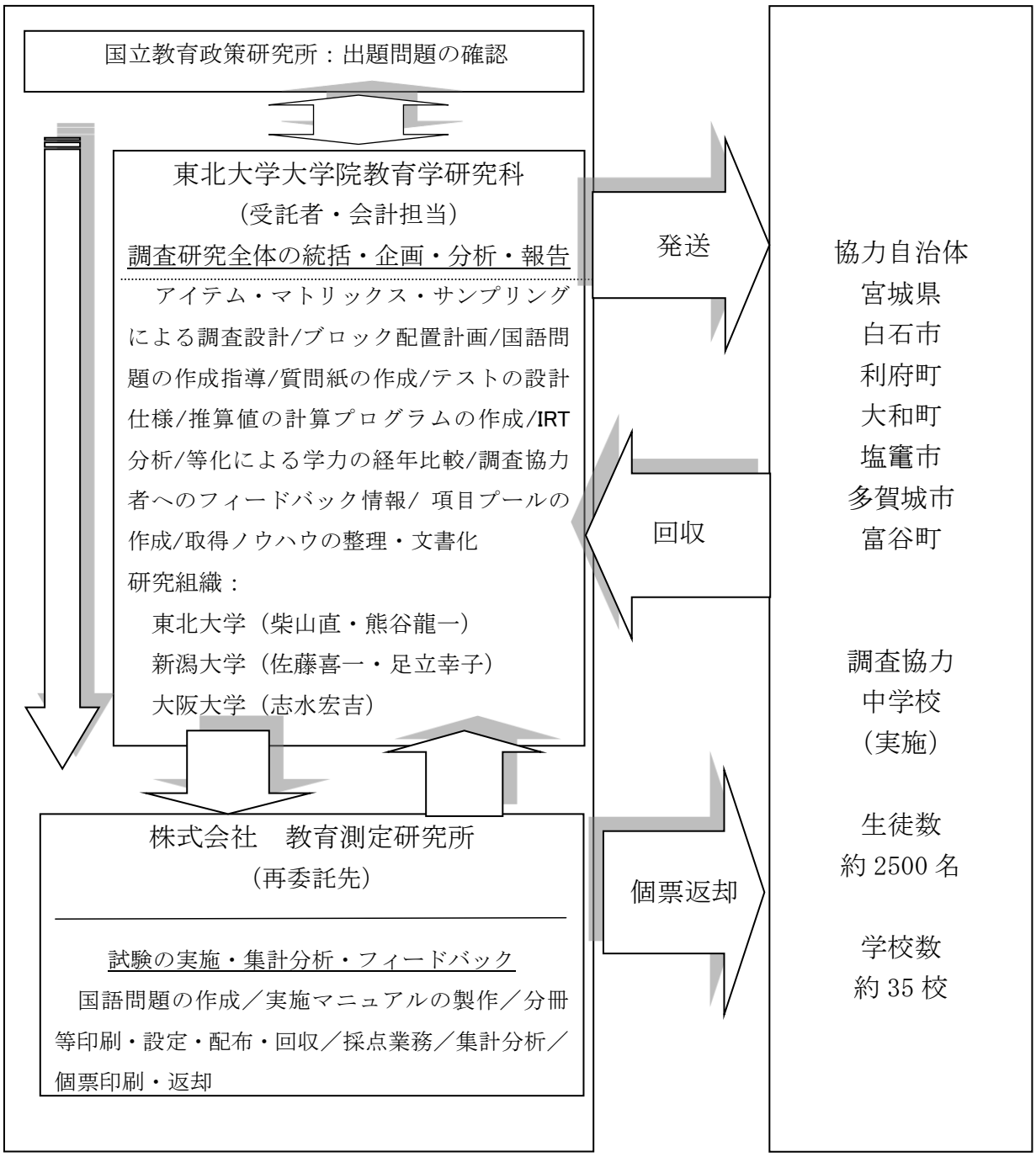
研究組織

研究代表	柴山　直	東北大学大学院教育学研究科
研究協力	熊谷　龍一	東北大学大学院教育学研究科
	佐藤　喜一	新潟大学入学センター
	足立　幸子	新潟大学教育学部
	志水　宏吉	大阪大学人間科学研究科
研究助手	中野　友香子	東北大学大学院教育学研究科
研究補助	佐藤　誠子	東北大学大学院教育学研究科
	宮田　佳緒里	東北大学大学院教育学研究科
	千葉　陽子	東北大学教育学部
	坂本　佑太朗	東北大学教育学部
事務担当	紙屋　雅子	

実施集計　株式会社教育測定研究所

作題助言　国立教育政策研究所

事業の実施体制図


国立教育政策研究所：出題問題の確認
東北大学大学院教育学研究科
（受託者・会計担当）
調査研究全体の統括・企画・分析・報告
アイテム・マトリックス・サンプリングによる調査設計/ブロック配置計画/国語問題の作成指導/質問紙の作成/テストの設計仕様/推算値の計算プログラムの作成/IRT分析/等化による学力の経年比較/調査協力者へのフィードバック情報/ 項目プールの作成/取得ノウハウの整理・文書化
研究組織：
東北大学（柴山直・熊谷龍一）
新潟大学（佐藤喜一・足立幸子）
大阪大学（志水宏吉）
株式会社　教育測定研究所
（再委託先）
試験の実施・集計分析・フィードバック
国語問題の作成／実施マニュアルの製作／分冊等印刷・設定・配布・回収／採点業務／集計分析／個票印刷・返却
発送
回収
個票返却
協力自治体
宮城県
白石市
利府町
大和町
塩竈市
多賀城市
富谷町
調査協力
中学校
（実施）
生徒数
約 2500 名
学校数
約 35 校

調査研究計画

【準備】

1. 出題問題の選定作業開始（数学・国語）
2. 質問紙調査項目の開発開始
3. 協力自治体への依頼および実施にあたっての諸事項の調整
4. 出題問題の確定（数学・国語）
5. 調査デザインのための項目配置計画の策定
6. 質問紙調査項目の確定
7. 協力校の確定
8. 分冊・解答用紙等のレイアウト確定
9. 試験実施マニュアルの作成

【実施】

10. 協力校への依頼・実施内容の説明
11. 調査用紙等細部点検・確認
12. 試験実施マニュアル・問題冊子・解答用紙の印刷
13. 配布準備，搬送
14. 調査実施

【分析】

15. 推算値計算用ソフトのプログラミング
16. データ入力作業・基礎統計量の集計
17. 調査協力者への個別フィードバック
18. IRT 分析等データ解析作業
19. IRT 分析・推算値等データ解析作業
20. 推算値を利用した分析
21. データ解析結果の整理

【文書化】

22. 開発したノウハウについてのまとめ
23. 報告内容に関する最終検討会（於東北大学）
24. 最終報告書の作成

実施経過

2012.08.03　各自治体へ直接訪問し協力依頼終了
2012.08.04　数学釣合い型不完備ブロックデザイン（BIBD）打合せ　於　教育測定研究所
2012.08.05　数学の分冊デザインの組み直し
2012.08.13～14　各自治体への協力依頼状の作成
2012.08.15　各自治体への協力依頼状の発送
2012.08.17　国語問題（小説）3 問作成
2012.08.19　国研へ，国語問題（小説）3 問のチェック依頼
2012.09.04　学校質問紙，生徒質問紙素案作成
2012.09.12　国研より国語問題（小説）のチェック返却
2012.09.13　国研へ，国語問題（小説）2 問修正版のチェック依頼
2012.09.18　国語問題冊子組み　質問紙修正
2012.09.19　国語の項目セット決定
2012.09.21　質問紙決定
2012.09.27　国語・数学　分冊完成
2012.09.28　質問紙完成
2012.10　各協力中学校で調査実施
2012.11.19　打合せ（進捗状況の確認・個人票の見方解説文の検討・BIBD 効果について
報告・報告書構成執筆分担案について検討）於　教育測定研究所
2012.11.28　素データ入力終了
2012.12.09　報告書目次案作成
2012.12.11　国語 GP 表完成
2012.12.12　打合せ　於　東北大　（国語の結果について検討）
2012.12.13　問題対応表，数学改変箇所一覧表の作成
2012.12.14　項目管理表作成
2012.12.17　打合せ（調査事業打ち合わせ）　於　新潟大学
打合せ（基礎集計作業等の進捗状況確認等）於　教育測定研究所
2012.12.18　質問紙データの入力終了
2012.12.19　報告書用諸資料の整理
2012.12.24　項目管理表の再修正
2013.02.04　打合せ（分析結果矛盾の原因究明作業）於　教育測定研究所
2013.03.11　打合せ（推算値の計算アルゴリズムの検討）　於　新潟大学
打合せ（分析データ突き合わせ結果の検証作業）於　教育測定研究所
2013.03.26　分析結果および考察に関す最終報告会　於　東北大学

目次

全国規模の学力調査におけるマトリックス・サンプリングにもとづく集団統計量の推定について

第１部　本調査の概要

1. 本調査の概要

1.1 本調査研究の意義と目的

この調査研究では平成 22 年度ならびに平成 23 年度に実施された文部科学省委託研究「学力調査を活用した専門的課題分析に関する調査研究」のうちの２つの調査研究，「全国規模の学力調査における重複テスト分冊法適用の試み」，および，「全国規模の学力調査における重複テスト分冊法の展開可能性について」によって獲得された重複テスト分冊法に関する種々の知見の上に，学力の経年変化ならびに集団統計量を捉えるための技術的基盤を確立することを目的とした．そのために，あらたに，釣合い型不完備ブロックデザイン(balanced incomplete block designs；BIBD)にもとづくマトリックス・サンプリング(matrix-sampling)，ならびに推算値(plausible values；PV)の二つの手法を導入した．

平成 25 年度から実施される，いわゆる全国学力・学習状況調査（きめ細かい調査）では，全数調査方式による部分とサンプリング調査による部分とに分割され，後者のサンプリング調査では経年変化などが可能となるような枠組みが取り入れられる予定である．通常，サンプリング調査を行う際には対象となる児童生徒を全国の母集団から，たとえば，層化多段抽出法などによって，推定すべき集団統計量の値に偏りが生じないように選ぶなどの工夫がなされる．これに加えて，学力調査の場合，児童生徒の標本集団を偏りなく選ぶと同時に，調査すべき学習領域すべてにわたって偏りなく出題する工夫が必要となる．児童生徒の母集団と同様に出題する問題項目（以後，項目；item）のいわば母集団（項目プール；item pool）を考え，そこから偏りなく項目を選ぶことによって，児童生徒集団に関する統計的な推測と同時にたとえば教育行政の観点から必要な学習分野がすべて予定通り修得されたかどうかの検討が可能となるようにするのである．このように児童生徒サイドと項目サイドの両方から標本抽出を考えることをマトリックス・サンプリング（Pophan,W.J(1993)）とよぶ．

児童生徒を対象としたサンプリングについては既存の標本調査法の適用が重要であるが，一方，項目に関しては，項目プールからサンプリングした後の，むしろ具体的な問題冊子の組み上げの際に，農学や化学，品質管理の分野で発展してきた実験計画法のうち，特に釣合い型不完備ブロックデザイン（BIBD）の適用が重要となる．すなわち，重複テスト分冊法で使用される分冊（booklet）を準備された項目から組み上げていく際に，複数の分冊によって幅広い調査領域をカバーしながらも，得られたデータに含まれるバイアスをなるべく小さくすることを目的に BIBD は導入されるのである．さらに，推算値は，異なる問題から構成される分冊によって求められた学力特性値（より一般的には尺度値）の推定誤差を補正しながら，質問紙で得られた回答パターンと学力特性値を結合するために導入されるものである．この推算値を介して種々の判断に必要なさまざま集団統計量の最終的な推定値が求められることになる．

本調査研究は，学力の測定モデル（Psychometric model)としては項目反応理論（Item Response Theory；IRT）モデルに基礎をおき，調査モデル(Survey model)としてはマトリックス・サンプリングを採用し，統計的に偏りのない事後分析を行うために推算値を導入するという，いわば 3 層構造から

成り立っている．さらに，調査ではあっても指導の側面を重要視する我が国における教育の独自性に配慮し，方法論上からは以下の 5 つの内容からこの調査研究を構成した．

（1）数学と国語において，NAEP で開発され PISA 等でも用いられている調査手法であるマトリックス･サンプリングの全国的な学力調査への適用ノウハウを修得しその効果について検証する．

（2）数学と国語における学力の集団的な経年変化を等化法によって捉える．その際，記述問題を組み込んだ場合の項目反応モデルにもとづく等化法の有効性を検証する．

（3） マトリックス・サンプリングを用いた調査において，学力調査に加えて質問紙調査を実施し，推算値を利用して学力と背景情報との関連を分析するノウハウを獲得する．

（4） マトリックス・サンプリングを用いた調査における調査協力者への結果のフィードバック方法の開発を，昨年度の数学に引きつづき国語においても行う．

（5）項目反応モデルをベースにして設計されたテスト結果の分析やその次のテストを作成する際に必須となる専門ソフトとして，EasyEstimation（熊谷 2009,2012）を採用する．

上の（1）から（3）が PISA 等の国際的な学力調査でも採用されている新しい調査方式に関するものである。また，（4）は調査であっても指導を重視する我が国の教育風土にあわせて，昨年度の数学に引きつづき，リーディング・リテラシーの測定を中心とした本年度の国語おいて試みるものである．さらに，（5）の EasyEsimation（熊谷 2009,2012）に関しては，このソフトがユーザーインターフェイスを初め，第 11 章で述べるように，他のソフトと比較して優れた特徴と使い勝手の良さがあるフリーソフトであることにより全面的に IRT 分析専用のソフトウェアとして採用した．実際，我が国ではすでに多くの研究機関等がダウンロードし，新しいテスト開発などに利用している．今後，我が国で大規模学力調査が本格的に実施されるようになれば，標準的な専門ソフトとして，学校現場なども含めて，より広く使われることになるであろう．

1.2 手法概要

1）マトリックス・サンプリングの導入と BIB デザインの採用

限られた時間の中でテストを実施するとき，たとえば学習指導要領を例にとると，1 回のテストでその全ての範囲を出題することは実質不可能である．また，記述形式の項目によってより高次の学力を測定しようとする場合には一項目あたりに必要な解答時間が長くなり，項目の数自体を減らす必要があり，ますます，調査できる範囲は制限されることになる．そのために，すべての対象領域からの多数の出題問題を準備し（それらの項目全体を項目プールとよぶ），それをいくつかの項目セットにわけ，さらにその項目セットを組み合わせることで，何冊かの分冊を構成し，その分冊を偏りなく実施することが求められる．このとき，各分冊における項目の位置や出現回数，互いの組合せ回数などに偏りが生じると正しい結果が得られない．そのため，分冊を組み上げるときに，実験計画法の分野で発展してきた釣合い型不完備ブロックデザイン(BIB デザイン)を採用することでその偏りをなくことを試みた．

２）等化法による学力の集団的な経年変化のモニターリング

平成 23 年度調査研究では数学において学習指導要領の改訂による新しい項目の導入，および項目特性が必ずしも良くなかった項目の差し替えによっても，学力の経年変化を捉えることができることを実証的に確認した．数学においては昨年度実施した 64 項目のうち精度の良い項目 52 個を，上で述べた BIB(7,7,3)デザインにしたがって採用し，テスト全体の測定精度を低下させないようにしながら，学力の時系列変化をモニターリングした．具体的には自治体ごと，学校ごとに尺度値の集団統計量をもとめ，昨年度のものと比較可能なようにすることを試みた．ただし，この報告書には自治体および学校の個別情報保護のため掲載していない．

また，国語については，昨年度実施した 12 個の大問（厳密にはバリエーションを含むと 13 個）のうち，必ずしも見込み通りの機能を果たさなかった大問を入れ替え，7 大問（項目セット）で 7 分冊を構成した上で，昨年度データから項目特性が推定できている 4 大問を基準（アンカーまたは係留と呼ばれる）にして，記述問題を含んだ場合の等化（equating）をあらたに試みる．これは多段階反応モデルにもとづく等化の試みと言うことになる．

３）生徒質問紙の実施と学力との相関分析

PISA 等の国際学力調査では生徒の家庭環境や学習条件等を調査し，学習到達度との関連性を分析するために実施されているものであるが，その統計分析に入る前段階で特殊な処理を行い，児童の尺度値をそのまま使うのではなく，推算値と呼ばれる統計量を求め，それと質問紙で得られたデータとを結びつけることでさまざまな分析を行っている．

母集団統計量を求める際には，分冊毎に項目構成が異なっていること，分冊に含まれる項目数が限定されていることなどが原因となって，生徒の尺度値から直接的には正確な母集団統計量の推定値が得られないことが知られている．とくに後者に関しては従来型の一冊子一斉方式のテストでも実は同様の問題を抱えている．

推算値の求め方は以下のようになる．まず x を項目反応パターン，θ を尺度値，f（x｜θ）を θ が所与の場合に項目パターン x が得られる項目反応確率とする．さらに θ の事前分布として平均 μ，分散 σ２の正規分布を仮定すると，x が得られた場合の事後確率 h（θ｜x）は

$$h(\theta|x) = \frac{f(x|\theta)g(\theta)}{\int f(x|\theta)g(\theta)dx}$$

で定義できる．PV はこの分布に従う乱数として得られる値である．与えられた分布関数に従う乱数を得る方法としては von Neumann の棄却法として知られる方法を採用した．また，そのための計算プログラムについても検証に資するためプログラムリストを本報告書に記載した．

具体的には，上の方法で，生徒一人につき M 個（通常は M=5）の PV を求め，そのひとつひとつにその生徒から得られた質問紙のデータを結びつける．したがって，最終的な統計分析の対象となるデータは形式的には，生徒数を n，質問項目数を m とすると n×(m+1)×M の，いわゆる 3 相 3 元データとなる．ここで，(m+1)は質問項目数に推算値を加えた数をあらわしている．このデータに基づく分

析例なども報告書に例示する．

４）個別フィードバック方法の開発

学力の時系列変化を捉えるためには問題は基本的に非公開にする必要がある．また，マトリックス・サンプリングなどの導入によって幅広い領域の調査を行うためには，複数の分冊に配置するためにきわめて多数の項目を準備しなければならない．それらの項目を，たとえば毎年作成し毎年公開していては，コスト・パフォーマンスが悪すぎる．また，非公開にすることで統計的に質保証された項目プールを構築することができれば，それらの項目を軸に，たとえば公開型の現行方式の学力調査と連動させながら，共通受検者デザインにもとづくテストリンキングの手法を使って，公開型の調査結果と非公開型の調査結果との対応づけによる，公開型のデータによる経年変動のモニターリングも技術的には可能になるなどのさまざまなメリットを得ることができるようになる．

しかし，我が国においては，たとえば本来的な意味での調査にのみ特化して合目的に実施し，その結果を目的通り教育施策に役立つレベルのみのフィードバックに限定して，調査に協力してくれた個々の児童・生徒になんら個別情報をフィードバックしないということは，学校サイドからすると，貴重な授業時間を割いて参加してくれた児童生徒の学力向上の指導に直接には役立たないという別の次元のデメリットが生じることを意味する．教育施策等の効果をみるために問題を非公開にすると，非公開にしたが故に出題された問題を使っての指導ができないと言うジレンマが生じるといってよい．

このジレンマを解消，ないし緩和するために，調査協力校からの強い要請に応えて過年度調査研究において，項目反応モデルの特質を活かした個別フィードバックを数学を中心に試みてきた．その結果，聞き取りではあるが，協力校からは一定の評価を得ることができている．

具体的には，児童・生徒に関して推定された尺度値 θ を使って，その児童・生徒が受けなかった項目についても，それらの項目特性が既知であれば，その項目を受けたと仮定した場合の期待正答確率が計算できるという項目反応モデルの特徴を利用する．このことによって，問題自体は非公開であるが，たとえば対象となる領域すべてにわたって，受けた項目については正誤情報を，受けなかった項目についてはその期待正答確率をフィードバックできることになり，十分とは言えないが指導に活かせる可能も担保できる．本年度はこの方法に関しても，とくにリーディング・リテラシーの調査が中心となる国語において学習指導要領との対応をとりながら，なるべく調査協力者である生徒の学力把握に役立つフィードバック方法となるよう工夫を試みた．

５）項目特性および学力特性値，推算値等の推定のための専用ソフトの採用

上記の４つの調査研究テーマを遂行していくための技術的基礎となっているのが項目反応モデル(IRT モデル)とよばれる一連の計量心理学的モデルである．通常の汎用型の統計ソフトでは必要な項目母数や尺度値などの推定ができない．そのため，たとえば NAEP や PISA ではそれぞれ PARSCALE や ConQuest と呼ばれる，それぞれの調査用にカスタマイズされた専用ソフトを準備している．

新しい調査方式を展開する場合，これらの商用版を利用することも選択肢の中に入るが，我が国においては，これらと同等の機能を備え，また優れたユーザーインターフェイスにより初心者にも使い

やすい EasyEstimation シリーズが一般向けに公開されている（熊谷, 2009）．過年度調査研究においてもこの専用ソフトを部分的に採用しながら, 商用ソフトである BILOG-MG やその後継のおなじく商用ソフトである IRTPRO で求められた種々の推定値との比較を行い，その結果を評価し，十分な機能をもつとの結論を得ている. また, それらの調査研究とは独立に当該分野の学会誌（熊谷, 2009, 2012））にも掲載されるなどその信頼性は第 3 者的にも担保されている.

さらに，問題非公開等の理由から，問題公開に変わる項目の品質保証の手段として，項目に関するさまざまな統計値を公開し，必ずしも当該分野に関する専門的訓練を受けているとは限らない第 3 者によって，それらのクロスチェック受けるなどの必要性を考えたときに，学術目的に限って自由に使える EasyEstimation シリーズの存在は大きい．このような理由から，実際に新しい調査方式を展開する状況を想定して，本調査研究では，専用ソフトとして全面的に EasyEstimation シリーズを採用した.

1.3 調査概要

調査対象は昨年度調査と同じく宮城県下の 6 つの自治体（うち一つは今年度あらたに協力が得られた自治体）の約 30 校 2500 名の中学 3 年生であった．実施は協力校の行事予定と重ならないようにすること，および時系列変化をとらえるためになるべく昨年度の実施時期に合わせる必要がある，などの理由から 10 月中旬であった.

実施教科は数学と国語であった．ただし国語に関してはリーディング・リテラシーに重点を置いた内容とした．出題する問題（以降，項目(item)と呼ぶ）については，過年度に出題し統計的な検証および内容的な妥当性の検討を経て品質が保証されたものを基本的には使用した．数学については過年度実施した 64 項目の中からの 56 項目，また国語については昨年度実施した 10 個のうち，非公開でかつ品質の高いものから順に残した 4 つの大問に加えて，3 つの大問を新しく作題し，あわせて 7 大問構成（各大問には 4 つの小問が含まれる）とした．なお，項目は項目プールを構成する必要上，本年度も引き続き非公開とするが，第 3 者による項目の質検証が可能なように，項目特性に関する統計値は可能な限り事後に報告する．ただし，国語に関してはリーディング・リテラシーに重点を置いた場合の問題内容の具体がわかるようにすべてを公開した.

1.4 実施手続き

本研究調査の実施内容は以下のとおりであった.

対象学年　中学３年生
対象人数　中学３年生：2541 名
実施校　　協力自治体の全中学校
実施時期　平成２４年１０月１日（月）～１０月１２日（金）
実施時間　上記の期間内の２授業時間＋質問紙調査のための１５分程度
（説明や配布，回収の時間 5 分を含めて 1 授業時間 45 分）

出題教科　数学・国語・質問紙
出題範囲　現行の学習指導要領の内容構成にあわせる
　　国語についてはＰＩＳＡ型のリーディング・リテラシー問題を含む
　　質問紙については学力の形成要因等
問題内容　国立教育政策研究所の助言を受ける
分冊数　　数学・国語とも 7 分冊／質問票 1 部
解答数　　中学校ともに一人当たり 30 問程度　問題冊子は回収
解答方式　記入式
個票返却　平成 23 年 12 月初旬　各学校へ各学校の結果とともに郵送で返却した

また，学校への負担をなるべく軽くするために，各学校では実施にかかる部分のみ担当とし，採点等はすべて研究組織側でおこなった．なお，返却した調査シートの主な内容としては，

1) 解答した分冊に含まれている問題についての正誤情報
2) 割り当てられなかった分冊に含まれている問題については推定正答確率の値
3) すべての問題に関して学習指導要領の項目事項を記載
4) 昨年度参加の学校には昨年度との経年比較結果もあわせて返却

であった．

第２部　理論的準備

2. 理論的準備

2.1 釣合い型不完備ブロックデザインの導入

全国的な学力調査のような大規模なアセスメントにおいては，参加する児童・生徒や学校に関する標本抽出の問題とともに，項目抽出の問題，すなわち，学習指導要領など調査対象とすべき領域をいかに偏りなく実施するテストの中に組み込むかという問題が存在する．従来の一冊子一斉実施型のテスト方式によって対象領域全体をカバーすることは，児童生徒への負担，実施コスト，時間的制限などの理由から実質的に不可能である．そこで，すべての対象領域からの多数の項目を準備し（それらの項目全体を項目プールとよぶ），それをいくつかの項目セットにまとめ，さらにその項目セットを組み合わせることで，何冊かの分冊を構成し，その分冊を偏りなく実施することを考える．このような方法をマトリックス・サンプリング（matrix -sampling[1]）と呼ぶ.

しかしながら，分冊を組み上げる際に，各分冊における項目セットの位置や出現回数，互いの組合せ回数などに偏りが生じると正しい結果が得られない．たとえばある項目セットが分冊の最後に必ず出現するような状況を考えると，児童生徒の疲労の効果によって，その項目セットに含まれる項目が本来もつ困難度よりもさらに難しい方向へバイアスがかかる可能性がでてくる等がその例である．言い替えれば，調査協力者の数，実施時間，実施コストがともに限られているという制約条件のもとで，得られる情報が最大かつ偏りのないものにする必要がある．具体的には，

a) どの項目セットも分冊全体を通して配置される回数が等しくなるようにする，
b) 項目セットの組み合わせパターンがたがいに同じ頻度で出現するようにする，
c) 冊子の中での項目セットの配置位置に偏りがないようにする，
d) 一つの冊子に含まれる項目数が同じになるようにする，

などが要求される．最後の d) については，項目セットの中に含まれる項目数をすべての項目セットで同一にしておくことで，各冊子に含まれる項目セット数を等しくすることと同値になる．このようなデザインを組むときに役に立つのが実験計画法にもとづく釣合い型不完備ブロックデザイン(balanced incomplete block design；BIBD)とよばれるものである．

2.1.1 釣合い型不完備ブロックデザイン

BIBD とは，一般に，v 個の処理が大きさ k の b 個のブロックで実行される実験計画のことを意味する．ここで k < v で，かつ b と k は b k が v の倍数になるように決められ，各処理が互いに同じ回数 r だけ出現し，どの処理の組みあわせも同じ数だけ λ 個のブロックに出現するデザインであり，簡単に BIB（v, b, k, r,λ）と標記される．λ のことを 2 つの処理の会合数と呼ぶ.

たとえばブロック計画 BIB（4,6,2,3,1）の場合，4 個の処理(v=4)をそれぞれ A,B,C,D とすると，大きさ 2 (k=2)の 6 つのブロック(b=6)は,

[1] 池田(2010)による“重複テスト分冊法”という訳語は実施形態に着目した場合のいわば意訳になる.

（A,B）（A,C）（A,D）（B,C）（B,D）（C,D）

のようになる．いずれの処理も全部で３回，いずれかのブロックにおいて出現し(r=3)，かつ任意の２つの処理の組合せはいずれも１回(λ=1)となっていることがわかる．このようにいわば出現回数等が全ての処理に対して均等となるようなデザインをさして釣合い型（balanced）と呼ぶ．また，ブロックの大きさ k が処理数ｖに等しい場合を完備な（complete）ブロック，それに対して，この例のようにその一部しか存在しない，すなわち k＜v の場合を不完備な(incomplete)ブロックと呼ぶ．

また，あるデザインが BIBD であるためには

$$bk = rv \quad \text{かつ} \quad \lambda(v-1) = r(k-1)$$

となることが必要である．実際，この例でも，6×2=3×4 かつ 1×(4－1)= 3 ×(2 －1)が成り立っていることが確認できる．ただし，BIBD となるための十分条件はわかっていない．また，例からも明らかなように，一般的な BIBD では上記 c)の条件が満たせない場合がほとんどである．そのため，本調査研究では，BIBD の特殊ケースであるユーデン方格法(Youden square design)を採用する．

2.1.2　ユーデン方格

ユーデン方格[2]とはラテン方格(Latin square design)から行や列を除くことによって得られるデザインのことである．ここでラテン方格とは，比較すべき処理の種類をｖ個とするとき，ｖ×ｖの正方形を考え，どの行にもどの列にも同じ処理が１個ずつ含まれるようにしたものである．たとえば，３種類の処理 A,B,C に対しての 3×3 のラテン方格は

A	B	C
B	C	A
C	A	B

B	C	A
C	A	B
B	C	A

A	C	B
B	A	C
C	B	A

など，また，４種類の処理 A,B,C,D に対しての 4×4 のラテン方格は

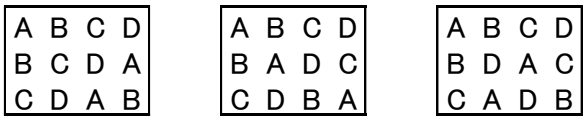
A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

A	B	C	D
B	A	D	C
C	D	B	A
D	C	A	B

A	B	C	D
B	D	A	C
C	A	D	B
D	C	B	A

などとなる．このうち 4×4 のラテン方格の，たとえば，第１列から第３列を取り出せば，

[2] ユーデン方格はフィッシャーとイェーツが 1938 年に導入したものであるが，その際，ユーデンを記念してこのように命名した（Upton 他(2010)の Latin square の項参照）．またユーデン「方格」となっているが実際には矩形であることにも注意が必要である．



がえられるが，それぞれが目的とするユーデン方格となっている．ここでは大規模な学力調査に利用することを念頭に，処理が A から G までの 7 つ存在する場合のユーデン方格を例にとって説明する．

1）まず表 1 のような 7×7 のラテン方格を考える．たとえば日照時間を列，土壌に含まれる水分の割合を行，肥料を A,B,C,D,E,F,G としたときの収穫量の多寡を比較するような場合である．ゆるやかに波打つ丘陵の斜面に広がった耕地を考えると，場所によって微妙に日照時間や土壌に含まれる水分の量などが異なっているのは自然である．そのようなときにその耕地を，たとえば，表 1 のように 7×7 の 49 の区画に分割しそこに 7 種類の肥料を施せば，列に関しても行に関しても各肥料を 1 回ずつ割り当てることが可能となる．

表 1　7×7 のラテン方格

A	B	C	D	E	F	G
B	C	D	E	F	G	A
C	D	E	F	G	A	B
D	E	F	G	A	B	C
E	F	G	A	B	C	D
F	G	A	B	C	D	E
G	A	B	C	D	E	F

2）次にこのラテン方格から第 3 行，および第 5 ,6, 7 行を抜き出した以下のようなデザインを考える．もはやすべての組合せはカバーできていないため不完備デザインとなっていることは明らかであるが，その一方で，上で述べた BIBD であるための条件は満たしていることがわかる．処理の組合せ回数（会合数）が λ= 2 であることから，パラメタ表示すると BIB(7,7,4,4,2)と書ける．さらに好ましいことに，一般の BIBD では必ずしも満足されない，処理の出現順序に関しても，いずれの処理もその出現する位置(position)が 4 つ行に関して 1 回ずつとなっていることもわかる．

表 2　4×7 のユーデン方格

C	D	E	F	G	A	B
E	F	G	A	B	C	D
F	G	A	B	C	D	E
G	A	B	C	D	E	F

3）同様に同じラテン方格から残りの部分である第 1，2，4 行を取り出すと以下のようなデザイン

になっていることがわかる. 不完備デザインになっていること, BIBD の必要条件を満たしていること, 出現位置が均等であることは４×７のユーデン方格の場合と同じである．ただし会合数は１（λ=1）となるため，パラメタ表示すると BIB（7,7,3,3,1）となる.

表３　３×７のユーデン方格

A	B	C	D	E	F	G
B	C	D	E	F	G	A
D	E	F	G	A	B	C

４）また，この場合に限っていえば，２）と３）で作成されたユーデン方格は互いに補集合となっている．本調査研究においては，２）で作成したユーデン方格は数学の分冊を準備するときに，３）で作成したユーデン方格は国語の分冊を準備する際に利用した．なお，ユーデン方格は PISA でも採用されているものである.

ただし，PISA などの国際学力調査でもユーデン方格法のことも BIBD と呼んでいるので，それにあわせて本調査研究でも特に断りのない限り，この名称を使う．また，実験計画法で使われている処理，ブロック，ブロックの大きさ等の用語を，大規模学力調査の文脈に合わせて理解しやすいように，それぞれ，処理を項目セット，ブロックを分冊，ブロックの大きさにたいして 1，2，･･･，k のように順序をつけ，それを位置と呼ぶこととする．なお，項目セットのことを PISA ではアイテム・クラスター(item cluster)と呼ぶこともある．さらに NAEP などでは，ユーデン方格を利用した分冊の組み上げデザインのことを分冊デザイン（Booklet design）と呼ぶこともある.

また，BIBD は項目セット数と分冊数および分冊ごとに項目セットを配置していく場合の位置(position)の数の特殊な組み合わせのもとでしか存在しないことが知られていて，たとえば，竹内(1989)，石井(1972a,1972b)の付表などを参照すればユーデン方格を含む BIBD の具体がわかる．しかしながら，現実的には PISA で採用されている 13 項目セット，13 分冊，4 位置，１会合数や，本研究で使用した 7 項目セット，7 分冊，3 位置，１会合数，または 7 項目セット，7 分冊，4 位置，2 会合数の３つのデザインくらいしか実用的なものはない.

さらにユーデン方格の場合，v = b，　k = r となるため，簡単のために，たとえば，上の３つのデザインを特に断りのない限り，それぞれ，BIB(13,4,1)，　BIB(7,3,1) ，　BIB(7,4,2)と略記することとする．すなわち，このような書き方をすれば BIB のうちのユーデン方格を示し，BIB(v=b, r=k,λ)という意味となる.

以上を，PISA を例にとって具体的に述べると以下のようになる．表中の分冊番号，位置番号以外の数値がアルファベットに代わって項目セットを表す番号となる.

表 4　PISA で採用されている BIB(13,4,1)デザイン

位置	分冊番号												
	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1	2	3	4	5	6	7	8	9	10	11	12	13
2	2	3	4	10	6	13	12	9	1	11	5	8	7
3	4	10	11	5	7	12	9	2	3	6	13	1	8
4	7	12	8	9	3	4	11	6	13	1	2	5	10

実際，表 4 に示されている BIB(13,13,4)デザインでは 13 個の項目セットは 13 冊の分冊のいずれかに必ず 4 回配置されており(上記の条件 a），項目セット間の組み合わせパターンはすべての組み合わせが 1 回ずつ出現している(条件 b)．さらに，いずれの項目セットも 4 つの位置のどこかに出現し(c)，いずれの分冊にも同数の項目セットが含まれていることがわかる(d).

また，PISA2006 では表 4 のデザインに基づき，13 個の項目セットに科学的リテラシーのための項目セットを 7 個（表 4 の項目セット番号では 1～ 7 に該当する），数学的リテラシーのための項目セットを 4 個（表 4 では 8～11），リーディング・リテラシーのための項目セットを 2 個（表 4 では 12～13）を配置している．これを休憩時間約 5 分をはさんで前後 1 時間ずつで，合計 2 時間をかけて 1 人の生徒が一つの分冊(booklet)を解くようにテストの設計がなされている．もし全ての項目を 1 人の生徒が回答するとすれば，6.5 時間となる。また PISA2006 のメイン調査であった科学リテラシーに関しては，各分冊の中に科学リテラシーへの態度・関心を尋ねるような質問項目も含まれている．

しかしながら，我が国において学校の通常時間割のなかで調査をする状況を考えると，同じ時間の中で，一つの分冊に含まれる複数の異なる科目の問題を解くような習慣がないこと，おなじくテストの中に態度をたずねるような項目を含めると，生徒がどちらを重要視してよいのかわからず，不必要な精神的負担をかける可能性があること，後述するように個人へのフィードバックに関して協力校などからの強い要請があるため，個人の尺度値の精度を確保する必要上ある程度以上の項目数を分冊の中に含めなければならないなどの理由から，1 校時の中でひとつの科目に限定してテストの設計をおこなった．

厳密にいえば，児童生徒のテストへの慣れや疲労効果などにより，たとえば数学を先に解いてから国語を解いた方が，その逆より国語の得点が系統的に良くなる可能性（いわゆる持ち越し効果；carryover effect）やその逆の可能性などを消去できないことになる．しかし，その効果はそれほど大きなものではないと経験的に見込むことができること，学校現場への負担をなるべく避けるためという理由により，本調査では数学と国語の実施順をランダムにするなどの制限は設けなかった．ただし，全ての分冊が偏りなく配付されるように，分冊の番号順に 1 冊ずつを並べたもの一つの単位とし，それを必要回数くりかえすことで，すべての生徒分の冊子を準備し，次に調査対象となるクラスの人数にあわせて最初から順にクラスごとに区切って配付の準備をした．

また国語に関しては，まず題材文を示した上で，次に，それに対して小問をいくつか構成する，い

わゆる大問形式をとる必要があるため，同じ時間のもとでは項目セット数を必然的に減らさざるを得ない．平成２３年度調査研究（柴山他，2011)の成果からその大問数は 3，大問を構成する小問数（項目数）は 4 というのが妥当であるとの結論をえているため，国語に関しても 7 分冊 7 項目セット 3 位置の BIBD を本研究では試みた．このことによって全ての項目を 1 人の生徒に実施するとすると数学については 3.5 校時, 国語については 2.3 校時になるものをいずれも一校時で実施することが可能となる。

数学と国語のそれぞれのユーデン方格の実際は表５，６に示すとおりである．これは先に説明した 7×7 のラテン方格からのユーデン方格の作成の際に得た互いに補集合の関係にある２つのデザインでもある．また，表 7 は国語で採用したユーデン方格（表６）をフィッシャーの表現によって表記したものである．この場合，表頭に項目セットの番号，表側に分冊の番号が入り，表中の数値は位置の番号を表すことになる．

表 5　本調査研究（数学）で採用した BIB(7,4,2)デザイン

位置	分冊番号						
	1	2	3	4	5	6	7
1	**3**	**4**	**5**	**6**	**7**	**1**	**2**
2	**5**	**6**	**7**	**1**	**2**	**3**	**4**
3	**6**	**7**	**1**	**2**	**3**	**4**	**5**
4	**7**	**1**	**2**	**3**	**4**	**5**	**6**

表 6　本調査研究（国語）で採用した BIB(7,3,1)デザイン

位置	分冊番号						
	1	2	3	4	5	6	7
1	1	2	3	4	5	6	7
2	2	3	4	5	6	7	1
3	4	5	6	7	1	2	3

表 7　フィッシャーによる表現（Fisher's representation）

		項目セット						
		1	2	3	4	5	6	7
分冊番号	1	1	2		3			
	2		1	2		3		
	3			1	2		3	
	4				1	2		3
	5	3				1	2	
	6		3				1	2
	7	2		3				1

2.1.3　ユーデン方格の構造

さらに表 7 を行列表記すると以下の行列 C のようになる．たとえば，この行列の第 2 行第 5 列には 1 が表示されているが，これは分冊２に項目セット５が含まれていることを示す．またその右横は 0

となっているが，これは項目セット6は分冊 2 には含まれていないことを示している．この行列は生起行列（incidence matrix）と呼ばれることもある．これを使ってユーデン方格の構造を検討しておく．

$$C = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

まず，CC′を求めると

$$CC' = \begin{bmatrix} 3 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 3 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 3 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 3 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 3 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 3 \end{bmatrix}$$

が得られる．これは分冊相互に共通に含まれる項目セットがどの分冊の組合せでも 1 セットずつあることを示している．また，対角要素は各分冊に含まれる項目セットの数を示しており，いずれの分冊についても 3 つの項目セットが含まれていることを示している. BIBD の一般的な表現を使えばいずれのブロックの大きさも３であり（ｋ＝３），処理の会合数はいずれも等しく１である（λ＝１）．

次に，C′C を求めると，

$$C'C = \begin{bmatrix} 3 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 3 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 3 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 3 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 3 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 3 \end{bmatrix}$$

がえられる．この場合，対角要素の数は各項目セットがいずれも等しく 3 回使われていることを示している．さらに，非対角要素がすべて１となっているが，これはいずれも項目セットもペアとなる回数は１回であることを示している．いいかえれば，いずれの処理の繰り返し数も３（ｒ＝３）である．

同様に数学のデザイン行列 D を考えると以下のようになる．

$$D = \begin{bmatrix} 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 & 0 \end{bmatrix}$$

デザイン行列 D が BIB である必要条件を満たしていることは明らかである．そこで，まず，DD′を求

めると

$$DD' = \begin{bmatrix} 4 & 2 & 2 & 2 & 2 & 2 & 2 \\ 2 & 4 & 2 & 2 & 2 & 2 & 2 \\ 2 & 2 & 4 & 2 & 2 & 2 & 2 \\ 2 & 2 & 2 & 4 & 2 & 2 & 2 \\ 2 & 2 & 2 & 2 & 4 & 2 & 2 \\ 2 & 2 & 2 & 2 & 2 & 4 & 2 \\ 2 & 2 & 2 & 2 & 2 & 2 & 4 \end{bmatrix}$$

が得られる．これは分冊相互に共通に含まれる項目セットがどの分冊の組合せでも 2 セットずつあることを示している．また，対角要素は各分冊に含まれる項目セットの数を示しており，いずれの分冊についても 4 つの項目セットが含まれていることを示している. BIBD の一般的な表現を使えばいずれのブロックの大きさも 4 である（k＝4）であり，会合数はλ＝2 となることがわかる．

一方，D′D を求めると，やはり

$$D'D = \begin{bmatrix} 4 & 2 & 2 & 2 & 2 & 2 & 2 \\ 2 & 4 & 2 & 2 & 2 & 2 & 2 \\ 2 & 2 & 4 & 2 & 2 & 2 & 2 \\ 2 & 2 & 2 & 4 & 2 & 2 & 2 \\ 2 & 2 & 2 & 2 & 4 & 2 & 2 \\ 2 & 2 & 2 & 2 & 2 & 4 & 2 \\ 2 & 2 & 2 & 2 & 2 & 2 & 4 \end{bmatrix}$$

がえられる．この場合，対角要素の数は各項目セットがいずれも等しく 4 回使われていることを示している．さらに，非対角要素がすべて 2 となっているが，これはいずれも項目セットもペアとなる回数は 2 回であることを示している．いいかえれば，いずれの処理の繰り返し数も 4（r＝4）である．

行列 C と D はいずれもユーデン方格となっているが，全ての条件が等しければ，調査のコストを考えたときには C の方がすぐれていることになる．具体的には，もし数学に限って，BIB(7,3,1)か BIB(7,4,2)のいずれかを選ぶとするならば，個々の生徒への負担をなるべく避けるという意味では BIB(7,3,1)で実施した方が 1 つの分冊の中の項目セット数を 4 から 3 に減らせる分, 実施時間もたとえば 40 分から 30 分になるなどの意味で好ましい．項目母数の推定精度への影響は大規模アセスメントの文脈を考えれば，十分なサンプルサイズをとっておけばほとんど問題はないレベルであろう．

なお，一般の BIBD，たとえば BIB(4,6,2,3,1)の場合，その生起行列 E は，

$$E = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}$$

となる．EE′を求めると

$$EE' = \begin{bmatrix} 3 & 1 & 1 & 1 \\ 1 & 3 & 1 & 1 \\ 1 & 1 & 3 & 1 \\ 1 & 1 & 1 & 3 \end{bmatrix}$$

となるが，一方で，E′E は，

$$E'E = \begin{bmatrix} 2 & 1 & 1 & 1 & 1 & 0 \\ 1 & 2 & 1 & 1 & 0 & 1 \\ 1 & 1 & 2 & 0 & 1 & 1 \\ 1 & 1 & 0 & 2 & 1 & 1 \\ 1 & 0 & 1 & 1 & 2 & 1 \\ 0 & 1 & 1 & 1 & 1 & 2 \end{bmatrix}$$

である．この場合，ユーデン方格法とは異なり，共通の項目セットを含まない分冊が生じることになる．

2.1.4　平成 23 年度数学の分冊デザインとの比較

さらに平成 23 年度の重複テスト分冊法の数学のデザインを検討すると，その分冊デザインは以下のようであった．これは等化デザインとしては NEAT（Nonequivalent groups with anchor test）デザインとなってはいるため，経年比較をとらえるためのテスト等化としてのデータ収集デザインとしては支障はないが，分冊デザイン上は BIBD ではない．

表 8　数学の分冊における項目セットの配置

分冊	位置1	位置2	位置3	位置4	位置5	位置6	位置7	位置8
S1	B3	B5	B11	B12	B7	B9	B15	B16
S2	B4	B6	B12	B13	B1	B8	B10	B16
S3	B5	B7	B13	B14	B1	B2	B3	B9
S4	B6	B8	B14	B15	B2	B4	B10	B11
S5	B7	B9	B15	B16	B4	B6	B12	B13
S6	B1	B8	B10	B16	B3	B5	B11	B12
S7	B1	B2	B3	B9	B5	B7	B13	B14
S8	B2	B4	B10	B11	B6	B8	B14	B15

具体的に検討するため，これを生起行列 F で表すと以下のようになる．

$$F = \begin{bmatrix} 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

さらに FF′をもとめると

$$FF' = \begin{bmatrix} 8 & 2 & 4 & 2 & 5 & 5 & 4 & 2 \\ 2 & 8 & 2 & 4 & 5 & 5 & 2 & 4 \\ 4 & 2 & 8 & 2 & 3 & 3 & 8 & 2 \\ 2 & 4 & 2 & 8 & 3 & 3 & 2 & 8 \\ 5 & 5 & 3 & 3 & 8 & 2 & 3 & 3 \\ 5 & 5 & 3 & 3 & 2 & 8 & 3 & 3 \\ 4 & 2 & 8 & 2 & 3 & 3 & 8 & 2 \\ 2 & 4 & 2 & 8 & 3 & 3 & 2 & 8 \end{bmatrix}$$

のようになる．確かにいずれの分冊も項目セットを等しく 8 つずつ含んでいるが，共通する項目セットの数は 2 から 8 までと不均等になっていることがわかる．また，F′F は

$$F'F = \begin{bmatrix} 4 & 2 & 3 & 1 & 3 & 1 & 2 & 2 & 2 & 2 & 1 & 2 & 3 & 2 & 0 & 2 \\ 2 & 4 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 0 & 2 & 4 & 2 & 0 \\ 3 & 2 & 4 & 0 & 4 & 0 & 3 & 1 & 3 & 1 & 2 & 2 & 2 & 2 & 1 & 2 \\ 1 & 2 & 0 & 4 & 0 & 4 & 1 & 3 & 1 & 3 & 2 & 2 & 2 & 2 & 3 & 2 \\ 3 & 2 & 4 & 0 & 4 & 0 & 3 & 1 & 3 & 1 & 2 & 2 & 2 & 2 & 1 & 2 \\ 1 & 2 & 0 & 4 & 0 & 4 & 1 & 3 & 1 & 3 & 2 & 2 & 2 & 2 & 3 & 2 \\ 2 & 2 & 3 & 1 & 3 & 1 & 4 & 0 & 4 & 0 & 1 & 2 & 3 & 2 & 2 & 2 \\ 2 & 2 & 1 & 3 & 1 & 3 & 0 & 4 & 0 & 4 & 3 & 2 & 1 & 2 & 2 & 2 \\ 2 & 2 & 3 & 1 & 3 & 1 & 4 & 0 & 4 & 0 & 1 & 2 & 3 & 2 & 2 & 2 \\ 2 & 2 & 1 & 3 & 1 & 3 & 0 & 4 & 0 & 4 & 3 & 2 & 1 & 2 & 2 & 2 \\ 1 & 2 & 2 & 2 & 2 & 2 & 1 & 3 & 1 & 3 & 4 & 2 & 0 & 2 & 3 & 2 \\ 2 & 0 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 4 & 2 & 0 & 2 & 4 \\ 3 & 2 & 2 & 2 & 2 & 2 & 3 & 1 & 3 & 1 & 0 & 2 & 4 & 2 & 1 & 2 \\ 2 & 4 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 0 & 2 & 4 & 2 & 0 \\ 0 & 2 & 1 & 3 & 1 & 3 & 2 & 2 & 2 & 2 & 3 & 2 & 1 & 2 & 4 & 2 \\ 2 & 0 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 4 & 2 & 0 & 2 & 4 \end{bmatrix}$$

となり，項目セットが使用される回数はいずれも等しく 4 回であるが，項目セット同士のペアの数は 0 から 4 と不均等となっていることは明らかである．このような不均衡は，たとえば，各学校や教育委員会の取り組み，あるいは学習状況調査などの外的情報と学力データを結びつけて分析しようとするとき，分冊を受けた集団や項目セットに割り当てられて集団が母集団からの偏りのない標本となっていることを担保できていないことを意味している．このような意味から大規模なアセスメントにおいては BIBD は分冊デザインとして本質的な役割を担うのである．[3]

BIBD については竹内(0989)，石井(0972a,0972b)などを参照されたい．また，大規模学力調査におけるさまざまなデザイン(Booklet Design)の理論と実際を詳細に解説した Frey ら(2009)，マトリックス・サンプリングにもとづくテストを作成・実施する際のコストを従来の方式と比較した Child ら(2003)の論文も参考になる．

[3]実際の効果については本報告書「７．３　BIBD 導入の効果」を参照のこと

2.2 推算値

一般に学力調査の結果を利用するときに，個人スコアに注目するのか，それとも集団統計量に注目するのかの区別は重要である．児童生徒ひとりひとりの学習指導が目的なら前者が利用されるものの，行政レベルで見たときのその施策の有効性や地域間格差，経済格差などのいわばマクロな視点での考察には後者の集団統計量が問題となる．これまで，我が国においては，その両者が明確に区別されていなかったために，必ずしもマクロなレベルにおける判断の際に必要となる十分な精度の集団統計量が準備されていたわけではない．

本節では，このような問題に対して，すでに国際的な学力調査において集団の能力分布を推定する際に利用される推算値（plausible value）についての概説を試みる．ただし，推算値に関しては邦文によるまとまった文献がないため，2.2.1～2.2.3 節では，Wu（2004）を主として参考としながら，推算値の必要性・定義・利用法について説明する．また, 2.2.4 節では, von Davier, Gonzalez, and Mislevy（2009）のシミュレーション結果を通して推算値を利用することの有用性を考察する．

2.2.1 推算値の必要性

推算値（plausible value）は，全米学力調査（National Assessment of Educational Progress, NAEP）1983-84（Beaton, 1987）のデータを分析するため，多重補完法（multiple imputation）に関する Rubin の研究に基づき，Mislevy, Sheehan, Beaton, Johnson によって最初に開発された．NAEP が補完法の導入を開始して以降，集団を対象とした調査の報告に推算値を利用することが推奨されてきた．推算値は, その後のすべての NAEP と TIMSS（Trends in International Mathematics and Science Study）でも利用されており，現在では PISA（Programme for International Student Assessment）や PIACC(Programme for the International Assessment of Adult Competencies)でも利用されるに至っている．

村木（2006）によれば，NAEP の目的は，個々人の能力推定値を算出することではなく，全米の児童生徒全体の学力分布，または人種などによる下位集団の学力分布をできるだけ正確に推定することである．たとえば，ある人種グループにおいて，この辺りの地域に住んでいる，このくらいの階級の人たちの能力分布の平均や分散がどうなっているのか，などを知るのが NAEP の目的である．

NAEP を含む国際的な学力調査では，試験のデザインに重複テスト分冊法が利用されているため，一人の受験者が解答する冊子の項目数は，個人別の報告を目的とする試験と比べてはるかに少ない．そのため，個々人の能力推定値から集団の能力分布を推定する方法では，分散の過大評価あるいは過小評価が生じ，集団間での正確な分散の比較はできない．そうなると，各集団の能力分布の平均値に統計的な差があるかどうかの判断にも支障をきたすことになる．その解決策の一つがベイズ統計の枠組みを用いた推算値の方法論である．

推算値は，項目数が比較的少ない場合でも，集団の能力分布の分散を正確に推定できるとともに，

能力分布のパーセンタイルも正確に推定できる．それゆえ，ある能力基準以上の学生の割合のように，一定得点以上の受験者の割合を予測する上でも推算値は重要な役割を果たす．

2.2.2　推算値の定義

通常，1, 2, 3 母数ロジスティックモデルなどの項目反応モデルでは，受験者の反応パターンと項目母数の推定値を用いて受験者の能力母数 θ を推定する．このとき，尤度関数の最大値を与える θ を推定値とする場合は最尤推定値，事後分布の最大値を与える θ を推定値とする場合は MAP（maximum a posteriori）推定値，事後分布の期待値を与える θ を推定値とする場合は EAP（expected a posteriori）推定値が得られる．それに対し，受験者における能力母数 θ の事後分布からの無作為標本を推算値という．図 2.2.1 に，最尤推定値，MAP 推定値，EAP 推定値，推算値の概念図を示す．

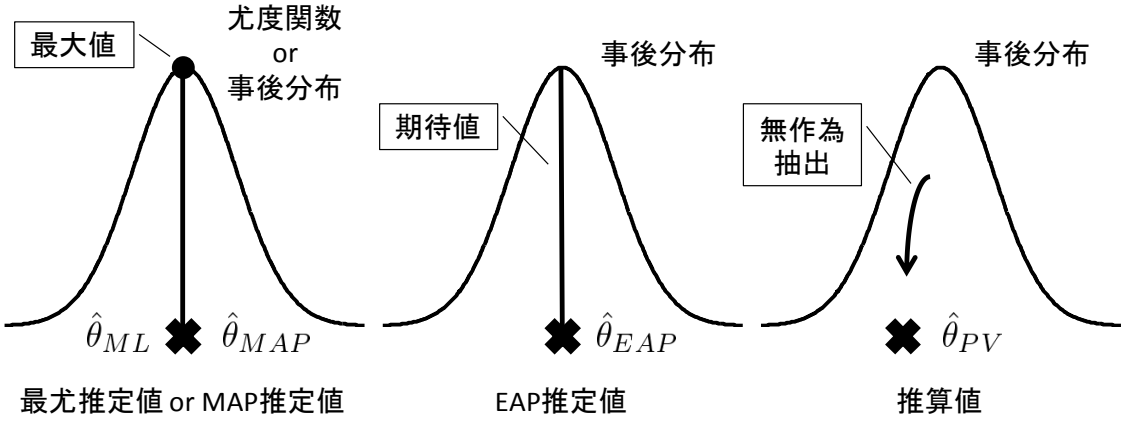


図 2.2.1　最尤推定値，MAP 推定値，EAP 推定値，推算値の概念図

数学的な説明のため，受験者の項目反応パターンを $\mathbf{x}$，能力母数を θ，尤度関数を $f(\mathbf{x}|\theta)$とする．さらに，能力母数 θ をベイズ推定するため，その事前分布として通常は正規分布 $g(\theta)\sim N(\mu,\sigma^2)$を仮定する．このとき，事後分布 $h(\mathbf{x}|\theta)$は，

$$h(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)g(\theta)}{\int f(\mathbf{x}|\theta)g(\theta)d\theta} \tag{2.2.1}$$

と表される．ある受験者の項目反応パターンが $\mathbf{x}$ であれば，その受験者の能力母数 θ の事後分布は(2.2.1)式で与えられる．この(2.2.1)式からの無作為標本が項目反応パターン $\mathbf{x}$ をもつ受験者の推算値である．

図 2.2.1 からも想像できるように，推算値は，個々人の能力推定値の一種と考えることができる．ただし，推算値には他の推定値と全く異なる性質があり，それは推定値が決定論的ではなく，確率論的であるということである．推算値は，事後分布からの無作為標本であるため，受験者の能力推定値についての情報を与えてくれるだけでなく，その能力推定値に付随する不確かさについての情報も与えてくれる．この点こそが推算値によって集団の能力分布の分散やパーセンタイルを正確に推定できる理由の一つである．なお，事後分布 $h(\mathbf{x}|\theta)$からの標本抽出には，棄却サンプリング（rejection sampling）（津田，1995 など）などを利用することができる．

2.2.3 推算値の利用

能力分布の集団統計量を知りたい場合，一つの方法としては，個々人の能力値を推定し，対象となる集団の平均，分散，パーセンタイルなどを計算する方法が考えられる．受験者の能力値が最尤推定値の場合，個々人の能力値の平均は，能力分布の母平均の不偏推定値であることが示されている．しかし，個々人の能力値の分散については，能力分布の母分散を過大評価してしまう．一方，受験者の能力値が EAP 推定値の場合，平均については能力分布の母平均の不偏推定値であるものの，分散については母分散を過小評価してしまうことが示されている．最尤推定値と EAP 推定値の両方の場合において，受験者数を増やしても分散推定値の偏りは消えないものの，項目数を増やせば分散推定値の偏りは小さくなる．

前節で説明したように，推算値は，受験者における能力母数 θ の事後分布からの無作為標本である．また，個々人の θ についての事後分布を集めた分布は，その集団の能力分布の推定分布を与える．それゆえ，対象となる集団の推算値の組は，その集団の能力分布からの無作為標本とみなすことができる．これは，非常に重要な考え方であり，推算値によって集団統計量の不偏推定値を得られることの根拠となっている．実際には，2 組以上の推算値を利用して集団統計量を推定するものの，対象集団のたった 1 組の推算値を用いるだけでも不偏推定値を得ることができる．これは，最尤推定値や EAP 推定値を用いて集団統計量を算出すると，分散推定値に偏りが生じてしまうのと対照的である．

米国・オーストラリアなどにおける国家レベルでの学力調査，PISA や TIMSS，PIAAC など国際的な学力調査においては，Little and Rubin（2002）の多重補完法に基づく推算値が標準的に用いられている．Little and Rubin（2002）では，一人の児童・生徒に対して 5 つの推算値を生成し，それらを使って集団における平均や分散などの統計的特性を推定することを推奨している．その際の推定方法としては，

[PV-R]　関心のある統計量を推算値の組ごとに計算し，それらの統計量を平均する．
[PV-W]　各受験者の推算値を平均し，1 組の平均値を用いて関心のある統計量を計算する．

の二つの計算方法が考えられる．

しかしながら，推算値の定義から，推算値の使い方としては前者の計算方法が正しく（”R”は Right），後者の計算方法は誤りである（”W”は Wrong）．そのため，統計量の計算量を減らそうとして，[PV-W]のように各受験者の推算値を平均してはいけない．図 2.2.1 からもわかるように，各受験者の推算値を平均することは，5 点を使って大雑把に EAP 推定値を計算していることと同じである．すでに述べたとおり，EAP 推定値による分散推定値は不偏推定値にならないので，[PV-W]の計算方法は明らかに誤りである．

K 組の推算値によって算出した関心のある統計量の補間分散は，群内分散と群間分散への分散分解の式のように，

$$\hat{V}_{IMP} = (1 + \frac{1}{K})[\frac{1}{K-1}\sum_i (M_{PVi} - \bar{M}_{PV})^2] + \frac{1}{K}\sum_i \hat{V}(M_{PVi}) \qquad (2.2.2)$$

と表される（Little & Rubin，2002）．ここで，M_{PVi} は推算値を用いて算出した集団 i における関心のある統計量，$\dot{M}_{PV}$ は関心のある統計量の集団についての平均値，$\hat{V}(M_{PVi})$は集団 i における関心のある統計量の誤差分散の推定量である．(2.2.2)式の正の平方根が補間の標準誤差に相当する．

前節で述べたように，NAEP が補完法の導入を開始して以降，集団を対象とした調査の報告に推算値を利用することが推奨されてきた．しかし，ここで強調すべきは，推算値が個人の能力値の推定については不向きなものであり，またそのような使用を目的としたものでもないということである．推算値は，能力値 θ の事後分布からの無作為標本なので，同じ得点パターンをもつ受験者が二人いたとしても，結果として異なる能力推定値が推定されてしまう．そうなると，二人の受験者から抗議を受けることは確実であり，統計的には問題がなくても，社会的には受入れられる話ではない．したがって，個々人の能力推定値（個人スコア）としては，これまで通り最尤推定値，EAP 推定値，MAP 推定値のような点推定値の方が向いている．その意味で，推算値は，あくまでも集団統計量について公式的な報告を行ったり，同じデータに関する二次的な分析を行ったりするためのツールであると言ってもよい．

その具体例の一つとして，推算値は，合否の分割点や対象集団における能力分布のパーセンタイルを点推定値よりも正確に推定できることがあげられる．たとえば，4 項目から構成されるテストがあれば，受験者の得点は 0, 1, 2, 3, 4 点のいずれかである．1 母数ロジスティックモデルを利用する場合，各得点につき一つの能力推定値が対応するので，図 2.2.2 のような能力母数 θ の事後分布が得られる．図中の横軸（θ）に向かって伸びている点線の矢印は，EAP 推定値を表している．

いま，能力値が-1 未満の受験者の割合に興味があるとする．EAP 推定値の場合，-1 未満の受験者の割合は，0 点をとった受験者の割合に等しい．実際，EAP 推定値は離散的であるため，0 点と 1 点の間のどんな分割点についても同じ割合が得られてしまう．それに対し，事後分布の曲線で囲まれた-1 未満の領域をみると，事後分布は連続的であり，すべての得点からの寄与があることがわかる．推算値は，事後分布からの無作為標本なので，各得点の事後分布からの寄与を EAP 推定値よりも正しく反

映させることができる．

このように，分割点未満の推算値の割合は，点推定値よりも正確な受験者の割合の推定値を与えてくれるのである．言い替えれば，推算値を用いることによって，点推定値がもつ離散的な性質に関連する問題を克服することができるのである．同様に，ここでは詳細を省くものの，パーセンタイル値の推定などにおいても，離散的な能力推定値の間を補間する必要があるという問題を推算値によって克服することができる．

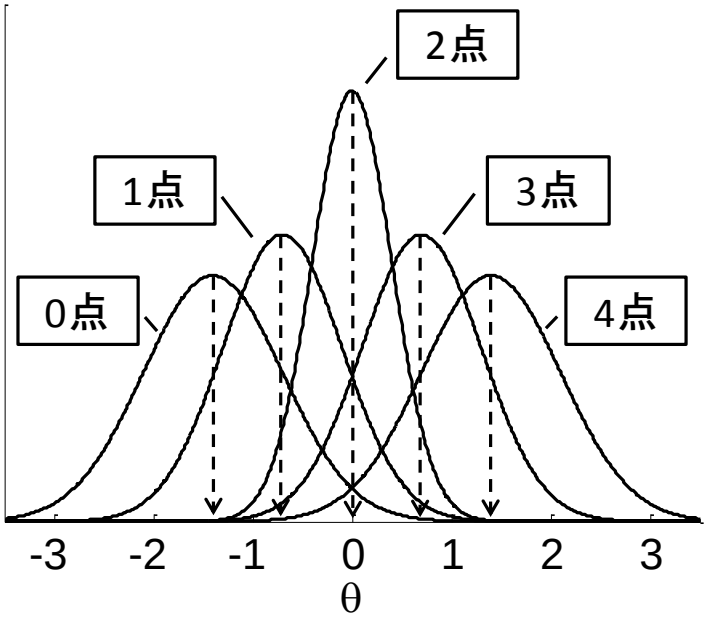


図 2.2.2　各得点の事後分布と EAP 推定値（Wu, 2004, Figure を単純化したもの）

2.2.4　推算値を用いる利点

von Davier et al.（2009）は，”What are plausible values and why are they useful?”という論文の中で，大規模調査において推算値が利用されない場合，あるいは正しく利用されない場合にどんな悪影響が生じるかを例示している．詳細は原論文に譲るものの，そこで報告されているシミュレーション結果から判断して，集団統計量の計算に推算値を用いる利点は以下のように整理できるであろう．

1）Von Davier et al.（2009）によるシミュレーションの範囲では，対象集団の能力分布の平均および標準偏差を推定する場合，真値にもっとも近い推定値が得られる方法は，推算値が正しく利用された場合であった．さらに，能力分布のパーセンタイル値をもっとも正確に推定できたのも，やはり推算値が正しく利用された場合であった．とくに項目数が 8 項目と少ない場合，その傾向は顕著であった．

2）一方，項目数が 16 項目，24 項目と増えるにつれて，個々人の能力推定値から集団の能力分布を推定する方法と推算値を用いて推定する方法との差は小さくなる傾向があることもシミュレーション結果から読み取れる．解析的に何項目であるとは決定できないものの，個々人の能力値を報告できるほどの項目数を利用できる場合は，集団の能力分布を推定するのに推算値を用いる必要はない可能性は残っている．

3）しかしながら，個々人の能力の最尤推定値と EAP 推定値を用いて集団の能力分布を推定しても，標準偏差については不偏推定値にならない結果がシミュレーションによって示されていることは重要である．また，大規模な学力調査の場合，調査結果が母集団の中の多くの人に影響を与える教育的政策に関する意思決定のために利用されることがある．たとえば，母集団の中で基準点に到達する学生の割合を計算する場合，わずか 2%の差によって数多くの学生の処遇が変わってしまうこともある．それらを考慮すると，全国的な学力調査において集団の能力分布を推定するときは，常に推算値を利用した方がよいと判断される．

以上の諸点は，全国的な学力調査において追加分析として，たとえば教育社会学的な観点からのアプローチを試みる際には，ほとんど顧みられることはなかった．しかし，国際的にも推算値の利用が標準となっていることもさりながら，何よりもまず，マクロな視点から我が国の教育施策に資する情報を得るには，推算値の方法論は，今後，必要不可欠となる調査分析ノウハウの一つである．

2.3　学力向上の要因を探るためのマルチレベル分析

2.3.1　階層化されたデータとマルチレベル分析

本研究でも扱っているような教育場面での大規模なデータでは，それが「階層構造」をなしていることが多い．たとえば，全国の中学生を母集団と想定したテストデータについて考えてみると，中学生個人は，いずれかの中学校に所属している．また各中学校には複数の学級があり，各中学生はいずれかの学級に所属していることになる（このような時に，各生徒は教室に「ネスト」されているという表現を用いることもある）．もう少し大きな視点で見ると，各中学校はいずれかの都道府県にネストされている．この階層構造を図式化すると図 1 のようになる．

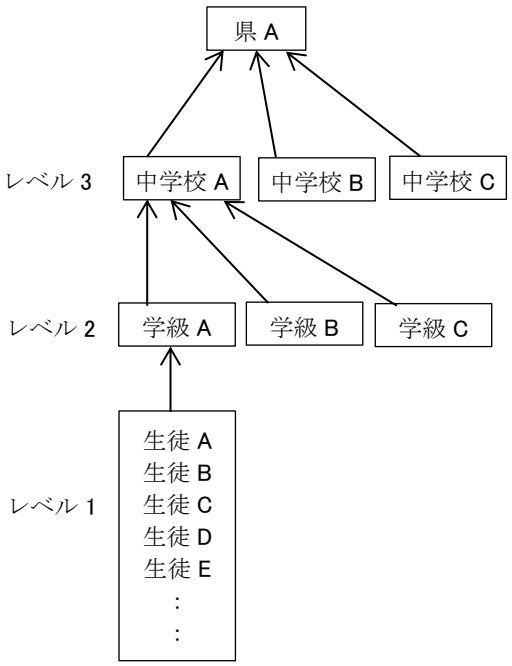


図 2.3.1　階層構造データの例

このような階層構造を持つデータに対して分析を行う場合，「階層」がどの程度影響をするのかを考慮する必要があるケースも多いであろう．たとえば，中学生におけるある「英単語テスト」と「英文読解テスト」の関係を調査するとしよう．この時，図 2.3.1 のような階層構造を持つデータにおいて，たとえば「学級」という階層が大きく影響を与える可能性がある．英語を担当する教師のパーソナリティや指導の進め方などが影響して，ある学級では「英文読解テスト」における「英単語テスト」の影響が大きく，別の学級では影響が小さいかもしれない．このように「学級」の影響がどの程度なのかを調べるために，一つの方法としては実験計画を組んで調査することが考えられる．厳密に実験計画を立てるためには，調査目的に照らし合わせて想定しうる教師の要因（パーソナリティや指導の進め方）を統制し，それらに無作為に生徒を割り当てるという必要がある．しかしながら現実場面では，調査段階では学級ごとにすでに生徒は割り当てられており，教師ごとに生徒を無作為に割り当てる，といったことが不可能であることが多い．このようなデータに対して，マルチレベル分析を用いるこ

とで，「学級（教師）」の影響および生徒の影響などを調べることが可能となる．

また，図 2.3.1 における各階層を「レベル」と表現する．レベル 1 が各生徒，レベル 2 が学級，レベル 3 が中学校となる．マルチレベル分析では，このレベルごとに異なる分析モデルを構築することで，その影響を検討することができる． レベルの数については，原理的には多数のレベル（階層）を表現することが可能ではあるが，現実の分析場面では，レベルの数が多すぎるとモデルが複雑になり，結果の解釈も難しくなるため，2 から 3 程度とすることがほとんどである．

2.3.2 ランダム切片モデル

先の中学生における「英単語テスト」と「英文読解テスト」の関係を検討するための例で，レベルを 2 つ（レベル 1 が各生徒，レベル 2 が中学校）にしたものを考えてみる．「英単語テスト」を変数 x_{ij} とし，「英文読解テスト」を y_{ij} として，両者の関係を検討することにする．ここで添え字 i は各生徒を表し，j は中学校を表す変数である．「英単語テスト」を説明変数（独立変数），「英文読解テスト」を従属変数とする単回帰モデルが考えられるが，マルチレベル分析では，以下のようなモデル式を立てる．

$$\text{レベル 1：} \boldsymbol{y_{ij} = \beta_{0j} + \beta_1 x_{ij} + e_{ij}}, \tag{2.3.1}$$

$$\text{レベル 2：} \boldsymbol{\beta_{0j} = \beta_0 + u_{0j}}, \tag{2.3.2}$$

レベル 1 の(2.3.1)式は，切片 β_{0j}，傾き β_1，誤差項 e_{ij} とする通常の単回帰モデルと全く同じである．ただし切片には，中学校を表す添え字 j が付されている．この切片 β_{0j} について，レベル 2 の(2.3.2)式でモデル化される．(2.3.2)式では切片 β_{0j} について，データ全体（全生徒）の切片 β_0 と j 番目の中学校の切片が β_0 から離れている程度を表す u_{0j} に分解している．通常の単回帰分析で誤差項 e_{ij} を平均 0，分散を σ_e^2 とする正規分布に従う確率変数（$e_{ij}\sim N(0,\sigma_e^2)$）として扱うことと同様に，$u_{0j}$ についても平均 0，分散を $\sigma_{u_0}^2$ とする正規分布に従う確率変数（$u_{0j}\sim N(0,\sigma_{u_0}^2)$）として扱う．マルチレベル分析では，各中学校が$\beta_0$ からどの程度離れているかを表す u_{0j} そのものよりも，「中学校」全体としての β_0 からの乖離を示す $\sigma_{u_0}^2$ が重要な指標となる（$\sigma_{u_0}^2$ が計算されたのちに，各 u_{0j} の値を推定することも可能である）．

(2.3.2)式を(2.3.1)式に代入し，

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_{0j} + e_{ij}, \tag{2.3.3}$$

と
して表現することもある．ここで β_0 や β_1 を「固定部（fixed part）」，u_{0j} や e_{ij} を「ランダム部（random part）」と呼ぶこともある．

以上のモデルが，切片部分についてランダム部を導入した「ランダム切片モデル」と呼ばれるものである．ランダム切片モデルについては，次のような状況が想定される．先の「英単語テスト」と「英文読解テスト」について，3 つの中学校のデータを散布図にしたものが図 2.3.2 である（実際のマルチ

レベル分析では，$\sigma_{u_0}^2$ の推定誤差を小さくするために，レベル 2（中学校）の数はもう少し大きい必要がある）．中学校ごとに相関係数を算出すると.57 から.76 となるが，データ全体での相関係数は.27 と大きく低下する．そして，データ全体での回帰直線（図中点線）に比べて，マルチレベル分析から得られる回帰直線（3 本の実線）の傾きが大きく違うことが見て取れる．通常の回帰分析では 1 本の回帰直線だけを考え，直線から各データの乖離をすべて誤差としていたが，ランダム切片モデルでは各中学校の平均値の違いによる影響を考慮に入れることが可能となる．

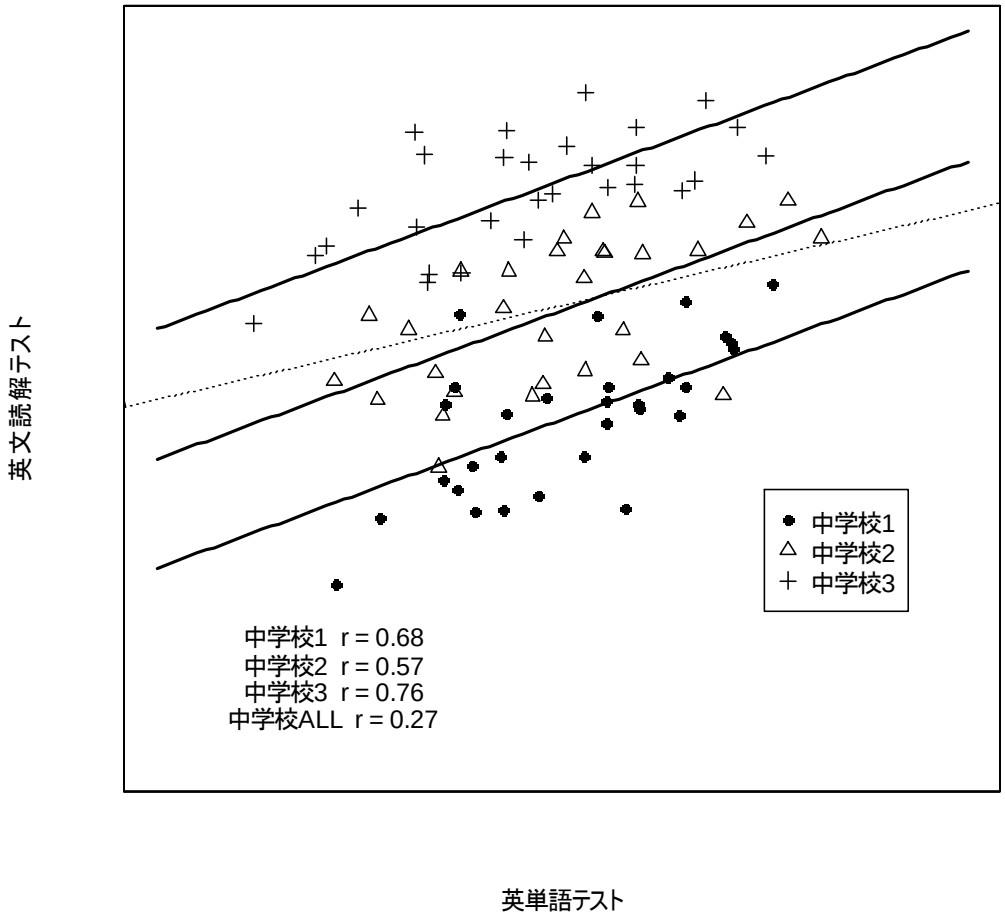


図 2.3.2　3 つの中学校の「英単語テスト」と「英文読解テスト」の散布図
（ランダム切片モデル）

2.3.3　ランダム係数モデル

前節では，(2.3.2)式のように切片 β_{0j} について，β_0 と u_{0j} への分解を行ったが，同様に係数部分についても，

$$\boldsymbol{\beta_{1j} = \beta_1 + u_{1j}}, \qquad (2.3.4)$$

として固定部とランダム部への分解を考えることができ，これをランダム係数モデルと呼ぶ．ただし，

通常は係数部分のみの分解ではなく，ランダム切片モデルも組み合わせて，以下のモデルを組み立てることが多い.

$$\text{レベル}1：\boldsymbol{y_{ij}} = \boldsymbol{\beta_{0j}} + \boldsymbol{\beta_1 x_{ij}} + \boldsymbol{e_{ij}}, \tag{2.3.5}$$

$$\text{レベル}2：\boldsymbol{\beta_{0j}} = \boldsymbol{\beta_0} + \boldsymbol{u_{0j}}, \tag{2.3.6}$$

$$\text{レベル}2：\boldsymbol{\beta_{1j}} = \boldsymbol{\beta_1} + \boldsymbol{u_{1j}}. \tag{2.3.7}$$

ここで各確率変数は,

$$e_{ij} \sim N(0, \sigma_e^2),\ \begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N(0, \Omega_u),\ \Omega_u = \begin{pmatrix} \sigma_{u_0}^2 & \\ \sigma_{u_{01}} & \sigma_{u_1}^2 \end{pmatrix},$$

に従うものとする．(2.3.5)式に，レベル 2 のモデルである(2.3.6)，(2.3.7)式を代入すると，

$$y_{ij} = \beta_0 + u_{0j} + \beta_1 x_{ij} + u_{1j} x_{ij} + e_{ij}, \tag{2.3.8}$$

のようになる.

2.3.4　null モデルと級内相関係数

説明変数を組み込まない以下のモデルを null モデルと呼ぶ.

$$\text{レベル}1：\boldsymbol{y_{ij}} = \boldsymbol{\beta_{0j}} + \boldsymbol{e_{ij}}, \tag{2.3.9}$$

$$\text{レベル}2：\boldsymbol{\beta_{0j}} = \boldsymbol{\beta_0} + \boldsymbol{u_{0j}}. \tag{2.3.10}$$

(2.3.9)式に(2.2.10)式を代入すると，

$$y_{ij} = \beta_0 + u_{0j} + e_{ij}, \tag{2.3.11}$$

が得られるが，この式の u_{0j} および e_{ij} の分散を用いて，以下の級内相関係数（intraclass correlation coefficient, ICC）が得られる.

$$ICC = \frac{\sigma_{u_{0j}}}{\sigma_{u_{0j}} + \sigma_{e_{ij}}}. \tag{2.3.12}$$

ICC は，全体の分散におけるレベル 2 の切片部の分散の割合を表す．ICC は 0 から 1 の数値になるが，1 に近づくほどレベル 2 の影響が大きいことになる．逆に ICC が 0 に近い場合には，レベル 2 の影響は小さいことになり，マルチレベル分析を行わなくてもよいことになる.

2.3.5　マルチレベル分析の可能性

（a）重回帰分析モデル

これまでは，説明変数１つで従属変数を説明する単回帰モデルを基にしたマルチレベル分析の例を示したが，説明変数が 2 つ以上の重回帰分析においても，同様のモデル構築を行うことができる．たとえば，説明変数を x_{1ij} と x_{2ij} の 2 つとすると以下のようなモデルが考えられる．

$$\text{レベル 1}：\boldsymbol{y_{ij} = \beta_{0j} + \beta_1 x_{1ij} + \beta_2 x_{2ij} + e_{ij}}, \quad (2.3.13)$$

$$\text{レベル 2}：\boldsymbol{\beta_{0j} = \beta_0 + u_{0j}}, \quad (2.3.14)$$

$$\text{レベル 2}：\boldsymbol{\beta_{1j} = \beta_1 + u_{1j}}, \quad (2.3.15)$$

$$\text{レベル 2}：\boldsymbol{\beta_{2j} = \beta_2 + u_{2j}}. \quad (2.3.16)$$

これは，レベル 1 の全ての切片，係数部にランダム部を割り当てており，フルモデルと呼ばれる．この他にも，x_{2ij} にはランダム部を割り当てずに(2.3.16)式を $\beta_{2j} = \beta_2$ とするなど，マルチレベル分析では，分析者が自由にモデルを構築することができる．

（b）モデルの比較

これまで述べたように，マルチレベル分析ではヌルモデルを含め，複数のモデルを構築することができる．通常，マルチレベル分析では最尤推定法により母数の推定を行うため，AIC や BIC といった適合度指標の比較，もしくは対数尤度を-2 倍したもの(-2 log-likelihood)を用いた尤度比検定などにより，モデルの比較・選択を行う．

（c）繰り返し測定データに対するモデル

レベル 1 について，これまでは各個人（先の例では中学生）としていたが，必ずしもその必要があるわけではない．たとえば，ある調査を同じ被験者に繰り返し実施した時は図 2.2.3 のようになるが，この場合はレベル 1 が各実施時期，レベル 2 が個人になる．

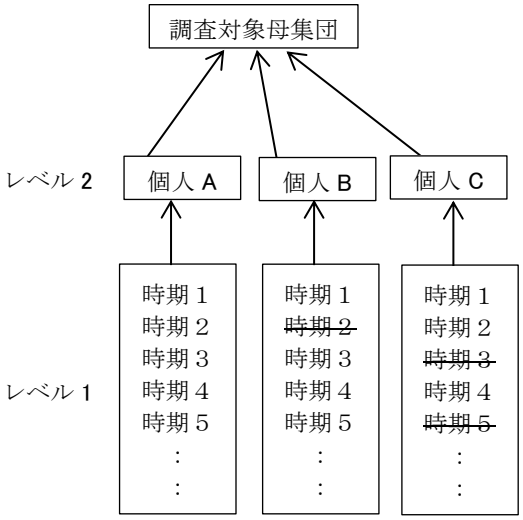


図 2.3.3　繰り返し測定デザイン

通常，このような繰り返し測定デザインの場合は，すべての被験者に対して，すべての実施時期で調査を行うことが望ましかった．しかしながら，マルチレベル分析では必ずしもその必要がないことが大きな特徴である（図 2.3.3 における個人Ｂの時期２や，個人Ｃの時期３，５など）．

（d）さらに発展的なモデル

本稿では扱わないが，マルチレベル分析ではこの他にも様々な発展的モデルが提案されている．カテゴリカルデータを扱うためのロジスティック回帰分析への応用，潜在変数を組み込んだマルチレベル因子分析モデルなど様々な応用例がある．

3. 名義反応モデルを利用した項目分析

全国的な学力調査などの大規模なテストにおいてパーフォーマンス・アセスメントが導入される場合，その採点はルーブリックなどにもとづく人間の主観的な判断が必要となる．その際，ルーブリックに示された規準間に順序がつけられない場合，その採点結果は名義尺度上にあると見なすことができる．IRT のうち名義反応モデルを利用すれば段階反応モデルと同様の形で項目分析などを置こうなうことができる．本研究調査においてはすべて事前に順序情報を与えた採点規準を用いたため，分析には名義尺度モデルを用いることはなかったが，今後の参考として，今回,リーディング・リテラシー問題によって得られた類型情報にこれを適用した例を示す．分析には EasyEstimation を利用した．

3.1 名義反応モデルの概要

平成２３年度文部科学省委託研究「全国規模の学力調査における重複テスト分冊法適用の試み」では，0 点，1 点，2 点のような段階をもつ多値型形式のデータについて，段階反応モデルを用いた分析が行なわれた．これらのデータは，それぞれの反応に対して順序性がつけられた（誤答＜正答，もしくは 0 点＜1 点＜2 点など）もの，いわゆる順序尺度データである．対して，順序性がない「名義尺度」データについて項目反応理論では Bock(1972, 1997）による名義反応モデル(nominal response model)を用いて，分析を行なうことができる．実際のテスト場面において，名義反応モデルは，

・多枝選択問題に対して，受験者が選択した選択枝情報をそのまま反応データとする場合，
・記述式問題などに対して，採点を順序尺度（0 点，1 点など）ではなく類型として採点したデータ，
・選択枝が順序性を持たない問題（例：「どの教科が好きですか？　１．国語，２．数学，3．英語」），

のような事例に活用することが考えられる．

名義反応モデルは，これまで扱ってきた項目反応理論の諸モデル（2 パラメタ・ロジスティック・モデルなど）と同様に，ある潜在特性値 θ を持つ受験者があるカテゴリに反応する確率を，項目特性関数で表現する．名義反応モデルでは，項目母数の数値自体から情報を得ることは難しく，項目反応カテゴリ特性曲線から，項目に関する情報を得ることになる（次節で具体的な分析結果を示す）．

当然のことながら，名義反応モデルにおいても項目反応理論の特徴（等化分析や，テスト情報量によるテストの精度表示など）を有しているが，先に挙げた項目反応カテゴリ特性曲線は GP 分析図に類似したものとなる．しかしながら，古典的テスト理論（正答数得点）の枠組みでは名義尺度上の項目は合計得点に組み入れることができないが，名義反応モデルではそのような項目をすべてモデル内で表現できることが最も大きな違いとなる．

3.2 名義反応モデルを利用した項目分析の実際

本節では，リーディング・リテラシーの問題について，名義反応モデルを利用した項目分析を行っ

た. まず, 多枝選択式問題の項目分析結果を示す. k01 の項目反応カテゴリ特性曲線をみると (図 3.1), リーディング・リテラシーの特性値が高くなるほど，正答である類型 4 を選択する確率が高く，その他の類型を選択する確率が低くなっている. それに対し, k02 の項目反応カテゴリ特性曲線は (図 3.2), リーディング・リテラシーの特性値が高いほど，誤答である類型 2 を選択する確率が高く，正答である類型 1 の選択確率は，低いままである．この結果より，k02 は，類型 2 が，判断を惑わせる，いわゆる「ひっかけ」の選択枝であったことが読み取れる．これは，本書の「4 章．数学と国語の信頼性」において，正答率及び点双列相関係数，GP 分析図から示唆された結果と一致するものである．

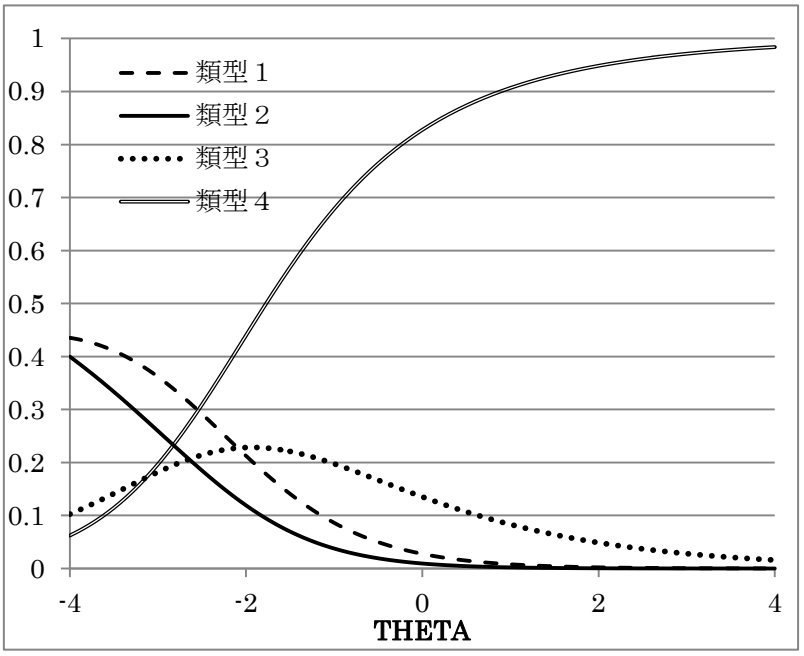


図 3.1 k01

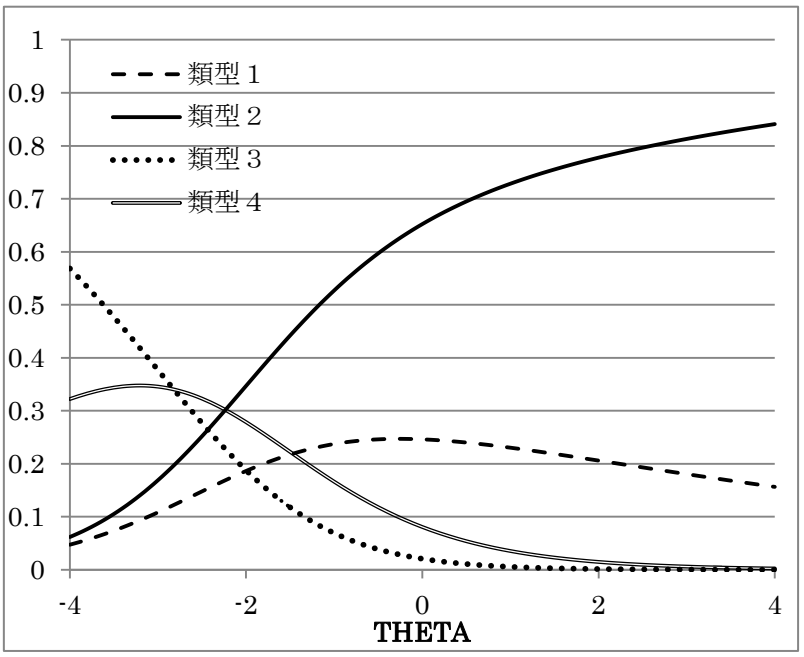


図 3.2 k02

続いて，k07 の項目反応カテゴリ特性曲線を見てみる．すると，k07 の正答は類型２であるにも関わらず，誤答である類型３のグラフが右上がりになっており，この類型が「ひっかけ」の選択枝であったと考えられる．これも，４章と一致する結果である．

本調査のリーディング・リテラシーに関する多枝選択式問題では，４つの選択枝のうち，正答が１つ含まれていた．受験者が選択した選択枝情報をそのまま反応データとすることによって，正答の類型だけでなく，複数の誤答の類型の特徴に関する情報も得られるということが，この名義反応モデルを用いた項目分析の大きな特徴の一つといえる．受験者の反応データを，二値データに変換してしまうと，項目パラメタの情報からしか情報を得ることができないため，識別力の低さについての情報は得られても，識別力が低かった理由については，本書の４章で行われているような複数の視点から考察していくことが必要になる．しかしながら，この名義反応モデルを用いた項目分析で描かれる，項目反応カテゴリ特性曲線をみることで，「ひっかけ」となる選択枝の存在を容易に発見できる．

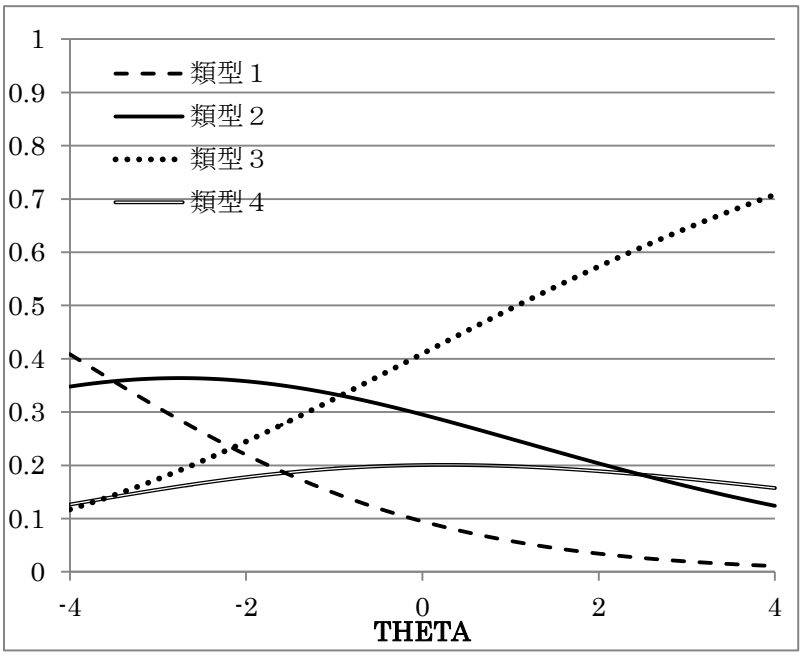


図 3.3 k07

続いて，記述式の問題の項目反応カテゴリ特性曲線を示す．k20 は，完全正答が類型１，部分正答が類型２であり，その他の類型が全て誤答である．図 3.4 をみると，部分正答の類型２は，途中まではきれいな項目反応カテゴリ特性曲線を描いているものの，途中から右下がりになっている．図 3.4 をみてわかるように，類型２が右下がりになり始める学力特性値と，類型１の傾きが大きくなる学力特性値がほぼ同じであることから，学力特性値が高いほど，部分正答（類型２）する確率は高くなるものの，一定のレベルよりも学力特性値が高くなると，部分正答（類型２）よりも，難度の高い完全正答（類型１）する確率が高くなることが示されている．

また，k20 の図 3.4 は，完全正答である類型１は右上がりになっているものの，途中で切れている．理論的に，学力特性値は平均 0，標準偏差１となるように推定されていることを考えると，本調査の受

験者集団に対し，類型１は難度が高い，すなわち，k20 の完全正答の基準が厳しすぎたことが推察される．さらに，類型３と類型５が類似した形状を示していることから，この２つの類型は，質的に類似した類型であったことが考察できる．

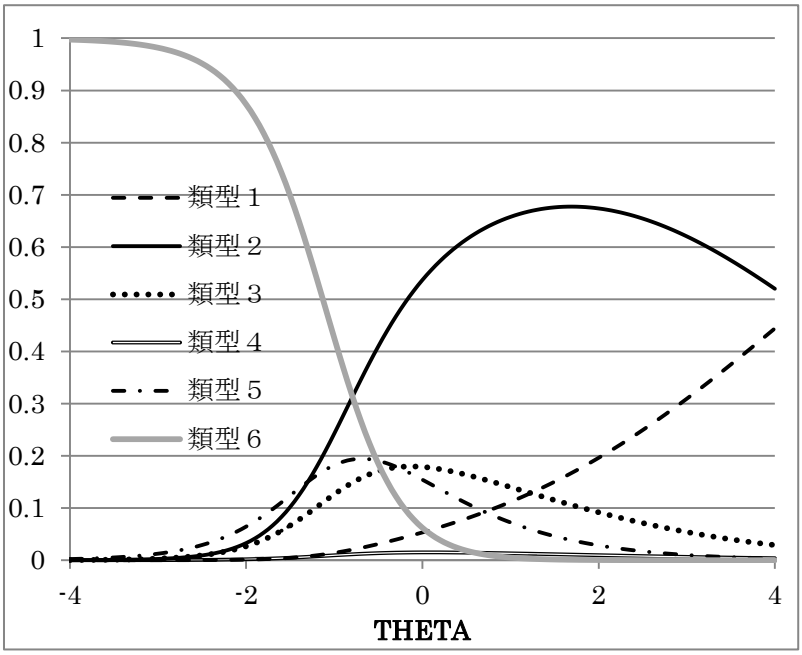


図 3.4 k20

k20 同様に，記述問題の例として，k24 を取り上げる．k24 の項目反応カテゴリ特性曲線（図 3.5）を概観すると，類型１と類型２，類型４と類型５が類似した形状を示しており，k24 の類型を，質的には「類型１と２」，「類型３」，「類型４と５」，「類型６」という大きく４つのまとまりに分類できると考えられる．実際に，k24 の６つの類型を上記した４分類に変換し，再度，項目反応カテゴリ特性曲線を描いたところ（図 3.6），「類型１・２」が二値型データを用いた項目特性曲線のような形状になり，図 3.5 と比べても，解釈のしやすいきれいな図となった．以上の結果を踏まえると，k24 の類型は，「類型１と２」，「類型３」，「類型４と５」と「類型６」の４つに分類できると判断することができる．

ここで，k24 の採点基準を確認してみると，順序性をもつ多値データへの変換ルールは，「類型１」が完全正答（２点），「類型２と３」が部分正答（１点），「類型４と５」が誤答（０点），類型６が無答であり，二値データへの変換ルールは，「類型１，２，３」が正答（１点），「類型４，５，６」が誤答（０点）となっている（k24 の類型の変換ルールは表 3.1 にまとめてある）．すなわち，k24 の名義反応を用いた項目分析の結果は，予め作成されていた採点基準及びデータの変換ルールと，齟齬の生じるものであった．よって，k24 の採点基準の妥当性を見直す必要があるといえる．このように，記述問題の場合には，名義反応を用いた項目反応カテゴリ特性曲線を描くことで，採点基準の適切さに関する情報を得ることもできる．

表 3.1　k24 の採点基準で想定された類型の分類と，項目分析から考察される類型の分類の比較

	採点基準		項目分析の結果からの考察
	順序性のある多値データ	二値データ	
類型１	完全正答（２点）	正答（１点）	グループ１
類型２	正答（１点）		
類型３			グループ２
類型４	誤答（０点）	誤答（０点）	グループ３
類型５			
類型６	無答（０点）		グループ４

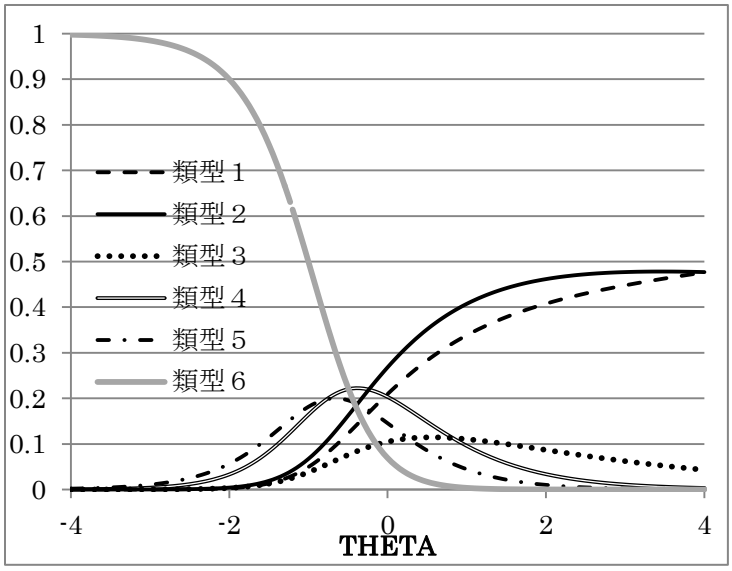


図 3.5. k24

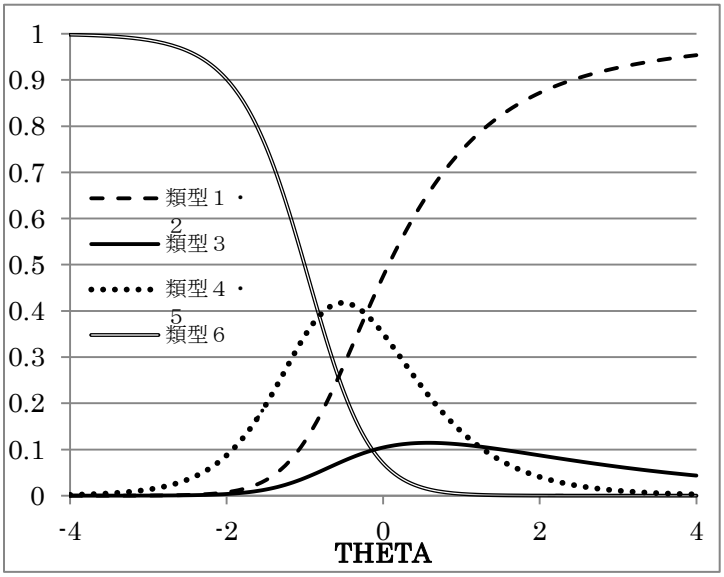


図 3.6 k24（採点基準修正版）