資料４

令和３年度『全国学力・学習状況調査』

経年変化分析調査

テクニカルレポート

文部科学省
MEXT
MINISTRY OF EDUCATION,
CULTURE, SPORTS,
SCIENCE AND TECHNOLOGY-JAPAN

文部科学省

総合教育政策局調査企画課学力調査室

目次

表目次

図目次

1　経年変化分析調査-序

「全国学力・学習状況調査」は、毎年悉皆で実施する調査（本体調査）及び３年に１度程度実施する「経年変化分析調査」と「保護者に対する調査」（補完調査）で構成されている。今後の全国学力・学習状況調査の方向性として、令和３年３月の「全国的な学力調査に関する専門家会議」において、①毎年、原則として悉皆で実施している調査と、②それを補完する調査である「経年変化分析調査」及び「保護者に対する調査」を国が実施すべき主要な調査の「二本柱」として位置付け、整理することが提言され、文部科学省では、この方向性を踏まえて、各調査の充実を図っていくこととしている。このことを受け、調査の透明性とデータの信頼性および今後の CBT 化等の技術的な発展性の担保のため、このテクニカルレポートでは経年変化分析調査で使用されている測定技術の詳細を解説・報告する。

経年変化分析調査は、悉皆による個々の児童生徒の学力状況の把握ではなく、全国的な学力の状況について、経年の変化を精緻に把握・分析し、国の教育施策の検証・改善に役立てることを目的としている。そのため、経年変化分析調査における直接の測定対象は全国の児童生徒母集団から抽出された標本であり、そこで使われている測定技術は、学力の測定モデル（心理計量モデル：psychometric model）としては項目反応理論（item response theory：IRT）を、調査デザインとしては標本調査法（sample survey method）を組み合わせた重複テスト分冊法（item-matrix sampling/matrix sampling）を基本としている。

PISA や全米学力調査（NAEP）等でも採用されている重複テスト分冊法の基本的な技術獲得は、平成 21 年度から平成 24 年度にかけて文部科学省委託調査研究として行われた。その後、平成 25 年度を初回として、平成 28 年度と令和３年度に重複テスト分冊法を全面的に採用した実施が行われ、個々の児童生徒の学力状況ではなく、国全体の学力分布の状況について、平成 28 年度調査を基準にした学力の経年変化を精緻に把握・分析した結果が報告されている。

これらの調査では、学力の測定モデルに IRT を採用していることを活かして、分冊(booklet)と呼ばれる問題構成が互いに異なる複数のテスト冊子を準備し、それらを別々の受検者集団に実施している。これにより、受検者の解答時間などの負担を軽減しながらも、以下のようなメリットを享受することが可能となっている。すなわち、

A)　多数のテスト問題(項目: item)を一度の調査で実施することで、従来の調査方法と比べ、測定内容の観点からより多面的かつ妥当な方式で学力分布の状況を調べられること、
B)　互いに異なる項目の組み合わせからなる分冊であっても、それぞれの分冊から得られた学力を共通の尺度上に布置して比較できること、
C)　過年度に実施した項目を一部含んだ分冊を複数準備することにより、高い精度で学力の経年変化の追跡ができること、

などである。さらに、標本調査法と組み合わせることにより、同一予算規模であっても、国全体の学力分布を高い精度で推定でき、教育施策を立案・遂行する上で必須の情報を系統的に得ることができる。

また、このレポートで報告された技術詳細は、将来的には、保護者調査や質問紙調査などの調査項目を分冊中に含めることによって、学力データと同時にそれらの情報をとることによる発展性や CBT 化などによるさらなる調査の効率化などを試みる際の基盤となることも期待される。そのため、今後の技術的な発展を予想して、現時点では調査報告等に使われていない推算値(plausible values：PV's)や項目反応理論モデル（IRT モデル）の一つである段階反応モデル（graded response model）などの理論的な解説もこのテクニカルレポートには含まれている。さらに将来的には IRT モデルを利用した多段階適応型テスト（computerized multistage testing：個々の受検者の能力や回答状況に合わせて出題する項目を逐次的に変えるテスト形式）なども組み込まれていくことも予想される。このテクニカルレポートはそのような発展のための最初のステップでもある。

2　テストの設計と開発

2.1　テスト理論

良いテストが持つべき特徴を整理すると、妥当性（validity）、信頼性（reliability）、有用性（usability）の３つが指摘できる。妥当性とは、テストによって得られる解釈が適切であるかどうかに関する特徴である。たとえば知能検査によってパーソナリティは測れない。このような場合に「パーソナリティ・テストとしてはこの知能検査には妥当性がない」という。学力検査と関心・意欲・態度のような人格評価とを混在させてはならないという評価の原則もこの特徴に由来する。また、信頼性とは、そのテストで測定されている結果がどの程度一貫している（consistent）かどうかに関する特徴である。練習効果など現実的な条件を捨象して、概念的抽象的に考え、もし同じ受検者が同じテストを受けても、受けるたびにその結果が異なってしまえばそのテストには信頼性がないことになる。一方、有用性とはそのテストのいわば使い勝手のよさや実用性を意味している。たとえどれだけ精度が高く理念的に優れたテストであっても、実施に莫大な費用や時間がかかるのでは、費用対効果（C/P 比）から判断してそのテストには有用性はない。

テスト理論（test theory）とは、一言でいえば、このような特徴をあわせもった高品質の心理検査を作成し、それを実施・運用するための基礎を与えるためのものである。すなわちあるテストが、

- そのテストで測定したい学力や性格などの心理学的特性を（妥当性の問題）、
- 本当に精度よく（信頼性の問題）、
- 現実的なコストのもとで（有用性の問題）、

測定できているかどうかを統計的に判断していく理論的基礎を与えるものである。

テスト理論には古典的テスト理論（classical test theory：CTT）と項目反応理論（IRT）の大きなふたつの理論体系がある。項目反応理論の方が比較的新しく提案されたモデルであるため、それ以前から存在していた理論に「古典的」という言葉を冠しているだけであって、古典的テスト理論が「もはや古びて役に立たない」理論であるという意味ではない。現在でもさまざまなテストの妥当性・信頼性・有用性を保証する技術体系として重要な役割をはたしている。しかし、この経年変化分析調査では将来の全国学力・学習状況調査の CBT 化を見据え、かつテスト実施運用の際の柔軟性から測定の技術的基礎として IRT を採用した。

2.2　学力分布と得点分布との区別

学力の調査において出題されるテスト問題（項目: item、このテクニカルレポートでは「試験」と「テスト」は基本的に同じ意味を指すが、適宜使い分ける）が年度間で異なれば学力の経年比

較はできない。観察されるテスト得点の変化が、受検者の学力の変化によるものなのか、項目の難易度(困難度)の変化によるものかの区別ができないためである(＝テスト得点の項目依存性)。通常のテストで使われているテスト得点にはこの両者が分かちがたく反映されている。そのため、極端な場合には、易しいテストを受けた学力の低い受検者と難しいテストを受けた学力の高い受検者のテスト得点が実際の学力の高低と逆転してしまう現象が起こる。その対処として、毎年同じ項目を繰り返し使うなどの方策が採られる。しかし、長期間にわたる学力調査では、例えば学習指導要領の改訂等の理由で過年度に実施した項目が使えなくなる、項目が公開できない、項目の場面設定が年月とともに古くなればやはり問題が使えなくなる、などの問題点が生じる。

こうした限界を克服する有効な手段として、項目反応理論モデル（IRT モデル）がある。IRT モデルは、従来の単なる加算方式によって求められるテスト得点とは違い、受検者の項目への正誤反応パターンにもとづいて、受検者の学力そのものを、実施した項目の困難度から明確に分離して扱うことができる。その原理を簡単に見ておこう。次の問題は現在の大学入学共通テストのいわば先駆けでもある共通第 1 次学力試験のある年の数学試験で出された 24 問のうちのある項目である。受検者数は約 30 万名であり、配点に重みを設けずに正答なら 1 点、誤答なら 0 点とする（このような得点のことを項目得点（項目反応）と呼ぶ）と、そのテスト得点（正答数得点）は 0 点から 24 点の範囲に入る。テスト得点ごとにこの項目に正答した受検者の割合を計算する。このテストで 0 点であった者の中にはこの項目が正答できたものは存在しないからその割合は

問題　半径√5/2 の円に内接する二等辺三角形 ABC において，AB=AC=2 とする．また，A を通るこの円の直径を AD とする．このとき，sin∠BAC=キ／クである．

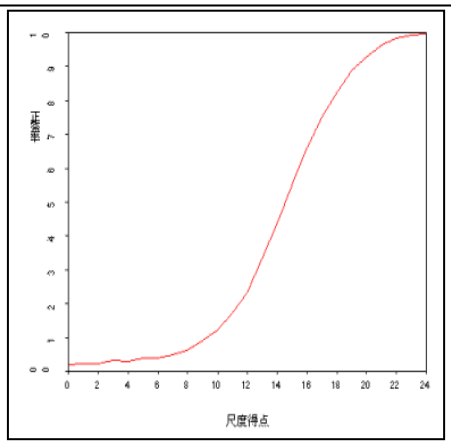


図 1：項目特性曲線の例（良い問題）

当然 0 となる。逆に満点の 24 点を取った者は全員この項目に正答しているからその割合は 1 となる。さらに、横軸にテスト得点、縦軸にテスト得点ごとの正答率をプロットし、それらを結べば図 1 の様な曲線が得られる。この曲線のことを項目特性曲線(item characteristic curve : ICC)と呼ぶ。当然ながら得点が高くなるほど正答率も通常高くなる。

次の項目も、共通第 1 次学力試験の国語で出題された項目である。この項目の項目特性曲線を描いたものが図 2 である。テスト得点が高い層よりも低い層や中間の層の方で若干正答率が高くなっていることが読み取れる。すなわちこの項目は調べたい国語の学力をうまく識別できていないという意味で悪い項目ということになる。

問題　傍線部「夙に」の意味として最も適当なものを，次の①～⑤のうちから一つ選べ.

①うまれつき　　②朝早くから　　③心のそこから

④ずっと以前から　　⑤十二分に

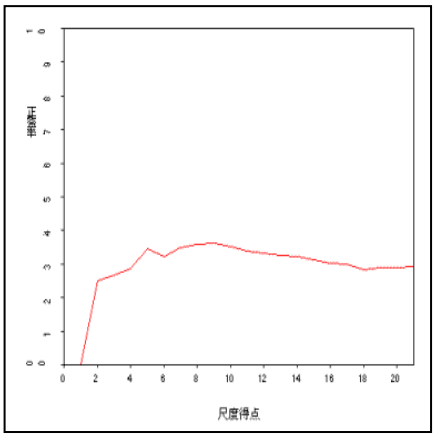


図 2：項目特性曲線の例（悪い問題）

ここまでの議論ではテスト得点と学力を同一視してきたが、IRT モデルにおいてはこの 2 つを明確に区別して扱う。すなわち、直接観測できない量としての学力を反映する潜在特性（または、受検者母数）θ を導入し、ある θ のもとである項目に正答する確率を表すことを考える。例えば、2 母数ロジスティックモデル（two-parameter logistic model）と呼ばれるものは、次式で記述される。

$$P(X_j = 1|\theta) = \frac{\exp\{Da_j(\theta - b_j)\}}{1 + \exp\{Da_j(\theta - b_j)\}} \tag{1}$$

この数式を使えば、項目の母数（識別力a_j、 困難度b_j）の組み合わせによって、図 3 に示すように様々な項目特性曲線をもつ項目を θ の軸上で扱えるようになる。点線で表された項目は実線で表された項目よりも難しく（困難度が高く）、実線の左側の曲線で表された項目は実線の項目より易しく（困難度が低く）、さらに実線を横切るような曲線で表された項目は実線の項目よりも学力の識別ができていない(識別力が低い)ことを表している。このように項目の困難度や識別力を IRT モデル中の母数として表現し項目特性曲線を記述することで、受検者の学力分布と

項目の困難度の高低とを分離し区別をしながら、ある特定のテスト冊子や受検者集団に依存することなく学力 θ を測定することが可能となるのである。

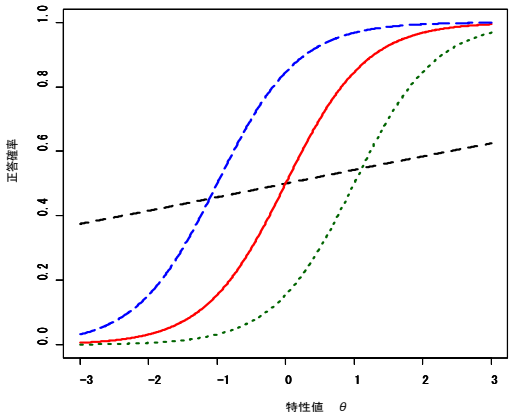


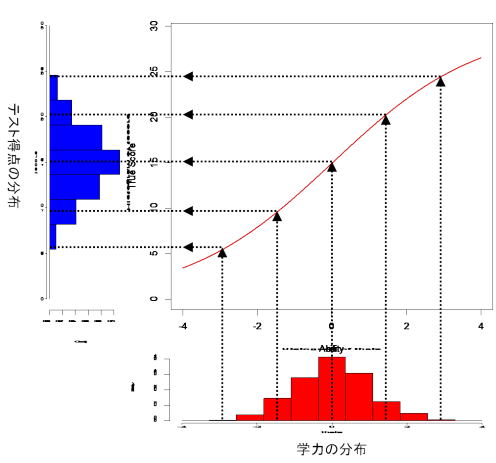
図 3：様々な項目特性曲線

受検者の学力分布と項目の困難度の高低との分離という点は、IRT の導入により、混同されがちなテストの得点分布と測定すべき学力分布とを分けて扱うことが可能となることを意味する。これを端的に示したのが次ページの図 4a〜4d である。図 4a が全体的に測定精度の高い（＝信頼性の高い）テストの場合に学力分布がテスト特性曲線と呼ばれるものを通してテストの得点分布に変換されるプロセスを示している。図 4b がそれとは逆に、テストが粗雑に作られた結果、測定精度が低い場合の様子を示している。どちらの図においても真に知りたい学力分布はまったく同じであるが、テストの性質によって手にできるスコアの分布が異なっていることがわかる。前者にくらべ後者は分布の広がり方がかなり縮まっており、児童生徒の学力の個人差の識別ができていない、したがって精度の低い（＝信頼性の低い）テストということになる。

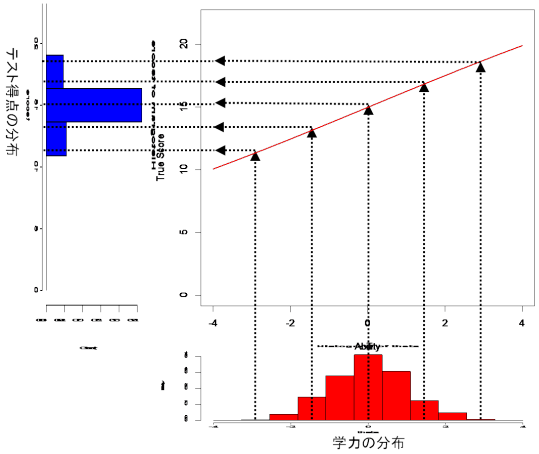
さらに、例えば資格試験や検定試験などのようにある水準以上の学力水準・能力水準であるか否かを判断する場合のテストにおいては、理想的には図 4c のような双峰分布（ふた山分布）にテスト得点の分布がなるようなテストを準備する必要がある。真に知りたい能力分布は図 4a、図 4b と同じであるにも関わらず、テストの作りによっては、テスト得点の分布がこのように変化する。いわゆる「学力のふた山化」現象などの議論が、テスト得点の分布をみての議論なのか、学力分布の形を推定しての議論なのかを明確に区別しておかないと判断を間違う可能性をこの図は示しているのである。

最後の図 4d は図 4a にくらべて全体的に難しい問題からなるテストを作った場合のテスト得

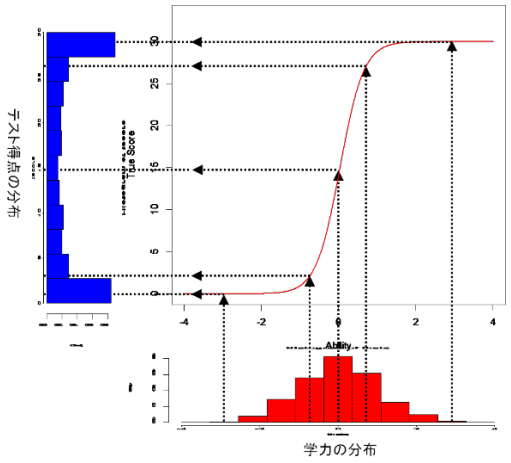
点の分布の様子を表したものである。能力の高い受検者を見いだす必要のある選抜試験などでとられるテストの作成方法である。能力分布がここでも同じであるにも関わらず、テスト得点の分布は得点の低い方に山ができていることがわかる。



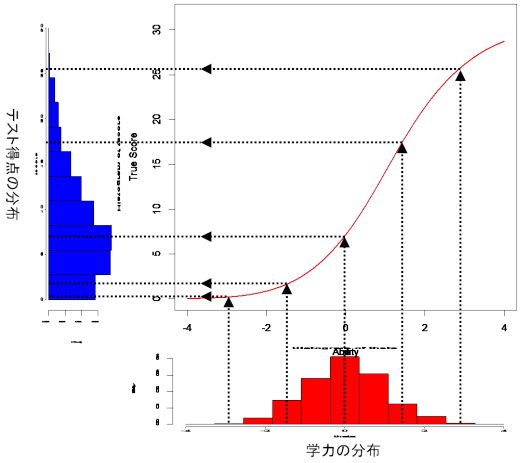
a.信頼性が高いテスト



b.信頼性が低いテスト



c.平均的な難しさの問題を多くした場合



d.難しい問題を多くした場合

図 4：学力分布とテスト得点の分布の関係

また、学力分布を表現する尺度（スケール）が構成できれば、次頁の図のように、実際に出題された問題の困難度をその尺度上で解釈し、識別力と合わせて、測定したい学力に対してその問題がどのように機能するかを、問題内容の質的妥当性とともに検討することも可能になる。

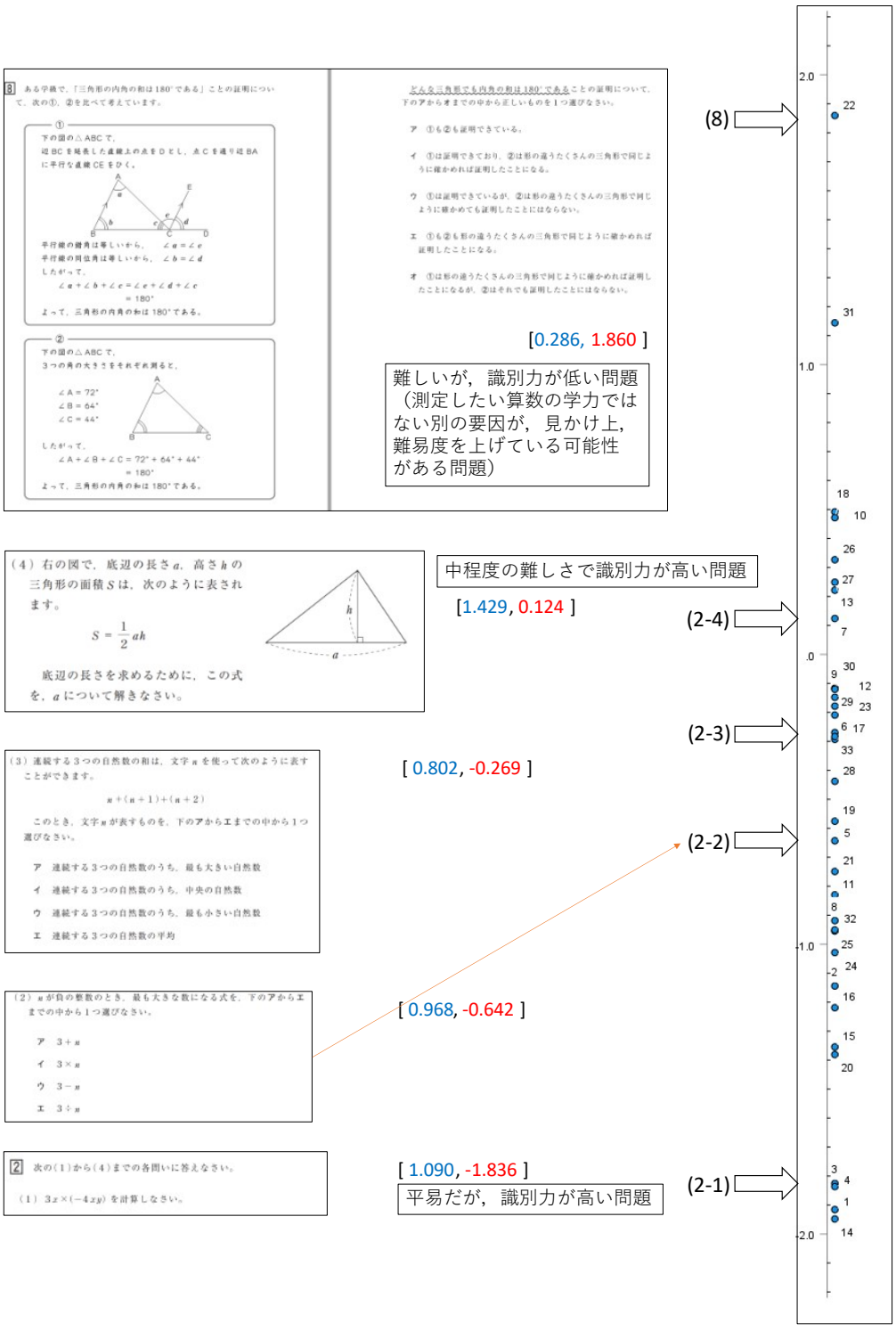


図 5：問題と識別力・困難度と尺度との関係

注：平成 21 年度全国学力・学習状況調査／中学・数学のデータより：[識別力、困難度]

2.3　学力測定のための心理計量モデル

2.3.1 項目反応理論（IRT）

テスト得点の項目依存性と項目困難度の標本依存性は、長い間、古典的テスト理論（CTT）の限界と認識されてきた問題であった。例えば、100 点満点のテストにおける同じ 70 点という得点であっても、難しい項目から構成されるテストの 70 点と易しい項目から構成されるテストの 70 点では実際の学力レベルが異なる。あるいは、正答率が同じ 20%の項目であっても、高い学力レベルの受検者集団における正答率 20%と低い学力レベルの受検者集団における 20%では実際の項目の難しさは異なる。このように、テスト得点で表される学力レベルはテストを構成する項目に依存し、正答率で表される項目困難度はテストを受検した受検者集団(標本)に依存する。

項目反応理論（IRT）は、CTT が抱えてきたそれらの問題に対し、項目反応モデル（以下、IRT モデル）という一つの解決策を与えた。IRT モデルは、一人の受検者が一つの項目に応答する際の正答確率をパラメトリックな関数としてモデル化したものである。IRT モデルには、受検者の学力を表す受検者母数と難しさなどの項目の特性を表す項目母数が含まれている。テストの結果から、項目に依存しない受検者母数と受検者集団に依存しない項目母数を推定することができる。このような IRT モデルの性質により、異なるテストを受検した受検者どうしの学力を同一尺度上で比較することが可能となる。その際、どんなテスト間でも学力を比較できるわけではなく、経年変化分析調査のように、事前に使用目的に沿うようにテストの構成を綿密に設計しておく必要がある。

すでに欧米では、IRT は学力調査をはじめ資格試験や適性試験などに広く利用されている。近年、日本でも IRT によるテスト運用が注目されつつあり、IRT によるテスト運用を利点の一つにあげる試験も多くなってきた。IRT によって運用されているテストの例として、NAEP（全米学力調査：National Assessment of Educational Progress)、LSAT（Law School Admission Test)、PISA（Programme for International Student Assessment)、TOEFL（Test of English as a Foreign Language)、情報処理技術者試験、医療系大学間共用試験医学系 CBT などがあげられる。

2.3.2　2 母数ロジスティックモデル

経年変化分析調査では、IRT モデルの代表格であり数理的にも取り扱いやすい 2 母数ロジスティックモデルを分析に用いる。2 母数ロジスティックモデルにおいて、θ をもつ受検者が項目 j に正答する確率 $P_j(\theta)$ は、

$$P_j(\theta) = \frac{\exp\left[Da_j\left(\theta - b_j\right)\right]}{1 + \exp\left[Da_j\left(\theta - b_j\right)\right]} \tag{2}$$

と表される。ここで、θは潜在特性値（latent trait)、能力母数（ability parameter）あるいは受検者母数などと呼ばれ、テストで測定しようとしている、学力などの何らかの構成概念（心理学的構成概念；psychological construct）に対応する。経年変化分析調査の場合、学力テスト（achievement test）を扱うことから、θを受検者母数と呼ぶこととし、θの具体的な推定値（estimate）を尺度値あるいはスコア（scale score）と呼ぶこととする。また、経年変化分析調査において測定しようとしている構成概念は各教科の学力である。θの定義域は$-\infty < \theta < \infty$であり、その値が大きいほどテストで測定しようとしている受検者の学力が高いことを示す。b_jは項目jの項目困難度母数（item difficulty）と呼ばれ、項目jの難しさを表す。b_jの定義域は$-\infty < b_j < \infty$であり、その値が大きいほど項目が難しいことを示す。a_jは項目jの項目識別力母数（item discrimination power）と呼ばれ、 $\theta = b_j$付近における受検者の尺度値の違いがどのくらい敏感に正答確率の違いに反映するかを表す。通常は$a_j > 0$が仮定され、その値が大きいほど$\theta = b_j$付近の尺度値をもつ受検者の個人差を明確に識別できることを示す。Dは尺度因子（scale factor）と呼ばれる定数であり、一般には(2)式で表わされるロジスティック曲線を正規累積曲線（normal ogive curve）に近似させるためにD=1.7（より正確には、D=1.702）が用いられる。

(2)式において、受検者母数θを横軸に、項目jへの正答確率$P_j(\theta)$を縦軸にとったグラフで表される曲線を項目特性曲線（item characteristic curve: ICC）と呼ぶ。図 6 を見ると、2 母数ロジスティックモデルの ICC の特徴を知ることができる。まず、$\theta = b_j$のとき正答確率がちょうど 0.5 になることがわかる。これは、(2)式の右辺に$\theta = b_j$を代入したとき$P_j(\theta) = 0.5$となることに対応する。また、項目 1 と項目 2 を比較すると、a_jの値を一定にしたままb_jの値を変化させると ICC が平行移動することがわかる。さらに、項目 1 と項目 3 を比較すると、a_jの値が大きいほうが$\theta = b_j$付近の ICC の傾きが急峻になることがわかる。

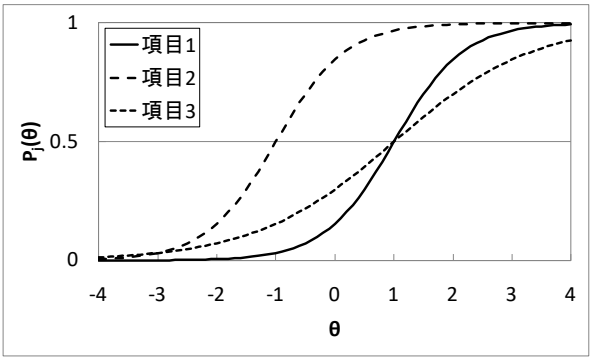


図 6：2 母数ロジスティックモデルの ICC の例

（注）項目 1 では a_1=1、 b_1=1、 項目 2 では a_2=1、 b_2=-1、 項目 3 では a_3=0.5、 b_3=1 である。

IRT モデルは、いくつかの前提条件の上にモデルが構成されている。特に重要な前提条件は、局所独立（local independence）の仮定と 1 次元性（uni-dimensionality）の仮定である。局所独

立の仮定とは、一つのテストにおいて、ある尺度値θをもつ受検者がある項目に正答する確率は他の項目に正答する確率の影響を受けないという仮定である。確率論的には、ある受検者が各項目に正答するのは互いに独立な事象であるということを意味する。1 次元性の仮定とは、一つのテストを構成する項目はただ一つの構成概念を測定するものでなければならないという仮定である。なお、2 母数ロジスティックモデルのように、受検者母数が 1 次元（1 変量）の IRT モデルでは、局所独立の仮定と 1 次元性の仮定とは同値である（Lord & Novick 1968）。実際には、モデルを適用する際に、テストの 1 次元性だけを何らかの方法でチェックすることが多い。

テストの 1 次元性は、様々な方法で確認することができる（Hattie 1985）。よく用いられる方法として、各項目間の四分相関係数行列（正答・誤答などの 2 値データどうしの相関係数行列）における固有値（主成分[=各能力次元]の分散を反映する量）のスクリープロットから判断する方法がある。例えば、テスト結果から図 7 のようなスクリープロットが得られたとする。図を見ると、第 1 固有値が突出しているとともに第 2 以下の固有値と比較して格段の差を生じていることがわかる。このような傾向が確認できるとき、単一の能力次元からテストの正答・誤答が説明できると判断され、当該テストにおいて 1 次元性の仮定は満たされていると判断できる。これ以外の客観的な基準としては、各項目間の四分相関係数行列の第 1 固有値の分散説明率が 20%以上あることが Reckase(1979)により推奨されている。

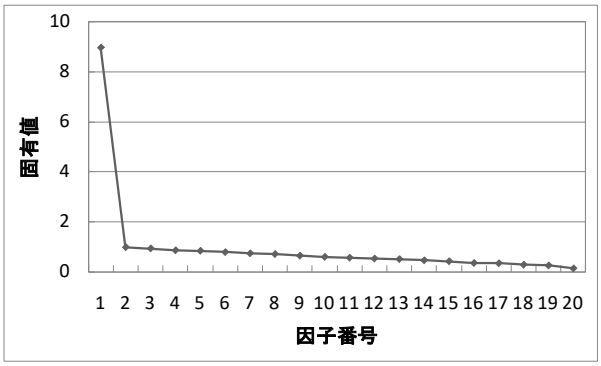


図 7：スクリープロットの例

（注）図中「因子番号」は固有値の大きさの順位のこと

2.3.3 項目情報関数とテスト情報関数

一般に、学力を測定する際に一つのテストの中で実施できる項目の数は限られている。またそのことを反映して、テストの項目は実施回や冊子によって変わるので、テストの結果得られたスコアがどの程度安定し信頼できるのかを知るために、テストの測定精度を評価することは重要である。CTT では、信頼性係数（reliability coefficient）を用いて受検者集団全体についての平均的・総合的な意味での測定精度を見積もることはできる。それに対し IRT では、IRT モデル

を導入することで、尺度値 θ の水準に応じた項目ごとの測定精度を表す指標を定義できる。その指標を項目情報関数（item information function）と呼び、

$$I_j(\theta) = \frac{P_j'^2(\theta)}{P_j(\theta)Q_j(\theta)} \tag{3}$$

と表すことができる。ここで、$P_j(\theta)$は尺度値 θ をもつ受検者が項目jに正答する確率である。$Q_j(\theta) = 1 - P_j(\theta)$は誤答確率である。また、$P_j'(\theta)$は θ についての$P_j(\theta)$の 1 次導関数（1 階微分）である。2 母数ロジスティックモデルを利用する場合は、

$$I_j(\theta) = D^2 a_j^2 P_j(\theta)Q_j(\theta) \tag{4}$$

で計算できる。

(3)式や(4)式で表される項目情報量は加算可能である。つまり、テストを構成する各項目の項目情報量を全ての項目に亘って加算すると、そのテスト全体の情報量となる。このような項目情報関数の単純和をテスト情報関数（test information function）と呼び、

$$I(\theta) = \sum_{j=1}^{n} I_j(\theta) = \sum_{j=1}^{n} \frac{P_j'^2(\theta)}{P_j(\theta)Q_j(\theta)} \tag{5}$$

と表すことができる。もちろん、2 母数ロジスティックモデルを利用した場合は(4)式の単純和となる。

図 8 に、前節の図 6 で利用した 3 項目の項目情報関数を示す。図 9 に、その 3 項目からなるテストを作成した場合のテスト情報関数を示す。図 8 を見ると、項目情報関数は尺度値と項目困難度が等しい$\theta = b_j$で最大値をとることや、項目識別力a_jが大きいほど鋭いピークをもつことがわかる。図 9 を見ると、テスト情報量はテストを構成する各項目の項目情報量の単純和になっている様子がうかがえる。

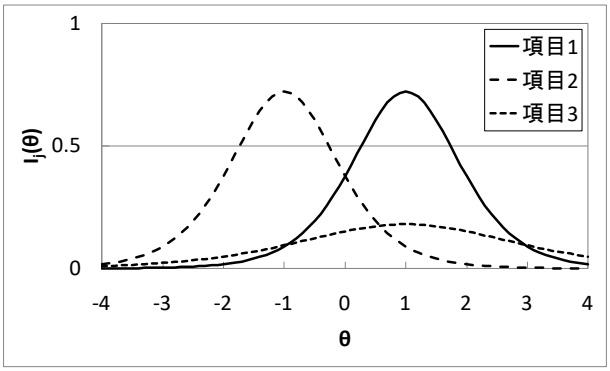


図 8：項目情報曲線の例

（注）項目 1 では a_1=1、 b_1=1、 項目 2 では a_2=1、 b_2=-1、 項目 3 では a_3=0.5、 b_3=1 である

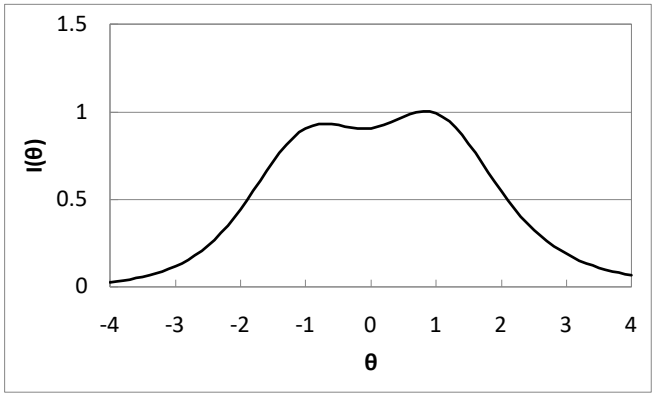


図 9：テスト情報曲線の例

(5)式のテスト情報量は、一般的な統計学の分野における Fisher 情報量と同一のものであり、テスト全体の測定精度と密接な関係がある。テスト情報量を用いると、次節で概説する θ の最尤推定値の標準誤差（ばらつき）を $1/\sqrt{I(\theta)}$ で見積もることができる（図 10)。すなわち、テスト情報関数をみれば、テストがどの付近の尺度値をどのくらい正確に測定できるのかが具体的にわかる。テスト情報量の大きい尺度値レベルが尺度値をより正確に測定できる部分であり、テスト情報量の小さい尺度値レベルが尺度値の測定精度が低くなる部分である。例えば、図 10 を見ると、当該テストはその受検者集団において平均的な尺度値レベルをもつ受検者の尺度値を他の尺度値より正確に測定できることが読み取れる。このような尺度値ごとの測定精度は、CTT では得られない情報であり、CTT に対する IRT の一般的な利点の一つになっている。

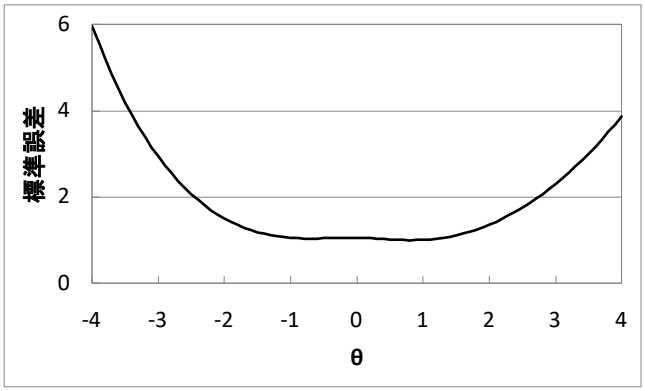


図 10：標準誤差の例

（注）テスト項目は、図 8 および図 9 と同じ

2.3.4 IRT における母数の推定方法

多肢選択式のテストなどでは、テスト結果を 2 値データ（1: 正答、0: 誤答）で表現できる。例えば、10 人の受検者が 5 項目からなるテストを受検したとすると、図 11 に示すような 10 行 5 列の行列が得られる。i行j列の要素は、受検者iが項目jに正答したか誤答したかを表している。このような 2 値の行列データは、項目反応データあるいは項目反応パターン（item response pattern）などと呼ばれる。経年変化分析調査では、n項目からなるテストをm人の受検者が受検したときの項目反応データを$X=\{x_{ij}: i=1,2,\cdots,m; j=1,2,\cdots,n\}$と記述することにする。

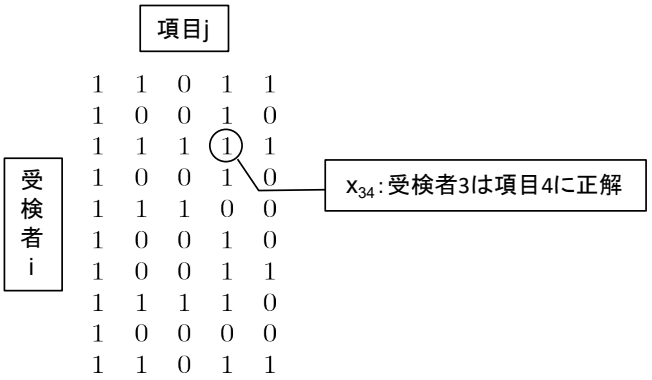


図 11：項目反応データの例

テストを IRT により運用するには、テストの結果として得られた項目反応データから IRT モデルに含まれる母数を推定する必要がある。その際、項目母数と受検者母数は別々に 2 段階で推定するのが現在の主流である。すなわち、項目母数を周辺最尤推定法（marginal maximum likelihood estimation: MMLE）によって推定したあと、得られた項目母数の推定値を利用して受検者母数を最尤推定法（maximum likelihood estimation: MLE）によって推定するという手順を踏む。IRT モデルの母数の推定は非常に煩雑であるため、母数を推定するための専用ソフトウェアは、BILOG-MG（Zimowski, Muraki, Mislevy, & Bock 2003）や EasyEstimation（熊谷 2009）をはじめとして、いくつか開発されている。

2.3.5 項目母数の推定

項目母数の推定には、ベイズ統計を応用した周辺最大事後推定法（marginal maximum a posteriori estimation: MMAPE）が利用されることが多い。その際、項目母数の推定値は MAP（maximum a posteriori）推定値として求められる。さらに、項目困難度母数と項目識別力母数に事前分布を設定しない場合の項目母数の推定法は、MMLE とみなすことができる。本節では、テスト結果の項目反応データから、項目母数を周辺最尤推定する方法について概説する。導出過程等の詳細については Baker and Kim（2004）や豊田(2005)が参考になる。

いま、m人の受検者がn項目からなるテストを受検し、図 11 に示すような項目反応データ$X=\{x_{ij}: i=1,2,\cdots,m; j=1,2,\cdots,n\}$が得られたとする。また、2 母数ロジスティックモデルなどの IRT モデルにより、尺度値θ_iをもつ受検者iが項目jに正答する確率が$P_j(\theta_i)$で与えられているとする。このとき、項目反応データXが得られる確率は、

$$P(X|\theta,\omega)=\prod_{i=1}^{m}\prod_{j=1}^{n}P_j(\theta_i)^{x_{ij}}Q_j(\theta_i)^{1-x_{ij}} \tag{6}$$

と表せる。ここで、$\theta=(\theta_1,\theta_2,\cdots,\theta_m)$は受検者母数ベクトルである。$\omega=(\omega_1,\omega_2,\cdots,\omega_n)$は項目母数ベクトルであり、2 母数ロジスティックモデルを利用する場合には$\omega_j=(a_j,b_j)$である。また、 $Q_j(\theta_i)=1-P_j(\theta_i)$は誤答確率である。Πは積を表す記号であり、前述した局所独立の仮定を反映している。

(6)式をθとωの同時尤度関数とみなし、その同時尤度関数を最大にするような(θ,ω)を最尤推定する方法は、同時最尤推定法（joint maximum likelihood estimation: JMLE）と呼ばれる。統計的推測における母数の推定値として、データのサンプルサイズが増えるにつれて推定値の誤差分散が小さくなり、推定値が真の値に近づいていくような一致推定量（consistent estimator)が求められるのが通常である。Neyman and Scott (1948) は、構造母数（structural parameter；IRT においては項目母数）が付随母数（incidental parameter：受検者母数）と同時に推定されるときには、構造母数（項目母数）は一致推定量にならないことを示した。このように JMLE には受検者の数を増やしても項目母数が一致推定量にならないという好ましくない性質がある（Andersen 1972）ため、現在では JMLE はほとんど用いられていない。

Bock and Lieberman (1970)の周辺最尤推定法（MMLE）は、一致推定量として項目母数を推定することができる。MMLE では、受検者母数に正規分布などの分布を仮定し、(6)式から局外母数（nuisance parameter; その推測に直接的な関心がない母数）である受検者母数を積分消去（integrate out）／周辺化（marginalize）して得られる周辺尤度関数を最適基準（最適化関数）としてこれを最大化する項目母数の値を求める。

受検者iがn項目からなるテストを受検したとき、その反応パターンが$x_i=(x_{i1},x_{i2},\cdots,x_{in})$となる確率は、局所独立の仮定から、

$$P(x_i|\theta_i,\omega)=\prod_{j=1}^{n}P_j(\theta_i)^{x_{ij}}Q_j(\theta_i)^{1-x_{ij}} \tag{7}$$

と表せる。ただし、$\omega=(\omega_1,\omega_2,\cdots,\omega_n)$は項目母数ベクトルである。

MMLE では、受検者集団の θ は母集団分布$g(\theta|\tau)$からのランダムサンプル（標本）であると仮定する。ここで、τは平均や分散など分布の形を決める母数ベクトルである。局外母数を取り除くため、$g(\theta|\tau)$を用いて(7)式から受検者母数を積分消去すると、

$$P(x_i|\omega,\tau)=\int_{-\infty}^{\infty}P(x_i|\theta_i,\omega)g(\theta_i|\tau)d\theta_i \tag{8}$$

という反応パターンx_iの周辺確率が得られる。簡便のため、これ以降は(8)式を$P(x_i)$と略記すれば、全ての反応パターン（項目反応データ）Xが得られる確率は、

$$L=\prod_{i=1}^{m}P(x_i) \tag{9}$$

と表せる。(9)式を項目母数についての関数としてみたとき、これを周辺尤度関数（marginal likelihood function）と呼び、その尤度関数の値が最大となるときの項目母数を推定値として計算する。

計算を容易にするために(9)式の両辺の（自然）対数をとれば、

$$\log L=\sum_{i=1}^{m}\log P(x_i) \tag{10}$$

と対数周辺尤度関数（log marginal likelihood function）が求められる。この対数周辺尤度関数を最大化するためには、(10)式を項目母数に関して 1 階偏微分した式を 0 とおいた尤度方程式（likelihood equation）を項目母数に関して解けばよい。2 母数ロジスティックモデルを用いた場合、a_jとb_jに関する尤度方程式は、

$$\frac{\partial}{\partial a_j}(\log L)=\sum_{i=1}^{m}\int_{-\infty}^{\infty}\left[x_{ij}-P_j(\theta_i)\right]\left(\theta_i-b_j\right)[P(\theta_i|x_i,\omega,\tau)]d\theta_i=0 \tag{11}$$

$$\frac{\partial}{\partial b_j}(\log L)=-a_j\sum_{i=1}^{m}\int_{-\infty}^{\infty}\left[x_{ij}-P_j(\theta_i)\right][P(\theta_i|x_i,\omega,\tau)]d\theta_i=0 \tag{12}$$

となる。

推定値を求めるためには、(11)式と(12)式からなる連立方程式を解く必要がある。しかし解析的には解けないため、積分や最適化の際に Newton-Raphson 法などの数値計算法に基づく反復計算が用いられる。具体的には、たとえば、積分に含まれる連続型の分布を離散的な分布で近似する数値積分などを利用することを考える。ここでは、θ 軸上に求積点X_h $(h = 1,2,\cdots H)$をとり、母集団分布$g(\theta|\tau)$を近似する。母集団分布として標準正規分布を利用する場合は、-4≦ θ ≦4 を対象として等間隔に求積点を区切ってもよいし、さらに効率的な方法として Gauss- Hermite 求積法などを用いてもよい。なお、係数$A(X_h)$はX_hの近傍における関数$g(\theta|\tau)$の値である。このとき、積分を含んでいた(11)式と(12)式はそれぞれ、

$$a_j \quad : \quad \sum_{h=1}^{H}\sum_{i=1}^{m}\left[x_{ij} - P_j(X_h)\right]\left(X_h - b_j\right)[P(X_h|x_i,\omega,\tau)] = 0 \tag{13}$$

$$b_j \quad : \quad -a_j\sum_{h=1}^{H}\sum_{i=1}^{m}\left[x_{ij} - P_j(X_h)\right][P(X_h|x_i,\omega,\tau)] = 0 \tag{14}$$

と近似できる。また、$P(X_h|x_i,\omega,\tau)$は、(7)式とベイズの定理より、

$$P(X_h|x_i,\omega,\tau) = \frac{\prod_{j=1}^{n} P_j(X_{h'})^{x_{ij}} Q_j(X_h)^{1-x_{ij}} A(X_h)}{\sum_{h'=1}^{H}\prod_{j=1}^{n} P_j(X_{h'})^{x_{ij}} Q_j(X_{h'})^{1-x_{ij}} A(X_{h'})} \tag{15}$$

と書き換えることができる。

Newton-Raphson 法に必要な(13)式と(14)式の 2 次偏導関数は、

$$H_j = \begin{bmatrix} \dfrac{\partial^2}{\partial a_j^2}(\log L) & \dfrac{\partial^2}{\partial a_j \partial b_j}(\log L) \\ \dfrac{\partial^2}{\partial b_j \partial a_j}(\log L) & \dfrac{\partial^2}{\partial b_j^2}(\log L) \end{bmatrix} \tag{16}$$

で与えることができる。ここでH_jは項目jの母数に関する、ヘッセ行列Hの部分行列である。

尤度方程式を Newton-Raphson 法で解く場合、反復計算のための更新則は、t回目の更新で得られた解をω_tとして、

$$\omega_{t+1} = \omega_t - H_t^{-1} g_t \tag{17}$$

となる。ここで、ω_tはサイズ$2n \times 1$の項目母数ベクトル、g_tは(11)式 、(12)式で定義されるサイズ$2n \times 1$の 1 次偏導関数ベクトル、H_t はサイズ$2n \times 2n$のヘッセ行列、H_t^{-1}はその逆行列である。

同じ最適化問題を Fisher のスコアリング法で解く場合は、ヘッセ行列の代わりに Fisher の情報行列を利用すればよい。Fisher 情報行列は、ヘッセ行列の期待値にマイナスの符号をつけた行列として計算できる。Newton-Raphson 法や Fisher のスコアリング法については、Kendall and Stuart (1979)などが参考になる。

なお、いずれの場合も初期値$\omega_0 = \left(a_j^{(0)}, b_j^{(0)}\right)$としては、たとえば、2 値反応データの場合 Lord et al.(1968)の heuristic method で使われる、

$$a_j^{(0)} = \frac{\rho_{jX}}{\sqrt{\left(1 - \rho_{jX}^2\right)}} \tag{18}$$

および

$$b_j^{(0)} = -\frac{\Phi^{-1}\left(p_j\right)}{\rho_{jX}} \tag{19}$$

が利用できる。ここで、ρ_{jX}は項目反応データXから求められる項目jに関する主因子解、p_jは項目jに関する項目正答確率、Φ^{-1}は逆正規累積分布関数である。

このように、Bock and Lieberman (1970)の周辺最尤推定法は画期的な手法ではあったが、数値積分や逆行列演算には大きな計算負荷がかかる。そのため、計算機の演算スピードの向上と EM アルゴリズムを用いた Bock and Aitkin (1981)の手法が開発されるまで、大規模テストデータへの MMLE の実用は待たなければならなかった。EM アルゴリズムの理論的展開は、Dempster, Laird, and Rubin (1977)においてなされたものである。このアルゴリズムのもつ利点は、期待対数尤度を求める E ステップ（expectation step）と、その対数尤度を最大化するような項目母数の値を計算する M ステップ（maximization step）を交互に繰り返すことで着実に最尤推定値を求めることができる点にある。Bock and Lieberman (1970)の MMLE と比較すると、尤度方程式を解く前に、以下の(20)式と(21)式で表される期待度数を求めるステップを挿入する点、および尤度方程式が期待度数を用いて書き換えられている点が改良されている。

[E ステップ]

項目母数の仮の値（反復 1 回目は初期値、それ以降は前回の M ステップで求められた値）を与え、求積点X_h $(h = 1,2,\cdots,H)$において項目jに解答する期待人数f_{jh}とその中での期待正答者数r_{jh}を計算する。

$$f_{jh} = \sum_{i=1}^{m} P(X_h|x_i,\omega,\tau) = \sum_{i=1}^{m}\left[\frac{L_i(X_h)A(X_h)}{\sum_{h'=1}^{H} L_i(X_{h'})A(X_{h'})}\right] \tag{20}$$

$$r_{jh} = \sum_{i=1}^{m} x_{ij}P(X_h|x_i,\omega,\tau) = \sum_{i=1}^{m}\left[\frac{x_{ij}L_i(X_h)A(X_h)}{\sum_{h'=1}^{H} L_i(X_{h'})A(X_{h'})}\right] \tag{21}$$

[M ステップ]

E ステップで計算したf_{jh}とr_{jh}の値を以下の(22)式と(23)式に代入し、尤度方程式を数値的に解く。収束基準を満たす場合は反復を終了し、満たさない場合は E ステップに戻る。

$$a_j \quad : \quad \sum_{h=1}^{H}\left(X_h - b_j\right)\left[r_{jh} - f_{jh}P_j(X_h)\right] = 0 \tag{22}$$

$$b_j \quad : \quad -a_j\sum_{h=1}^{H}\left[r_{jh} - f_{jh}P_j(X_h)\right] = 0 \tag{23}$$

なお、推定された項目母数の標準誤差は、推定が終了した時点での Fisher の情報行列の逆行列における対角要素の平方根として求められる。

Bock and Aitkin (1981)の EM アルゴリズムを用いた周辺最尤推定法（MMLE with EM algorithm：MMLE-EM）は，BILOG-MG（Zimowski, Muraki, Mislevy, & Bock 2003）をはじめ、PARSCALE、EasyEstimation（熊谷 2009）などの専用の IRT 分析ソフトウェアで利用できる。また、θ の母集団分布$g(\theta|\tau)$には標準正規分布を仮定していることが多い。なお、項目反応モデルの母数推定については、Baker and Kim (2004)が非常に詳しい。

2.3.6 項目母数推定の際のサンプルサイズの影響

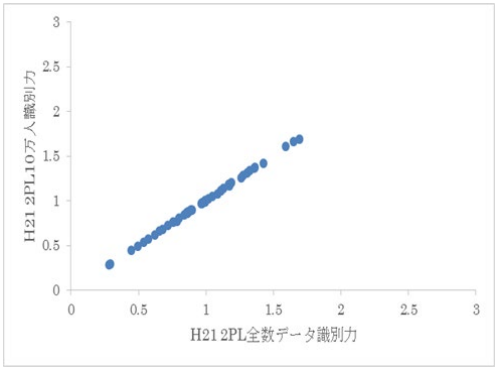
本節では、経年変化分析調査実施のために、項目母数の推定の際にサンプルサイズとしてどの程度が必要となるかをシミュレーションによって確認した。まず、項目識別力、項目困難度ともに全数データにおける推定値を基準にとる。具体的に、平成 21 年度全国学力学習状況調査の本体調査中学数学の全数データから無作為抽出したデータにもとづきそれぞれの項目識別力と項目困難度を求めた。その際のサンプルサイズは、100,000 人、10,000 人、2,000 人と順次 3 通り

に変化させた。次ページに、全数データにおける推定値と無作為抽出したデータにおける推定値をプロットして比較した図を掲載する（図 12)。原点を通り傾きが 1 の数直線上にプロットされている傾向が見られれば推測がうまく機能していることを示唆する。

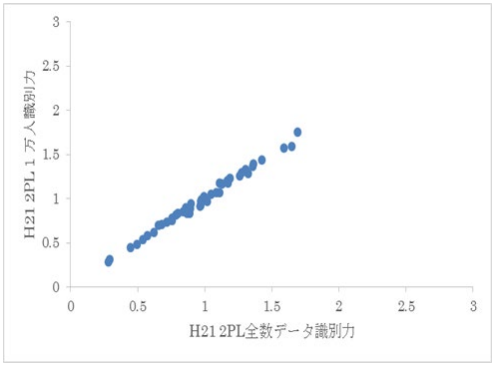
2 母数ロジスティックモデルを採用した場合には、項目困難度に関しては、サンプルサイズが 2,000 人でもほぼ全数データの推定値を復元できているが、項目識別力に関しては少なくとも 1 万名程度が必要なことがこれらの図から予想できる。標本調査法などに基づく数理的な根拠づけが必要ではあるが、このシミュレーション結果から判断して、経年変化分析調査の場合のサンプルサイズとしては少なくとも 10,000 人程度を準備することが望ましいであろう。

項目識別力

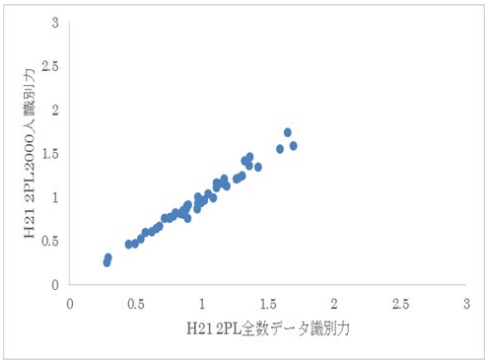
サンプルサイズ：100,000



サンプルサイズ：10,000

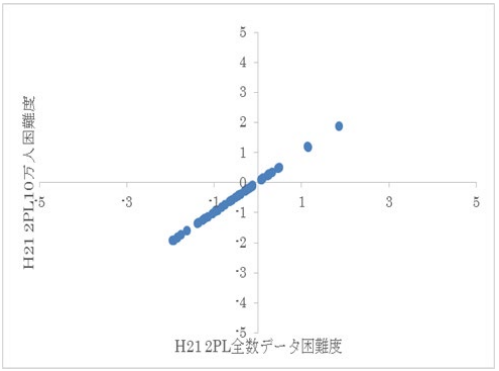


サンプルサイズ：2,000

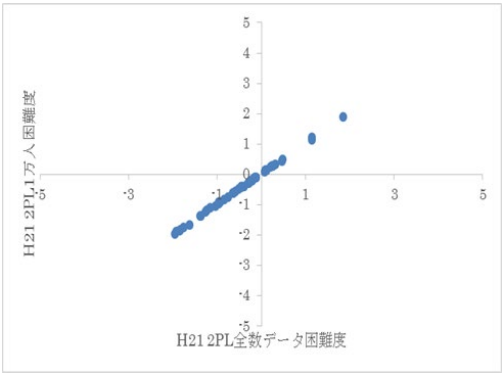


項目困難度

サンプルサイズ：100,000



サンプルサイズ：10,000



サンプルサイズ：2,000

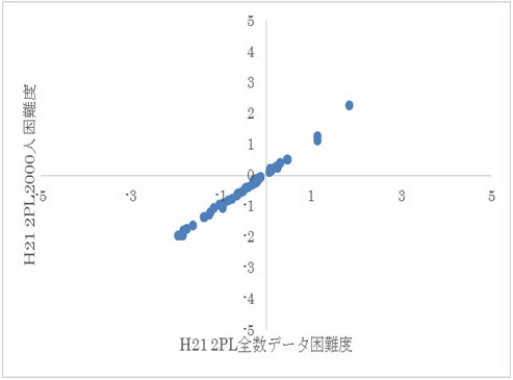


図 12：項目母数推定の際のサンプルサイズの影響

2.3.7 受検者母数の推定

IRT 分析ソフトウェアである BILOG-MG や EasyEstimation などでは、受検者母数の推定法として、最尤推定法（maximum likelihood estimation: MLE）、最大事後（maximum a posteriori: MAP）推定法、期待事後（expected a posteriori: EAP）推定法の三つが提供されていることが多い。本節では、項目母数が所与のときに受検者母数を最尤法によって推定する手続を、村木(2011)にしたがって概説する。

受検者iがn項目からなるテストを受検したとき、その反応パターンが$x_i = (x_{i1}, x_{i2}, \cdots, x_{in})$となる確率は、局所独立の仮定から、

$$L(x_i|\theta_i) = \prod_{j=1}^{n} P_j(\theta_i)^{x_{ij}} Q_j(\theta_i)^{1-x_{ij}} \tag{24}$$

と表せる。ここで、$P_j(\theta_i)$は尺度値θ_iをもつ受検者iが項目jに正答する確率であり、2 母数ロジスティックモデルを利用したときには(3)式で表される。また、$Q_j(\theta_i) = 1 - P_j(\theta_i)$は誤答確率である。(24)式の両辺の対数をとっても最大値をとるθ_iは変化しないので、計算を容易にするために(24)式の両辺の（自然）対数をとれば、

$$\log L\,(x_i|\theta_i) = \sum_{j=1}^{n} \left[x_{ij} \log P_j(\theta_i) + \left(1 - x_{ij}\right) \log Q_j(\theta_i)\right] \tag{25}$$

と対数尤度関数（log-likelihood function）が計算できる。この対数尤度関数を最大化するためには、(25)式をθ_iに関して 1 階偏微分した式を 0 とおいた尤度方程式をθ_iに関して解けばよい。(25)式の 1 次偏導関数は、

$$\begin{aligned} \frac{\partial \log L(x_i|\theta_i)}{\partial \theta_i} &= \frac{\partial}{\partial \theta_i} \sum_{j=1}^{n} \left[x_{ij} \log P_j(\theta_i) + \left(1 - x_{ij}\right) \log Q_j(\theta_i)\right] \\ &= \sum_{j=1}^{n} \frac{x_{ij}}{P_j(\theta_i)} \frac{\partial P_j(\theta_i)}{\partial \theta_i} + \sum_{j=1}^{n} \frac{1 - x_{ij}}{Q_j(\theta_i)} \frac{\partial Q_j(\theta_i)}{\partial \theta_i} \end{aligned} \tag{26}$$

と計算できる。(24)式の$P_j(\theta_i)$が 2 母数ロジスティックモデルであれば、

$$\frac{\partial P_j(\theta_i)}{\partial \theta_i} = D a_j P_j(\theta_i) Q_j(\theta_i) \tag{27}$$

と、1 次偏導関数が計算できる。 このように偏微分の結果が数学的に整理された形で表現できることも IRT モデルとしてロジスティック関数を用いる利点の一つである。

実際に、(27)式を 0 とおいた尤度方程式は、非線形方程式となるために解析的には解くことができない。コンピュータなどを用いて数値的に解くことが必要で、その際に Newton-Raphson 法や Fisher のスコアリング法などの数値計算法がよく利用される。Newton-Raphson 法では、反復計算を繰り返してθ_iの推定値を計算する。$t+1$回目の反復計算における更新則は、

$$\left[\hat{\theta}_i\right]_{t+1} = \left[\hat{\theta}_i\right]_t - \left[\frac{\partial^2 \log L(x_i|\theta_i)}{\partial \theta_i^2}\right]_t^{-1} \left[\frac{\partial L(x_i|\theta_i)}{\partial \theta_i}\right]_t \tag{28}$$

で与えられる。初期値としては、n項目からなるテストにおいて受検者iの正答数得点をXとするとき、 $\left[\hat{\theta}_i\right]_0 = \log\left[X/(n-X)\right]$とするとよい（log は自然対数）。2 母数ロジスティックモデルを利用する場合、(25)式の 2 次偏導関数は、

$$\frac{\partial^2 \log L\,(x_i|\theta_i)}{\partial \theta_i^2} = \sum_{j=1}^{n} D^2 a_j^2 P_j(\theta_i) Q_j(\theta_i) \tag{29}$$

となる。Fisher のスコアリング法では、(29) 式の代わりに Fisher の情報関数を(28)式の右辺第 2 項に当てはめて反復計算する。Fisher 情報関数は、(29)式の期待値にマイナスの符号をつけたものである。そのため Newton-Raphson 法と Fisher のスコアリング法の更新則は一致する。

受検者母数の最尤推定値の漸近的な標準誤差は、(5)式のテスト情報量を用いて、

$$SE_{\theta_i} = [I(\theta_i)]^{-\frac{1}{2}} \tag{30}$$

の関係式から計算できる。2 母数ロジスティックモデルの場合は、

$$SE_{\theta_i} = \left[D^2 a_j^2 P_j(\theta_i) Q_j(\theta_i)\right]^{-\frac{1}{2}} \tag{31}$$

となる。なお、テストの項目数nが大きくなるほど、受検者母数の最尤推定値は受検者の真の尺度値に近づき、その真の標準誤差も(30)式の値に近づく。この漸近的性質は、項目数が 20 以上の場合に実用的な意味で利用可能であることが知られている。

BILOG-MG などでは、受検者母数の最尤推定値を求めるのに Fisher のスコアリング法が採用されている。デフォルトの設定では、事前にコマンドファイルに設定した反復回数の最大値に達するか、全てのθ_i $(i = 1,2,\cdots,m)$においてt回目と$t+1$回目の反復における推定値の差が 0.01 未満になったときに反復計算は終了するようになっている。推定結果を利用する際には、出力をよく見て推定値が問題なく収束したものであることを確認する必要がある。

最尤推定法は、反応パターンx_iによっては推定値が収束しない場合がある。また、受検者が全問正解あるいは全問不正解の場合はその受検者の推定値を求めることができない。それに対し、MAP 推定法と EAP 推定法では、事前分布の設定により、どのような反応パターンでも対応する推定値を求めることができる。この点では、MAP 推定法と EAP 推定法のほうが最尤推定法より優れているようにみえる。しかし、最尤推定値が収束せず発散するのは、反応パターンなどに何らかの検討すべき課題が含まれている可能性があるとも考えられる。MAP 推定法や EAP 推定法を用いると、そのような問題が見過ごされてしまう危険性は否定できない。経年変化分析調査では、各推定法の長所と短所を総合的に考慮し、また PISA 等のクラスターとは異なり各分冊の中でもある程度の項目数が確保されていることから、一致性の観点も踏まえて、受検者母数の推定法として最尤推定法を利用することとした。

2.4 推算値

一般に学力調査の結果を利用するときに、各受検者の尺度値に注目するのか、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への評価や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を見越した集団統計量が考慮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ている推算値（plausible value：PV's）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、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、2.4.4 節では、von Davier, Gonzalez, and Mislevy (2009)にもとづき推算値を利用することの有用性を考察する。

2.4.1 推算値の必要性

推算値は、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、多重代入法（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 が代入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 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、個々の児童生徒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（母集団）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。例えば、ある人種や民族、または特定の地域や階級に相当する下位集団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。

一方、PISA などの国際的な学力調査では、テストの設計（デザイン）に重複テスト分冊法が利用されている。一人の受検者が解答する冊子の中にはクラスター（cluster）と呼ばれる項目群がありその群の中の項目数は、個々人の能力推定を目的とするテストと比べてはるかに少ない。具体的には、一つの冊子の中のあるクラスターには科学的リテラシーに関する項目群、別のクラスターには数学的リテラシーに関する項目群、さらに別のクラスターには質問項目群などが含まれ、それぞれのクラスターは 5 項目程度から構成されている。そのため、個々人の能力をクラスター5 項目程度で正確に推定することには限界がある。その結果、そこから求められる集団統計量の精度にも疑問が残る。このような問題への解決策となる一つの方法論がベイズ統計の枠組みを用いた推算値法である。

推算値は、各受検者が解答する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児童生徒の割合のように、一定の得点以上の受検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。

2.4.2 推算値の定義

通常、IRT モデルでは、受検者の反応パターンと項目母数の推定値を用いて受検者の能力母数 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

（MLE)、事後分布の最大値を与える θ を推定値とする場合は MAP 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 推定値が得られる。それに対し、受検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 13 に、最尤推定値、MAP 推定値、EAP 推定値、推算値の概念図を示す。なお、事後分布が左右非対称になれば、MAP 推定値と EAP 推定値は異なる値を一般に示す。

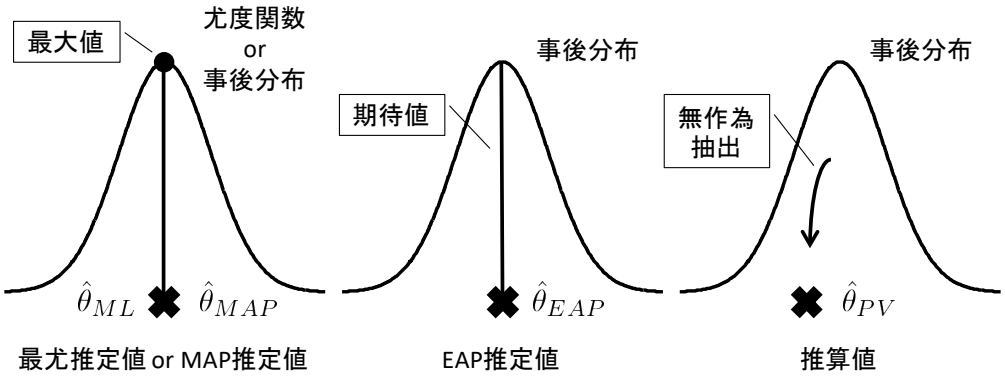


図 13：最尤推定値、MAP 推定値、EAP 推定値、推算値の概念図

数学的な説明のため、受検者の項目反応パターンをx、能力母数を θ、尤度関数を$f(x|\theta)$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$g(\theta)$~$N(\mu,\sigma^2)$を仮定する。このとき、事後分布$h(\theta|x)$は、

$$h(\theta|x)=\frac{f(x|\theta)g(\theta)}{\int f(x|\theta)\,g(\theta)d\theta}\propto f(x|\theta)g(\theta) \tag{32}$$

と表される。つまり、ある受検者の項目反応パターンがxであれば、その受検者の能力母数 θ の事後分布は(32)式で与えられる。この(32)式からの無作為標本が項目反応パターンxをもつ受検者の推算値である。

図 13 からも分かるように、推算値は、個々の能力推定値の一つのサンプル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、個々の標本は各受検者の尤もらしい推定値とは必ずしもならないが、それが受検者の能力推定値だけでなく、それに付随する推定の不確かさについての情報も与えてくれる。この点がシミュレーション研究などにもとづき、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ための基盤を与える根拠とされている。なお、事後分布$h(\theta|x)$からの標本抽出には、棄却サンプリング（rejection sampling）（津田 1995）などを利用することができる。

2.4.3 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、それを直接用いて集団の平均、分散、パーセンタイルなどを計算することが考えられる。受検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、受検者の能力値が EAP 推定値の場合、その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 推定値のそれぞれにおいて、受検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。

推算値は各受検者における能力母数 θ の事後分布からの 1 組の無作為標本である。また、個々人の 1 組の無作為標本を集めた分布は、その集団の能力分布の推定結果を与える。逆に言えば、個々人の推算値の組は、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、対象集団内の個々人から得られるたった 1 組の推算値を用いるだけでも（分散に関する）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、分散の推定値に偏りが生じてしまうことと対照的である。

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)の多重代入法（multiple imputation）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] 各受検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 2 つの計算方法が考えられる。

しかしながら、推算値の定義を踏まえると、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、全体の計算量を減らそうとして、[PV-W]のように各受検者の推算値を平均してはいけない。図 13 からもわかるように、各受検者の推算値を平均することは、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推定した関心のある統計量の分散（集団統計量の推定の誤差分散）は、群内分散と群間分散への分散の分解式のように、

$$\hat{V}_{IMP} = \left(1 + \frac{1}{K}\right)\left[\frac{1}{K-1}\sum_i \left(M_{PV_i} - \bar{M}_{PV}\right)^2\right] + \frac{1}{K}\sum_i \hat{V}\left(M_{PV_i}\right) \tag{33}$$

と表される（Little & Rubin 2002）。ここで、M_{PV_i}は$i(1,2,\cdots,K)$組目の推算値を用いて算出した集団統計量、$\bar{M}_{PV}$は全ての組に亘って計算された集団統計量の平均値、$\hat{V}\left(M_{PV_i}\right)$は$M_{PV_i}$の誤差分散の推定量である。(33)式の正の平方根が標準誤差に相当する。

前節で述べたように、NAEP が代入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（多重代入法に基づく）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、個々の推算値はその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検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、二人の受検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の推定や分布推定で活用されるべき量であり、それに基づ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際の有用な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。例えば、4 項目から構成されるテストがあれば、受検者の得点は 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（＝識別力母数が項目間で同じと仮定したモデル）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 14 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 未満の受検者の割合に興味があるとする。EAP 推定値の場合、-1 未満の受検者の割合は、0 点をとった受検者の割合に等しい。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、能力分布のパーセンタイルのより正確な推定に活用することができる。

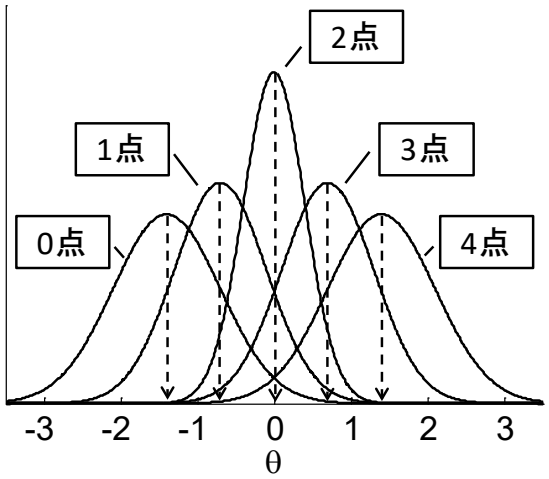


図 14：各得点の事後分布と EAP 推定値

（注）Wu,2004,Figure を単純化した図になる

2.4.4 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のよう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、集団統計量の計算に推算値を用いる利点とその実際について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（[PV-R]）であった。とくに（各受検者が解答する）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直接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。解析的な評価は難しいものの、個々人の正確な能力

値を報告できるほどの多数の項目を利用できる場合は、集団の能力分布を推定に必ずしも推算値を用いる必要はない可能性がある。現行の経年変化分析調査では分冊あたりの項目数が20程度あることも踏まえて、取り扱いの点でも利点の多い最尤推定値を用いることとした。

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（または分散）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。また、経年変化分析調査の場合、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。例えば、母集団の中で基準点に到達する児童生徒の割合を推定する場合、わずか1%や2%の推定結果の違いだけであっても結果として数多くの児童生徒や学校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、経年変化分析調査のように集団統計量や能力分布の推定に関心があるテストに際しては、将来的には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、例えば教育経済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、推算値は、今後、大規模学力データに関して必要となる分析ノウハウの一つである。

2.5　受検者母数推定値による集団統計量の性質

ここでは（推算値を利用せずに）個々の受検者母数の推定値を直接用いて集団統計量（平均、標準偏差）を算出した場合、どの程度のバイアス（偏り）が生じ、また異なる推定法の間で推定値にどのような相関関係や差異があるのかをシミュレーションで確認する。

2.5.1 シミュレーション手続

シミュレーションの手続は下記に示すとおりである。なおこのシミュレーションでは繰り返し発生による結果の安定性の評価は行っていないためおよその傾向のみの把握となる。

① 2母数ロジスティックモデルを真のモデルとして、 20項目、$n = 1000$人の項目反応データを数値的に発生させる。
② その際、aパラメタ（項目識別力）は平均が-0.2、 SD(標準偏差)が0.2である対数正規分布より、bパラメタ（項目困難度）は標準正規分布より乱数を発生させて設定する。
③ また、真の受検者母数θは$N(0.8, 0.8^2)$から発生させる。

④ 項目母数は既知として前項で発生させたものを利用し、EasyEstimation を用い、MLE、EAP、MAP、PV（PV-R: PV 数=5 組）、POP により、θ の推定値（もしくは母集団統計量）を計算した。MLE は全問正答、全問誤答の反応パターンを示した受検者に対する EasyEstimation による補正がある。また、POP とは EasyEstimation の一つのオプション指定であり、これが指定されると EasyEstimation は Mislevy & Bock (1982)による潜在特性値の分布推定"estimation of the latent distribution"を実行する。これは得られたデータから事前分布を逆に逐次的に推定する方法である。

2.5.2 シミュレーション結果

推定方法ごと(MLE、EAP、MAP)に、1,000 人の平均値、標準偏差（POP の場合は、推定された潜在特性値の分布から直接計算）を示したものが表 1 となる。表中の TRUE は、「方法」で発生させた θ の真の値を用いた場合の結果であり、受検者数(=1000)が有限であることを反映して、母平均(=0.8)と母分散(=0.8^2)の値とは一致しない PV (PV-R)については、5 つの PV の組ごとに平均値・標準偏差を計算して、その 5 つの値の平均値を便宜的にとったものである。EAP、MAP、PV については、事前分布として標準正規分布 $N(0, 1^2)$とした。

表 1：推定方法ごとの集団統計量の違い

	TRUE $N(0.8, 0.8^2)$	MLE	EAP	MAP	PV	POP
Mean	0.829	0.873	0.722	0.694	0.719	0.831
SD	0.809	0.929	0.722	0.698	0.817	0.797

受検者母数の推定法の違いによって以下のような特徴があることを指摘できる。

A) TRUE は、設定値なのでほぼ $N(0.8, 0.8^2)$どおり。
B) MLE は、平均値がやや大きい。また SD については大きくなっている。
C) EAP、MAP は、平均値が事前分布の平均(0.0)に引き寄せられ、SD も縮小して 0.8 より小さくなる。この現象は縮小推定（shrinkage estimation）として知られる。MAP の方がその影響が大きい。
D) PV も平均値は事前分布の平均(0.0)に引き寄せられる。一方、SD は影響しない。これが PV を用いるメリットの一つとされる。
E) POP は、ほぼ TRUE と同じ値になっている。

また、真の θ の値と各推定法における推定値の差異は以下の通りである。$\sqrt{1/n\sum(\hat{\theta}_i - \theta_i)^2}$で定義される RMSE では EAP が最小となる一方、$1/n\sum(\hat{\theta}_i - \theta_i)$で定義される BIAS では MLE が最小

となっていることがわかる。

表 2：真の θ の値と各推定値間の差違

	MLE	EAP	MAP	PV1	PV2	PV3	PV4	PV5
RMSE	0.456	0.393	0.402	0.538	0.563	0.549	0.535	0.544
BIAS	0.044	-0.107	-0.135	-0.106	-0.111	-0.115	-0.109	-0.107

さらに真の θ の値および各推定法の推定値間の相関係数と散布図はそれぞれ表 3、図 15 に示す通りである。相関係数の高さから MLE、EAP、MAP の間にはほぼ一対一対応の関係があることが観察できる。しかし全問正答または全問誤答の反応パターンについての処理の違いも反映された結果、MLE と EAP、MAP の間にはやや曲線的な関係が見られる一方で、EAP と MAP の間には直線的な関係が認められる。

表 3：受検者母数推定値間の相関係数行列

	TRUE.	MLE	EAP	MAP	PV1
TRUE.	1.0000	0.8725	0.8840	0.8841	0.7888
MLE	0.8725	1.0000	0.9908	0.9902	0.8753
EAP	0.8840	0.9908	1.0000	1.0000	0.8824
MAP	0.8841	0.9902	1.0000	1.0000	0.8823
PV1	0.7888	0.8753	0.8824	0.8823	1.0000

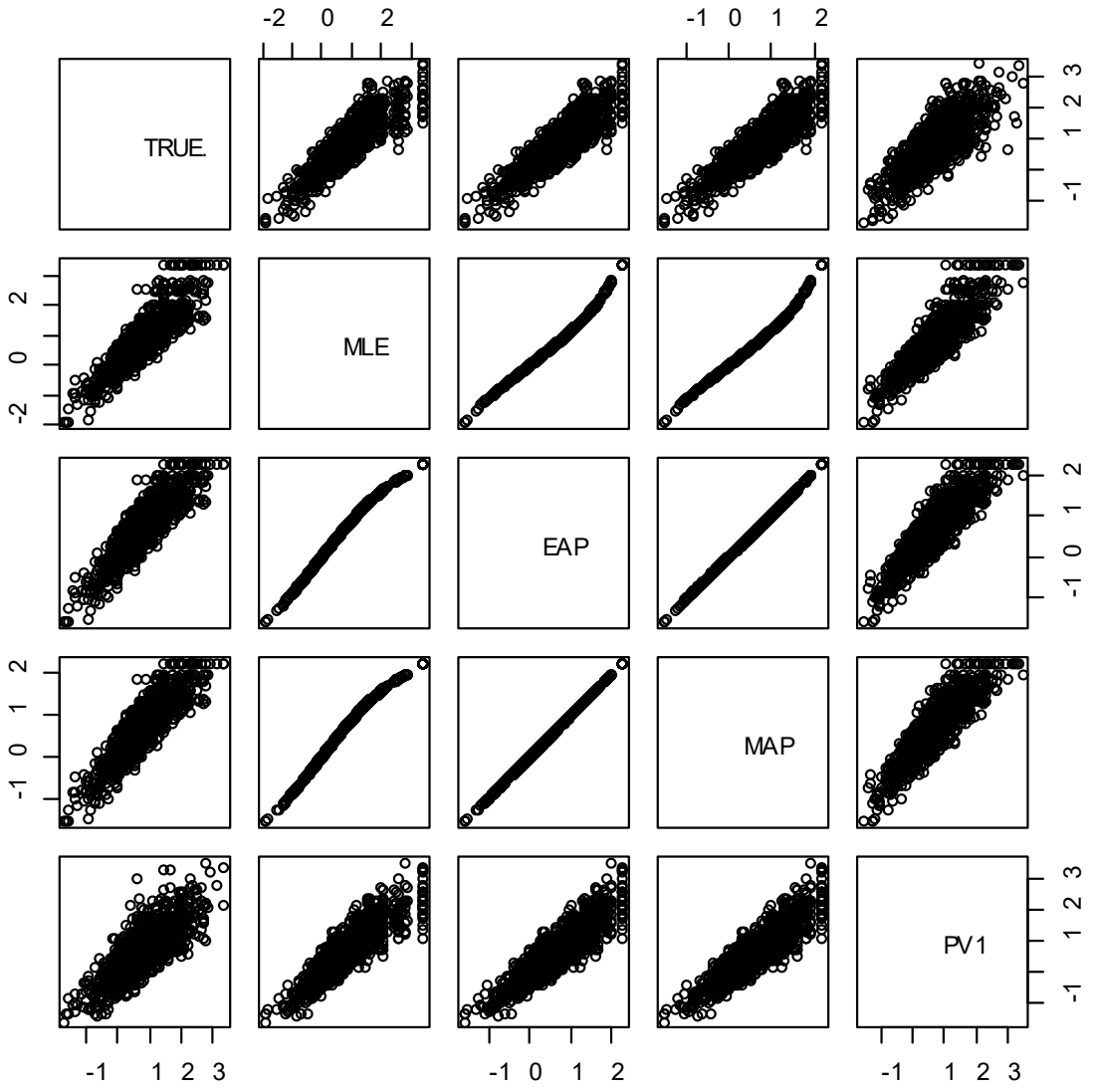


図 15：受検者母数の推定値間の散布図

2.6　経年比較のための尺度等化

2.6.1 スコアを比較するとは

含まれる項目の種類が異なる 2 つ（もしくはそれ以上）のテスト得点は、そのままでは相互に比較をすることができない。受検者の学力が同一であったとしても、難しい項目群で構成されたテストではその得点は低くなり、易しい項目群で構成されたテストではテスト得点は高くなる。項目の困難度が等しいテストであっても、学力が高い受検者集団が受検したテストの平均得点は高くなり、学力が低い受検者集団においては平均点は低くなる。そこで、このような 2 つのテスト得点を比較可能にするための手続きが必要となる。 Holland & Dorans (2006)ではこの手続きを総称してリンキング（linking）と呼び、さらにその目的やテストの性質により、“predicting”、”scale aligning”、“test equating”の 3 つの下位分類を設けている(図 16)。

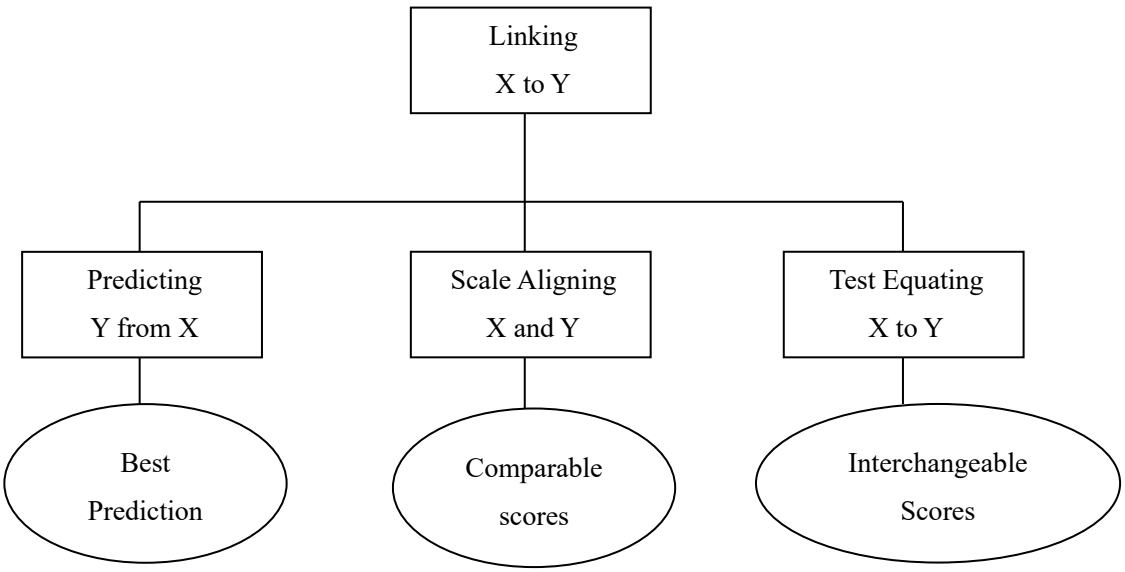


図 16：リンキングの下位分類

（注）Holland & Dorans(2006）より引用

ここで“predicting（予測）”とは、名称通りテスト得点 X からテスト得点 Y を予測する関係の場合のリンキングである。X から Y の方向のみを扱う「非対称性(asymmetry)」が特徴であり、具体的には入学試験の得点から、入学後の成績を予測する場合などが挙げられる。“predicting”では、2 つのテスト得点間で“最良な予測(best prediction)”ができるかどうかが問題となる。

“scale aligning（尺度調整）”（または、scaling）では、テスト X の得点とテスト Y の得点を相互に関係づけることになる。テスト X の 60 点はテスト Y の 40 点に相当し、テスト Y の 50 点はテスト X の 75 点に相当する、というような関係づけを行うのである。これは言い換えれば、

テスト X とテスト Y を“共通尺度（common scale）”上で表現しているとも言える。scale aligning では 2 つのテスト得点が“比較可能な（comparable）”な関係である。さらに scale aligning は、2 つのテストで測定しようとしている内容（構成概念）が同じかどうか、2 つのテストの信頼性が等しいかどうかなどにより、“calibration（較正）”、“concordance（対応づけ）”などさらなる下位分類が存在する広い概念でもある。

“test equating（テスト等化）”では、2 つのテスト得点が“相互に交換可能（interchangeable）”な関係を条件とする。“相互に交換可能”な関係について次のような例を考えてみる。例えば、小学校 6 年生向けの算数のテストと 4 年生向けのテストを相互に関係づけたとする。このとき、小学校 6 年生向けテストの 40 点が 4 年生向けテストでは 70 点に対応するという結果が得られたとしても、実際に 4 年生向けテストで 70 点を得た 4 年生児童が、6 年生向けテストを受検したときに 40 点を得ることができるわけではない。現実には学習内容の履修・未履修などが大きく関係するが、それ以外にも、a）測定内容（構成概念）が両テスト間で異なること、b）6 年生向けテストには 4 年生向けの（易しい）問題項目がないため信頼性が低下する、などが影響するからである。このような状態は、“相互に交換可能”な関係ではないとされる。2 つのテスト得点が“相互に交換可能な”状態であるためには、2 つのテストが測定する構成概念が同一であり、かつ両者の困難度が等しいという状況が必要となる。また、Dorans et al.(2000)がガイドラインとしてまとめたように、

a. 測定対象となる構成概念が同一であること(the equal construct requirement)
b. 信頼性が等しいこと（the equal reliability requirement）
c. 対称性が保たれていること（the symmetry requirement）
d. どちらのテストを受けても同等であること（the equity requirement）
e. 母集団不変であること（population invariance requirement）

の 5 つの条件をすべて満たす場合のみを等化と呼ぶ場合もある。

2.6.2 垂直等化と垂直尺度化

なお、学力の幅広い発達段階に共通して使える尺度をもとめる方法としては「垂直尺度化」がある。この手法も、従来は困難度に着目して「垂直等化（vertical equating)」として分類されてきたが、“相互に交換可能”かどうか、また Dorans et al.(2000)の 5 条件を満たさない等の観点から「垂直尺度化（vertical scaling)」として、Scale Aligning の下位分類の 1 つに区分されるようになっている。この手法の限界としては学年が進むにつれて、測定対象としている「学力」の質的変化を同一の尺度で捕捉できなくなるため複数の学年をまたぐ 1 次元尺度の妥当性に問題が出てくることである（Kolen & Brennan 2014)。

このように Test Equating は、Linking の 3 つの下位分類の中で最も強い制限を必要とするものである。このテクニカルレポートでは以後 Test Equating を「等化」と呼ぶこととする。

2.6.3 IRT における等化

素点（配点に重みを設けずに正答なら 1 点、誤答なら 0 点とする）によるテストの等化手続きとしては、古典的テスト理論の枠組みの中で、線形等化法や等パーセンタイル等化法などがよく用いられてきた（池田 1994)。しかし、IRT によるテスト分析が普及するに伴い、等化の議論は IRT の枠組みの中でも行なわれるようになってきた（村木 2011)。

IRT の枠組みにおける等化では、等化を行う 2 つのテストについてそれらを繋ぐ何らかの情報が必要となる。通常、これらは 2 つのテストを同時に受検する「共通受検者」や、2 つのテストに共通に含まれる「共通項目」などを設定することによりなされる。さらには、テストを実施する様々な状況・制限によりこれらの方法を組み合わせて用いることもある。このように、等化されるテストについて、どのようなデータ収集デザイン（data collection design）を採用するかは非常に重要な問題となる。

等化のデザインが決定されると、実際にどのような計算手続きで 2 つのテストを等化するかが次の問題となる。計算手続きには、等化係数を用いる方法や、同時尺度調整法などいくつかの方法が存在するが、それらのうちいずれを選択するかは等化デザインと大きく関わっている。ただしここで気をつけなければいけないのは、等化デザインと等化の計算手続きとが必ずしも 1 対 1 で対応しているわけではないことである。ある等化デザインでテストが実施されたときに、複数の方法で等化の計算を行うことが可能な場合も多いのである。また、等化デザインは、受検者数、受検者への負担、テスト実施時のコスト、テストに含まれる項目数など様々な実施上の制約の下で決定していかなければならないことにも注意が必要である。

2.6.4 等化のためのデータ収集デザイン

先に述べたように、IRT における等化のためには 2 つのテストを繋ぐ何らかの情報が必要となる。通常これらを実現するために、いわば基本型とでもいうべき、いくつかのデータ収集デザイン（等化デザイン；equating design）が提案されている。それぞれのデザインごとに長所・短所があり、現実場面では様々な制約状況の中で、どのようなデザインを組むのか決めていくことになる。また、重複テスト分冊法では調査精度を担保するためにこれらのデザインがすべて組み込まれている。

（1） 単一グループデザイン

いちばん単純なデータ収集デザインとして単一グループデザイン（single-group design : SG）

がある（図 17）。これは調査対象となる母集団からの抽出標本としての受検者グループ P にテスト X とテスト Y を同時に実施することでデータを収集する。同じ受検者グループが受けているため、学力分布は同一であることから、テスト X とテスト Y に関するテスト得点の分布の統計的な性質の差異はテスト X とテスト Y の性質に帰着されることを利用して等化を行える。なお、SG デザインは共通受検者デザイン（common-subjects design）とも呼称されることがある。

グループ	テスト X	テスト Y
P	○	○

図 17：単一グループデザイン（SG）

SG デザインでは、受検者の人数を多くすることで、等化の精度を上げることができる。後述するアンカーテストデザインでは、等化の精度を上げるためにはアンカーとなる共通項目の数を多くしなければならないが、1 つのテストに含めることができる項目数などの制限を受ける。単一グループデザインでは、原理的には受検者の人数をある程度以上確保できれば安定した等化を行うことができることから、NAEP でも用いられている（村木 2006）。一方、現実場面においては、2 つのテストをほぼ同時に受検する集団を確保することは、様々なコストの面で難しい場合も多い。

（2） 等価グループデザイン

SG デザインは、等化を行う 2 つのテストを同一受検者グループに受検させるデザインであった。このとき、受検者は 2 つのテストを同時に受検する必要があるため受検者への負担が過重になるというデメリットが生じる。そのため、現実の実施場面では、調査対象となる母集団から標本抽出された異なる 2 つの受検者群を準備し、その各々にテスト X またはテスト Y を実施することを考える。これが等価グループデザイン（equivalent-groups design : EG）である（図 18）。同一母集団から互いに独立に標本抽出された受検者群であるため、期待される学力分布も同一と見なせることを利用し等化を行うことができる。また、各群の受検者はテスト X かテスト Y のどちらかひとつだけを受ければ良いため、SG デザインよりは受検者への負荷が軽減されるメリットがある。

グループ	テスト X	テスト Y
P_1	○	
P_2		○

図 18：等価グループデザイン（EG）

（3）カウンターバランスデザイン

テストを受ける際には一般に長時間の解答時間から生じる疲労や問題の出題順による正答率への影響などのいわゆる持ち越し効果（carry-over effect）が存在する。このような効果を相殺して等化精度を高めるために工夫されたのがカウンターバランスデザイン（counterbalanced design：CB）である（図 19）。CB デザインでは母集団 P からの異なる 2 つの独立標本としての受検者群 P_1、P_2 を考え、かつテスト X とテスト Y の両方を、受検者群間で順番を入れ替えて実施することで持ち越し効果を打ち消し、等化精度を高めている。ただし、同じ受検者がテスト X とテスト Y を受けなければならないため、受検者への負荷という点では SG デザインと同じデメリットを抱えている。

グループ	テスト X	テスト Y
P_1	1	2
P_2	2	1

図 19：カウンターバランスデザイン（CB）

（注）カウンターバランスデザインの表中の数字は実施順を示す

（4）アンカーテストを伴う不等価グループデザイン（NEAT）

アンカーテストデザイン(anchor test design：AT)は、等化を行う 2 つのテストの他に、それらを繋ぐもう 1 つのテスト（係留テスト、アンカーテスト: anchor test）を利用して、等化を行うデザインである（図 19）。これは、共通項目デザイン（common-items design）と SG デザインを組み合わせた方法と見ることもできる。すなわち 2 つの受検者群にそれぞれテスト X とテスト Y を実施することに加え、両群に共通して実施するテスト A を設ける。テスト X とテスト A に関しては SG デザイン（テスト Y とテスト A に関しても同様）となっている一方、テスト A に関しては 2 つの受検者群が共通受検者となっている。

アンカーテストデザインでは共通項目から情報を獲得できるため、この点を活かしさらに実用的にしたデザインとしてアンカーテストを伴う不等価グループデザイン（non-equivalent groups with anchor test design : NEAT）がある。

グループ	テスト X	テスト A	テスト Y
P	○	●	
Q		●	○

図 20：アンカーテストを伴う不等価グループデザイン（NEAT）

(注)P、Q はそれぞれ互いに異なる母集団からの標本を表す

NEAT デザインはアンカーテスト A の情報を使ってテスト X とテスト Y の等化を行うことができる上、受検者集団がかならずしも同一母集団から抽出された標本である必要がないため、受検者数の確保の点でメリットがある。さらにテスト A を過重にしなければ、受検者への負荷も SG デザインや GB デザインほどはかからない。しかしその一方で、テスト A に含まれる項目数が非常に少ない（例えば 10 項目以下などの）場合、等化精度に問題が生じるというデメリットもある。言い換えれば、共通項目の数が多い方が等化の精度が高まるが、1 つのテストの項目数に制限（テスト X とテスト A を合計した項目数に制限）がある場合など、共通項目の確保という点から NEAT デザイン（アンカーテストデザインを含む）を用いることが難しい場合も多々ある。

以上、基本となる等化のためのデータ収集デザインを説明してきたが、等化デザインを決定する際に等化精度の担保の上で重要なのは、共通項目数や共通受検者数など、テストをつないでいる共通の情報をできるだけ多くすることである。 しかし、一方で、等化デザイン決定においては、どのような状況でも対応できるというような決定的なものは存在しない。精度、実施可能性、コストなど様々な条件を勘案して構築することが必要不可欠である。さらに等化にあたっては、次項で述べるどの等化方法を使うかもその都度収集されたデータから専門的かつ総合的な検討と判断が求められると言うことも重要な経験的事実である。

なお、経年変化分析調査は中長期にわたって継続実施されることが予定されている。その結果、実施ごとに行う等化の誤差などが累積していくことが予想できる。そのため、どこかの時点で基準尺度の較正（目盛りあわせ）は必要である。そのことも見通して等化デザインを組む必要がある。

2.6.5 さまざまな等化方法

等化デザインが決定された後、そのデザインに従ってテストデータが収集される。そして、得られたテストデータを用いて、実際に等化の計算が行なわれる。等化デザインと同様に、等化の計算手続きにおいても多種多様な方法が提案されている。以下に、その代表的なものを紹介する。

（1）等化係数を用いた方法

IRT においては、受検者母数といくつかの項目母数からなる項目特性関数により、項目に対する正答確率を表すことになる。代表的な IRT モデルである 2 母数ロジスティックモデルでは、尺度値 θ をもつ受検者が項目jに正答する確率$P_j(\theta)$を、

$$P_j(\theta) = \frac{\exp\left[Da_j(\theta - b_j)\right]}{1 + \exp\left[Da_j(\theta - b_j)\right]} \tag{34}$$

として表す。このとき受検者母数 θ について、

$$\theta^* = K\theta + L \tag{35}$$

と線形変換を施したものを θ^*とする。同時に項目母数a_j、b_jについても、

$$a^* = \frac{a}{K} \tag{36}$$

$$b^* = Kb + L \tag{37}$$

のように線形変換したものをa^*、b^*とすると、

$$P_j(\theta) = \frac{\exp\left[Da_j(\theta - b_j)\right]}{1 + \exp\left[Da_j(\theta - b_j)\right]} = \frac{\exp\left[Da_j^*(\theta^* - b_j^*)\right]}{1 + \exp\left[Da_j^*(\theta^* - b_j^*)\right]} = P_j(\theta^*) \tag{38}$$

のように線形変換を施す前と後とで正答確率は変わらない。このことはすなわち、学力を表現する尺度値 θ について、K、Lを用いて原点と単位とを任意に決定できることを表している。通常の IRT モデルを用いた分析では、分析に用いたデータセットにおける受検者集団の母集団について、平均 0、標準偏差 1 となるように基準化した上で項目母数の推定などが行なわれる。前節の例で言えば、テスト X、テスト Y はそれぞれ独立に分析した状態では、それぞれの母集団分布の平均が 0、標準偏差が 1 となっており、直接それらを比較することができない。そこで、(35)式を用いて、例えばテスト Y の原点と単位をテスト X のそれに合わせるような作業を行うことで、両テストを比較可能にするのである。このときに用いているK及びLを等化係数と呼ぶ。

等化係数を用いた等化では、いかにしてこの等化係数を定める（推定する）のかが問題となる。等化係数の推定にも様々な方法が提案されているが、なかでも最も簡便な方法として Marco(1977)による mean / sigma 法が挙げられる。この mean / sigma 法は、テスト X 及びテス

トＹに含まれる共通項目について、それぞれデータセットから推定された困難度母数b_X、b_Yの平均μ_{b_X}、μ_{b_Y}、標準偏差σ_{b_X}、σ_{b_Y}を用いて、$K = \sigma_{b_Y}/\sigma_{b_X}$、$L = \mu_{b_Y} - K\mu_{b_X}$として等化係数の推定を行うものである。なお共通受検者デザインの場合には、2 つのテストから得られる共通受検者の尺度推定値θ_X、θ_Yの平均μ_{θ_X}、μ_{θ_Y}、標準偏差σ_{θ_X}、σ_{θ_Y}を用いて、同様に$K = \sigma_{\theta_Y}/\sigma_{\theta_X}$、$L = \mu_{\theta_Y} - K\mu_{\theta_X}$として mean / sigma 法を適用することが可能である。

なお、mean / sigma 法の他にも、2 つの項目特性関数の差を最小にする方法（Haebara 1980）や、テスト特性関数の差を最小にする方法（Stocking & Lord 1983）、共通受検者の項目反応パターンを利用して等化係数を最尤推定する方法（野口 1986）、mean / sigma 法において推定尺度値の分散に含まれる誤差成分の大きさを推定し取り除くという補正方法（野口・熊谷 2011）など様々な方法が提案されている。

（２）同時尺度調整法

同時尺度調整法（concurrent calibration）は、等化デザインの下で得られたテストデータ全体を用いて、同時に項目母数の推定を行うことで等化を行う方法である。アンカーテストデザインにおいて、得られたデータを不完全データ行列（incomplete data matrix）として見たとき、項目が提示されない部分（未提示項目：not presented items）が存在する。このデータ行列全体について項目母数の推定を行うことで、データ全体が 1 つの尺度上で表現されることとなり、等化が実現される。なお、未提示項目については、項目母数推定時の尤度関数には含まれない。さらに、多母集団モデルを用いた計算ができるソフトウェア（BILOG-MG（Zimowski, Muraki, Mislevy, & Bock 1996）や EasyEstimation（熊谷 2009））などでは、特定のテストＸ（もしくはテストＹ）を受検した集団を基準集団（その集団分布の平均を 0、標準偏差を 1 とする）として分析を行うことができる。単一集団での分析の場合には、データ行列に含まれる集団全体の母集団分布について、平均を 0、標準偏差を 1 とすることとなる。

（３）項目固定法

項目固定法（fixed items method）は、等化したいテストの項目母数推定時に、すでに得られているテストデータから推定した項目母数を既知として利用し、等化を行う方法である。例えば、アンカーテストデザイン（共通項目デザイン）における項目固定法の手続きは以下のようになる。

１）テストＸの項目母数を推定する。

２）テストＹの項目母数を推定する際に、共通項目部については１）で得られた項目母数に「固定」して推定する。このとき、母集団分布の平均を 0、標準偏差を 1 とするような制限は必要としない。

以上の手続きにより、テストＹの項目母数は全てテストＸ上の尺度で表現されることとなる。この計算手続きにおいても、ソフトウェアとしては BILOG-MG や EasyEstimation を用いることができる。特に EasyEstimation では、項目固定を行う手続きについてマウス操作を前提とした GUI により、簡便に分析を行うことが可能である。

2.6.6 基準尺度の構成

この節では平成 28 年度基準の尺度を構成する際の等化係数の推定および等化法の選択について行った手続の概略を記録しておく。

1）平成 25 年度（H25）と平成 28 年度（H28）の年度間共通項目について、推定された項目母数をプロットし、等化可能性について検討を行った。(35)式にあるように、等化係数を用いる方法では母数間の線形的な関係が想定されるが、各年度の項目困難度の推定値をそれぞれ縦軸と横軸に配してプロットするとほぼ直線上に並ぶため、全共通項目を使用してそのまま等化を行って問題ないと判断した。

2）年度間共通項目の項目母数推定値を用いて、 Mean-Mean 法、Mean -GMean 法 （Mean-Mean 法において、項目識別力の幾何平均を用いる方法)、 Mean-Sigma 法、Haebara 法、Stocking-Lord 法の 5 つの方法 （以下、それぞれ MM、MGM、MS、HB、SL と略記する）で H25 の能力母数の尺度を H28 の能力母数の尺度に等化するための等化係数を求めた。

3）MGM 法以外の推定には R のパッケージである plink ver.1.3.1 を使用した。別のプログラム STUIRT ver.1.0 を用いて、結果が許容範囲内で一致することを確認した。MGM 法については、別に R 関数を作成しそれを利用した。こちらは、独立に 2 名の分析者がプログラムを書き、結果が一致することを確認した。HB 法と SL 法については、以下の条件を設定して推定を実行した：

条件 I：項目特性関数・テスト特性関数にかけるウェイトは、各年度・各教科の項目母数推定時の BILOG のアウトプット(*.PH2 ファイル）の値を使用（求積点の数＝41）する。

条件 II：目的関数については、等化の対称性を考慮し、標準化は行わない。

4）共通項目について、H28 の項目母数推定値と、等化後の H25 の項目母数推定値をプロットして等化方法による違いを比較検討した。また、各等化法について、Leave-One-Out 交差妥当

化（LOOCV）を行い、共通項目を 1 つだけ除外した場合の等化係数を求めてこれを各共通項目について繰り返し、その違いを記録した。等化係数のばらつき（標準偏差）が小さいほど、データとして与えられた特定の項目母数推定値の影響を受けにくい、頑健な推定ができていることの証拠と捉えることができる。

5）各教科ごとの検討については以下の通りであった。

国語：いずれの方法でも、(35)式の傾き（K）は 1 より小さく、切片（L）はマイナスの値となった。MM、MGM、MS では K、L の値とも大きめ、HB および SL では相対的に小さめとなった。項目母数推定値のプロットでは、いずれの方法でも目立った違いは見られなかった。LOOCV によれば、K、L ともに最も頑健であったのは HB、最も頑健でなかったのは MM であった。

算数：MM および MGM では傾き（K）が 1 を超え、その他の方法では 1 より小さくなった。切片（L）に関しては、MM および MGM がプラスに大きく、それ以外の方法ではプラスであるものの MM および MGM よりは小さな値となった。項目母数推定値のプロットでは、いずれも方法でも目立った違いは見られなかったが、項目困難度推定値のプロットより、等化係数の傾きは 1 より小さくなるのが適切と推察されたため、MM および MGM の結果を採用するのは無理があると考えられた。LOOCV によれば、K、L ともに最も頑健であったのは MGM、最も頑健でなかったのは MS であった。

6）総合的判断

上記教科別の結果より、

・国語と算数では最も頑健であった／なかった等化法が異なる、

・両教科とも、最も頑健であった方法が、5 つの方法の中で最も極端な等化係数推定値を与えた（国語では HB、算数では MGM）、

となった。一方で、教科や、（将来的に）年度を超えて同じ等化法を使うことを考えると、異なる状況に対応できる汎用性も重要な観点となる。以上より、

・今回の調査データに関しては、いずれの教科においても、他の方法と比較して中庸な等化係数推定値および LOOCV の結果を与えた、

・多くの使用実績がある（米国では、最も標準的な等化法として用いられている）、

SL 法による結果を採用するのが妥当と判断した。

7）ただし、これは今回の分析結果に基づく判断であり、今後同様の調査を行う場合には、各等化法のパフォーマンスを都度比較し、SL 法を継続的に使用していくのが適切かどうかを慎重に判断していく必要がある。中学校国語・数学においても同様。

8）令和 3 年度に関しては、総合的に判断しすべての科目において、平成 28 年度項目母数を固定しての項目固定法を採用した。推定の際のソフトウェアは EasyEstimation を利用した。

2.7　重複テスト分冊法の導入

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（item pool）とよぶ）、それらをいくつかのユニット（PISA の呼称は cluster[1]）にまとめ、さらにそのユニ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）あるいはアイテム・マトリックス・サンプリング（item-matrix-sampling、以下「重複テスト分冊法」[2]）と呼ぶ。

しかしながら、分冊を組み上げる際に、各分冊におけるユニ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。例えばあるユニ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、そのユニ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）どのユニ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）ユニ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）冊子の中でのユニ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、

---

[1] PISA ではひとつの分冊の中にリーディング・リテラシー、数学的リテラシーや質問紙項目などを組み込んでいるため、それらをクラスター（cluster）と呼ぶが、経年変化分析調査では分冊の中には同一教科の問題しか含まれていないため問題（項目）のまとまりと言う意味で「ユニット」と呼ぶことにした。

[2] 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。

などが要求される。最後の d）については、ユニットの中に含まれる項目数をすべてのユニットで同一にしておくことで、各冊子に含まれるユニ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。

2.7.1 釣合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bk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。

例えばブロック計画 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、

（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($\lambda = 1$)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、経年変化分析調査では、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。

2.7.2 ユーデン方格

ユーデン方格[3]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、$v \times v$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。例えば、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、例えば、第1行から第3行を取り出せば、

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

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

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

が得られるが、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4のような7 × 7のラテン方格を考える。例えば日照時間を列、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

[3] 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 et al. 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。

の耕地を、例えば、表 4 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 4：7×7 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$\lambda = 2$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 つの行に関して 1 回ずつとなっていることもわかる。

表 5：4×7 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さきの 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 6：3 × 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。経年変化分析調査の開発研究においては、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、それにあわせて経年変化分析調査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、処理をユニット、ブロックを分冊、ブロックの大きさにたいして$1,2,\cdots,k$のように順序をつけ、それを位置（一つの分冊の中での出現順序）と呼ぶこととする。なお、ユニットのことを 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 はユニット数と分冊数および分冊ごとにユニ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、例えば、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 ユニット、13 分冊、4 位置、 1 会合数や、本研究で使用した 7 ユニット、7 分冊、3 位置、 1 会合数、または 7 ユニット、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、例えば、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。

経年変化分析調査では国語と算数/数学に関しては PISA と同一の BIBD を採用した。表 7 に示す分冊番号、位置番号以外の数値がユニット（PISA ではクラスター）を表す番号となる。

表 7：経年変化分析調査で採用されている 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 7 に示されている BIB(13,13,4)デザインでは 13 個のユニットは 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)、ユニ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、いずれのユニットも 4 つの位置のどこかに出現し(条件 c)、いずれの分冊にも同数のユニットが含まれていることがわかる(条件 d)。

また、PISA2006 では表 7 のデザインに基づき、13 個のユニットに科学的リテラシーのためのユニットを 7 個（表 7 のユニット番号では 1〜7 に該当する)、数学的リテラシーのためのユニットを 4 個（表 7 では 8〜11)、リーディング・リテラシーのためのユニットを 2 個（表 7 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、同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こと、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 1 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、例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まれること、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ユニット数を必然的に減らさざるを得ない。一連の開発研究からその大問数は 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。そのため、国語に関しても 7 分冊 7 ユニット 3 位置の BIBD を開発研究では試みた。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 1 校時で実施することが可能となった。経年変化分析調査の国語／算数・数学で用いられている 13 分冊方式の場合であると、ひとりの児童生徒に全問解答してもらうとすれば 6.5 校時かかる計算になる。

なお、開発研究の中で試みられた、数学と国語のそれぞれのユーデン方格の実際は表 8、9 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 10 は国語で採用したユーデン方格（表 9）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、表頭にユニットの番号、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 8：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 9：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 10：BIB(7,3,1)デザインのフィッシャーによる表現（Fisher's representation）

		ユニット番号						
		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

2.7.3 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。例えば、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 にユニット 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、これはユニット 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、Cの転置行列をC'とし、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ユニ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ユニットの数を示しており、いずれの分冊についても 3 つのユニ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（$\lambda = 1$）。

次に、$C'C$を求めると、

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

が得られる。この場合、対角要素の数は各ユニ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、これはいずれもユニットもペアとなる回数は 1 回であることを示している。言い換えれば、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ユニ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ユニットの数を示しており、いずれ

の分冊についても 4 つのユニ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$\lambda = 2$となることがわかる。

一方、$D'D$を求めると、やはり、

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

が得られる。この場合、対角要素の数は各ユニ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、これはいずれもユニットもペアとなる回数は 2 回であることを示している。言い換えれば、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 つの分冊の中のユニット数を 4 から 3 に減らせる分、実施時間も例えば 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、例えば 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、共通のユニットを含まない分冊が生じることになる。

3　経年変化分析調査の標本抽出方法

経年変化分析調査の標本抽出法は令和 3 年度から同時に実施された保護者調査と同じ方法を採用している。概略は以下のとおりである。詳細は「別冊 （標本抽出方法)」を参照のこと。

3.1　抽出方法

学校を抽出単位とした層化集落抽出法によって選ばれた以下の学校数を対象とする。

小学校： 国語、算数それぞれ全国で 300 校（合計で全国から 600 校）

中学校： 国語、数学、英語それぞれ全国で 250 校（合計で全国から 750 校）

各層への学校数の割当は母集団学校数による比例割当とし、層内での学校の抽出は 1 番目は無作為に、そして 2 番目以降は通し番号に従って一定間隔で抽出を行う系統抽出法により抽出する。

3.2　層の構成方法

層を構成するための変数の候補は以下の 3 変数とする。

都市規模： 学校の所在地の都市規模

指定都市、中核市、人口 10 万以上、人口 10 万未満の 4 区分とした。

学校規模： 調査対象学年の学級数による学校規模

小学校では 2 学級未満（小)、2 学級（中)、3 学級以上（大）の 3 区分とする。

中学校では 4 学級未満（小)、4 ～ 5 学級（中)、6 学級以上（大）の 3 区分とする。

学力層： 平成 31（令和元）年度全国学力・学習状況調査結果に基づく学力層

小学校は国語と算数の学校別平均正答率を合計し、四分位数で平均正答率の合計が大きい方からA層、B層、C層、D層の 4 区分とする。

中学校は国語と数学と英語の学校別平均正答率を合計し、四分位数で平均正答率の合計が大きい方からA層、B層、C層、D層の 4 区分とする。

3.3　測定モデルにおける下位集団の取り扱い

なお、NAEP 等では上記の調査モデル（survey model）における下位集団（subpopulation）に対応する形で測定モデル（psychometric model）の方にもそれぞれの下位集団に対応させた事前分布を仮定している。これを潜在回帰モデル（latent regression model）あるいは能力回帰モデル（ability regression model）の θ への組み込みとよぶが、経年変化分析調査ではこの方式は採用していない。下位母集団の属性定義が不安定で困難なこと、等化作業の煩雑化やそれにともなう等化誤差の混入を避けるためである。将来的には例えば保護者調査との接合が実現すればそちらで収集する属性情報を利用することになるであろう。

4　調査の実際

4.1　データ収集デザイン

4.1.1 国語と算数・数学

本体調査データと経年変化分析調査データとの共通受検者集団を考慮したデータ構造は図 21 に示すとおりである。まず平成 25 年度データに着目すると、本体調査ブロックの全数データのうちに経年変化分析調査も受けた集団が存在する。それを図 21 では濃い網掛けで表現している。実際には標本抽出をしているが、わかりやすいように図中では標本抽出された集団をまとめて表現した。この集団は平成 25 年度経年変化分析調査の 2 分冊を受けている。どの分冊を受けるかは学校ごとに無作為で振り分けているため、平成 25 年度はデータ収集デザインとしては、同一母集団から抽出された統計的には等質な 2 つの標本集団が別々の冊子を受ける等価グループデザイン（equivalent groups design：EG）となっている。一方、平成 28 年度は本体調査ブロック側での抽出手続きは同じであるが、経年変化分析調査ブロック側での実施方式が BIBD に基づく重複テスト分冊方式となっている。さらに平成 25 年度と平成 28 年度の経年変化分析調査に着目すると、項目には両年度とも実施された項目が含まれている。これを共通項目とする。実際には各分冊ともそれらの項目が含まれているが、描画すると複雑になるため、わかりやすいように、図 21[4]では項目欄のみにその様子を描いている。

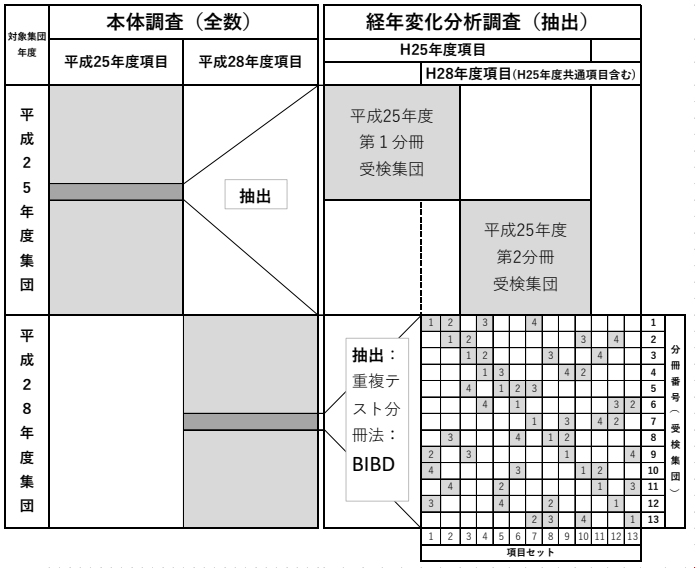


図 21：全国学力・学習状況調査におけるデータ収集デザインの概略図

[4] BIBD の表についてグレーに塗りつぶされているマスは、それに対応する列番号のユニットが、対応する分冊番号に含まれることを示す。なお、マス目内の番号は、その分冊内でのユニットの順番を示す。

図 21 に示されている H28 年度経年変化分析調査の BIBD はフィッシャーの行列表記となっている。同じものを分冊デザインで表現したものが図 22 となる。

		分　冊　番　号												
		分冊1	分冊2	分冊3	分冊4	分冊5	分冊6	分冊7	分冊8	分冊9	分冊10	分冊11	分冊12	分冊13
分冊内の位置	1番目	1	2	3	4	5	6	7	8	9	10	11	12	13
	2番目	2	3	4	10	6	13	12	9	1	11	5	8	7
	3番目	4	10	11	5	7	12	9	2	3	6	13	1	8
	4番目	7	12	8	9	3	4	11	6	13	1	2	5	10

図 22：国語、算数・数学における分冊デザイン

(注)表中の番号は、ユニット番号を表す

4.1.2 英語

英語については、「聞くこと」、「読むこと」、「書くこと」、「話すこと」の 4 技能での調査を実施した。その解答方式は、「聞くこと」、「読むこと」、「書くこと」の 3 技能調査は筆記方式、「話すこと」調査は PC・タブレット等の ICT 機器を活用した音声録音方式となる。「話すこと」調査がパフォーマンスアセスメントとなることから、実施コストがかかり、13 分冊からなる重複テスト分冊法は採用できない。そのため、国語/算数・数学とは異なり、データ収集デザインとしてはなるべく等化や標本抽出の精度を保つ一方、コスト削減のため、等価グループデザイン（EG）とアンカーテストデザイン（AT）とを組み合わせた等化デザインを採用した。

具体的には、R3 年度実施のデザインを例にとると、同一母集団から標本抽出した 2 つの標本集団に対して、互いに共通の問題（アンカーテスト：A）が含まれる 2 つの異なる分冊を、各学校には、それぞれの分冊を受ける生徒が約半数ずつになるように配付した。「聞くこと」、「読むこと」、「書くこと」のいわゆる 3 技能の調査と「話すこと」調査の分冊（機器を使用したため、正確にはプログラム）は同じ分冊番号になるように実施する。具体的な分冊デザインは以下の通りである（図 23）。

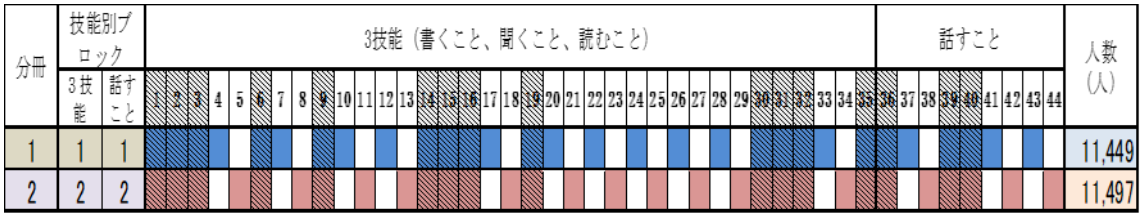


図 23：英語調査の分冊デザイン

（注）青は分冊１に入れられている項目番号を示す
赤は分冊２に入れられている項目番号を示す
網掛けは分冊１と分冊２との共通項目である

4.1.3 分冊配付方法の具体

各分冊の配付方法は国語/算数・数学を例にとると、小学校、中学校とも下記の要領で、学級毎に冊子を梱包し、分冊毎の学校、学級、児童生徒数の偏りがないように配付した。

1）まず、任意の最初の学校をＡ学校（２学級、１学級あたりの人数 21 人）とする。
2）そのＡ学校の最初の学級（クラス１）に対して、13 分冊を１分冊ずつ繰り返し、
　　分冊１、分冊２、･･･、分冊 13、分冊１、分冊２、･･･、分冊８
　として学級 21 人分の冊子を順次並べていき、その学級に割り当て多分を最後に梱包しまとめる。この学級は 21 人なので最後は冊子番号８の分冊となる。
3）次の学級（クラス２）に移り、分冊９から開始して、
　　分冊９、分冊 10、･･･、分冊 13、分冊１、分冊２、･･･、分冊 13、
　　分冊１、分冊２、分冊３
　のように同じく各分冊をならべていく。分冊番号９から開始し 21 人学級のため最後は分冊３となる。
4）さらに２番目の学校であるＢ学校（３学級、１学級あたりの人数 30 人）にうつり、
　　クラス１：分冊４、分冊５、･･･、分冊 13、分冊１、･･･、分冊 13、分冊１、
　　･･･、分冊７
　　クラス２：分冊８、分冊９、･･･、分冊 13、分冊１、･･･、分冊 13、分冊１、
　　･･･、分冊 11
　　クラス３：分冊 12、分冊 13、分冊１、･･･、分冊 13、分冊１、･･･、分冊 13、
　　分冊１、分冊２
　と順次並べていき、クラスごとに梱包する。
5）上記と同じ手順で、以下、抽出校数分を繰り返す。

なお、上記の作業は現在は手作業に頼らざるを得ないため、時間・コストがかかる。将来的には、経年変化分析調査を含む全国学力・学習状況調査の CBT 化が進むことによるこの作業に関する効率化や経費削減が期待されている。

4.2　尺度構成

令和 3 年度経年変化分析調査では、IRT に基づき、平成 25 年度及び平成 28 年度調査の実施データから、すべての項目（問題）の項目母数（困難度や識別力）を平成 28 年度基準の測定尺度上に等化し（測定尺度の構成：尺度の原点は 0、単位は 1)、この尺度に基づいて令和 3 年度調査の学力の測定を行った。このことにより、もし学力分布に何らかの変化があれば、平成 28 年度の学力分布を基準とした変化であると判断できる。なお、令和 3 年度新出の問題もあるが、調査実施後、あらためて新出問題を平成 28 年度尺度上に等化した。具体的には IRT の分析ソフトウェアとして EasyEstimation2.1.6 を利用し、平成 28 年度既出の項目母数を固定した項目固定法を用いて、新出問題の項目母数は周辺最尤法により推定した。

4.3　項目母数の推定結果

項目母数の推定結果は以下に示すとおりである。下記の図 24 は小学校の国語と算数に関する項目母数をプロットしたものである。X 軸には項目困難度をとり、Y 軸には項目識別力をとっている。尺度は学力推定値にあわせて原点を 0、単位を 1 とした。上が散布図、下が同じ散布図であるが、各プロット点から X 軸の方向へ垂直に下ろした垂線を描いた垂線プロットとなっている。垂線の長さが高いほど識別力が高く、垂線の位置が困難度に対応している。なお、わかりやすいように下図の Y 軸は上図の Y 軸に対して縮小して描いている。図 25、図 26 も同様である。

学力推定値の範囲は標準正規分布でおよそ－4.0 から 4.0 を想定しているが、いずれの学年教科についても、困難度が低い方に項目が分布していることがわかる。これは高い学力層の識別が中位の学力層に対する識別よりも精度が相対的に落ちることを意味している。分冊に含まれる項目を差し替え、困難度の高い項目をどの学年・教科についても少数個増やすことで尺度全体の測定精度はさらに向上することを意味している。

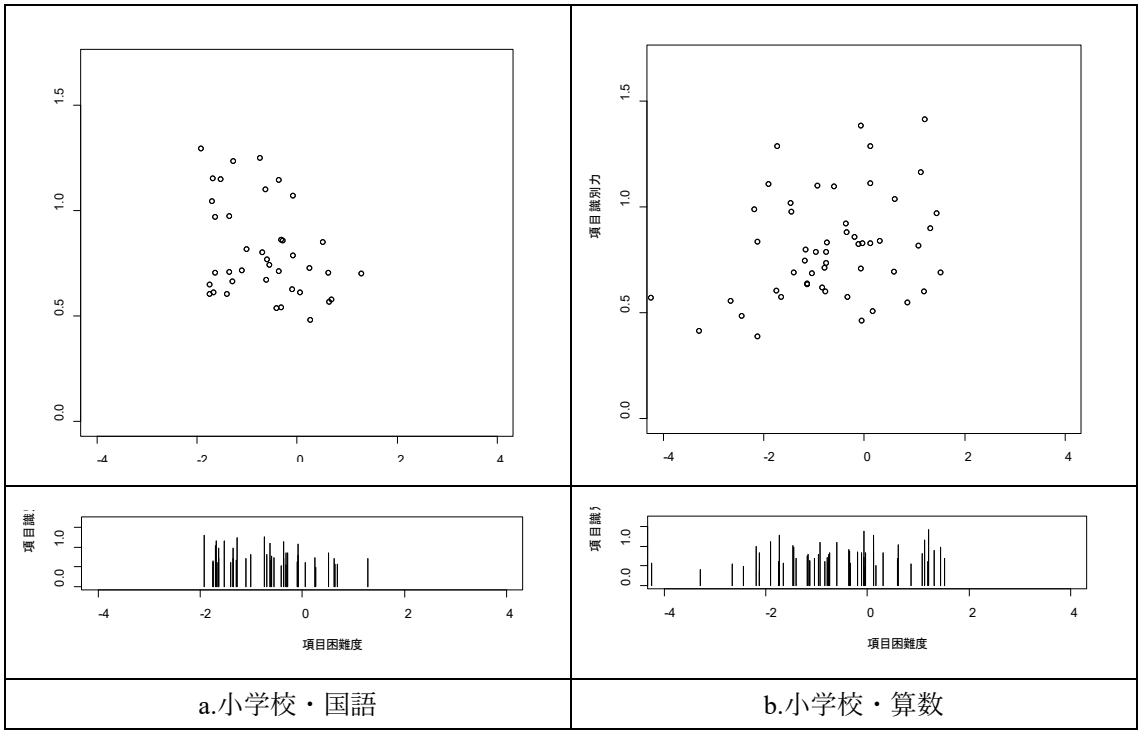


図 24：項目母数の散布図（小学校：国語・算数）

(注)上図：散布図・下図：垂線プロット・X 軸：項目困難度・Y 軸：項目識別力

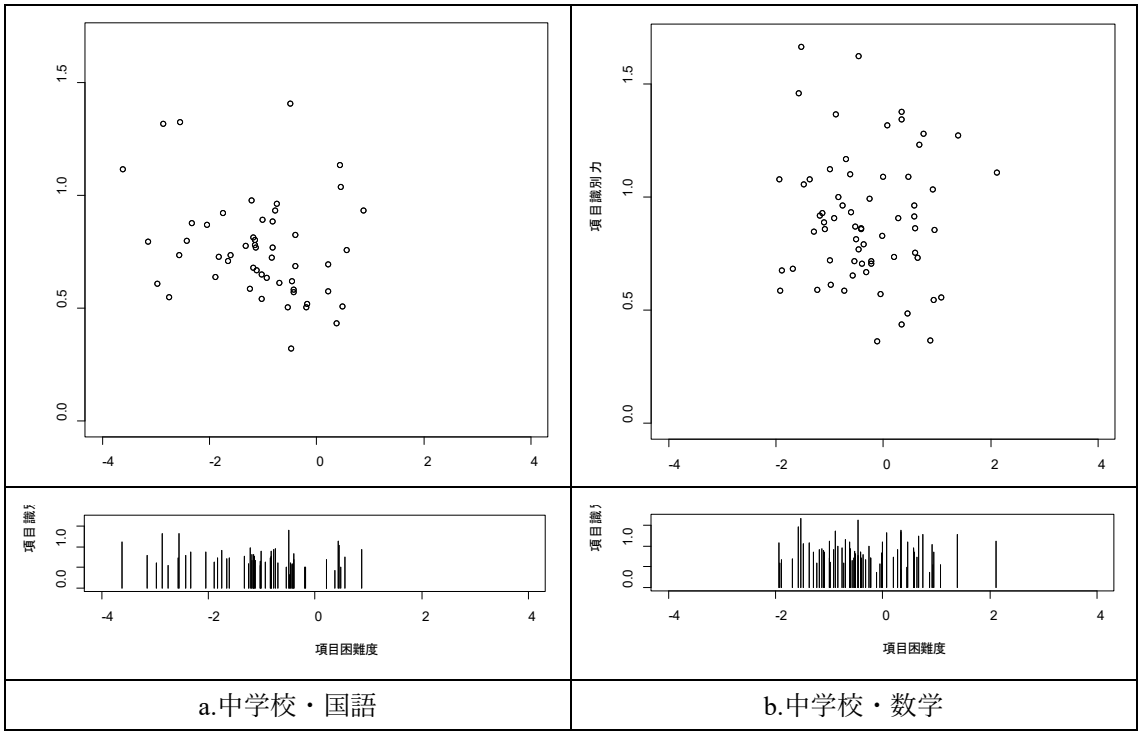


図 25：項目母数の散布図（中学校：国語・数学）

(注)上図：散布図・下図：垂線プロット・X 軸：項目困難度・Y 軸：項目識別力

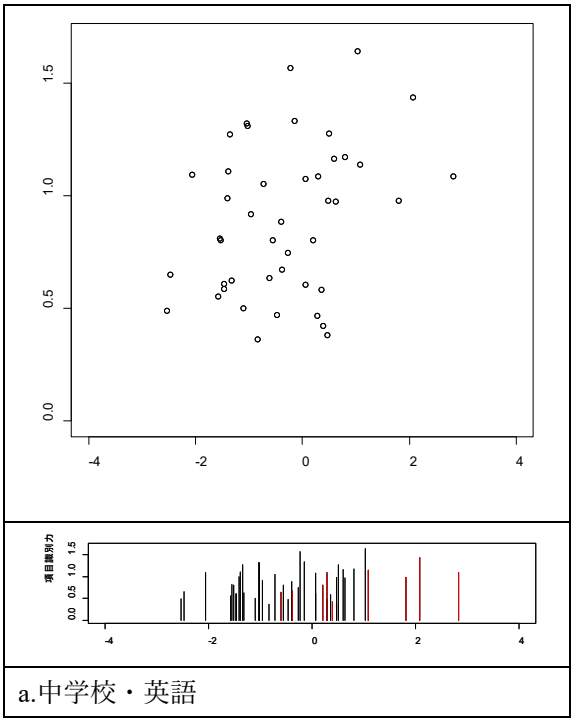


図 26：項目母数の散布図（中学校：英語）

(注)上図：散布図・下図：垂線プロット・X 軸：項目困難度・Y 軸：項目識別力（赤は「話すこと」）

4.4　経年変化分析調査と本体調査の精度

経年変化分析調査では重複テスト分冊法を採用しているため、調査全体では項目数が本体調査に比べて多くなっている。そのため、経年変化分析調査の方が測定精度を高くできる。そのことを示したのが下の図 27 である。青が令和 3 年度の経年変化分析調査、赤が平成 21 年度の本体調査である。学年・教科はいずれも中学校・数学である。一番上の図がテスト情報量曲線、次が測定誤差、その下 2 つが項目識別力と困難度の垂線プロットである。設計通り経年変化分析調査の方が全域にわたって精度良く測定できていることが確認できる。

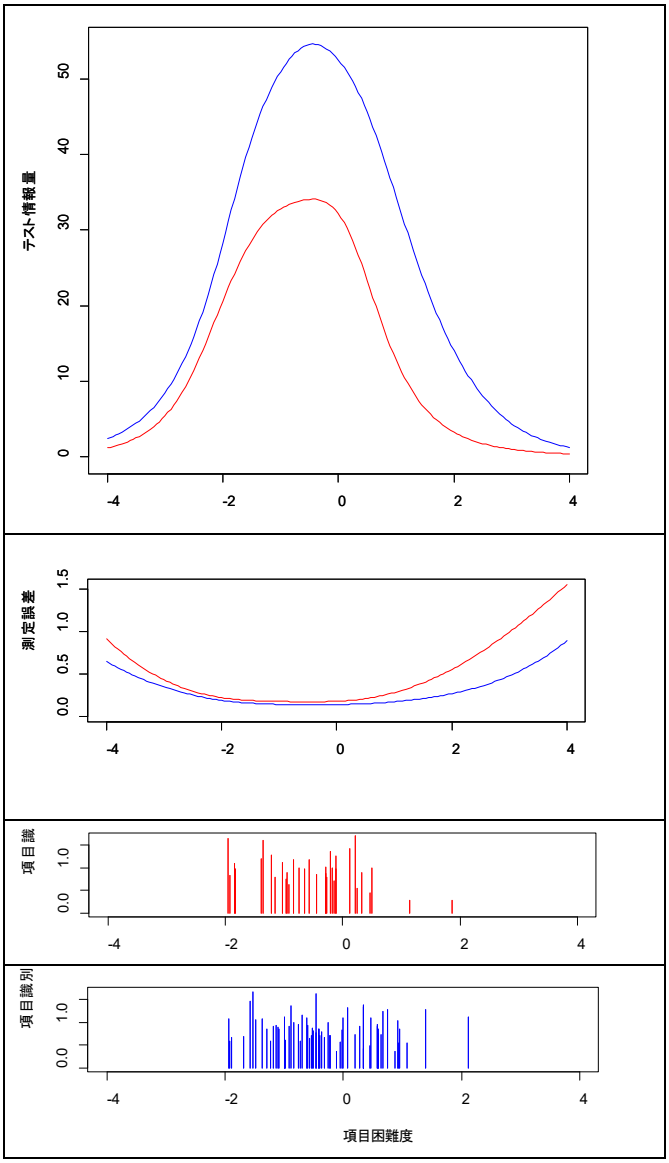


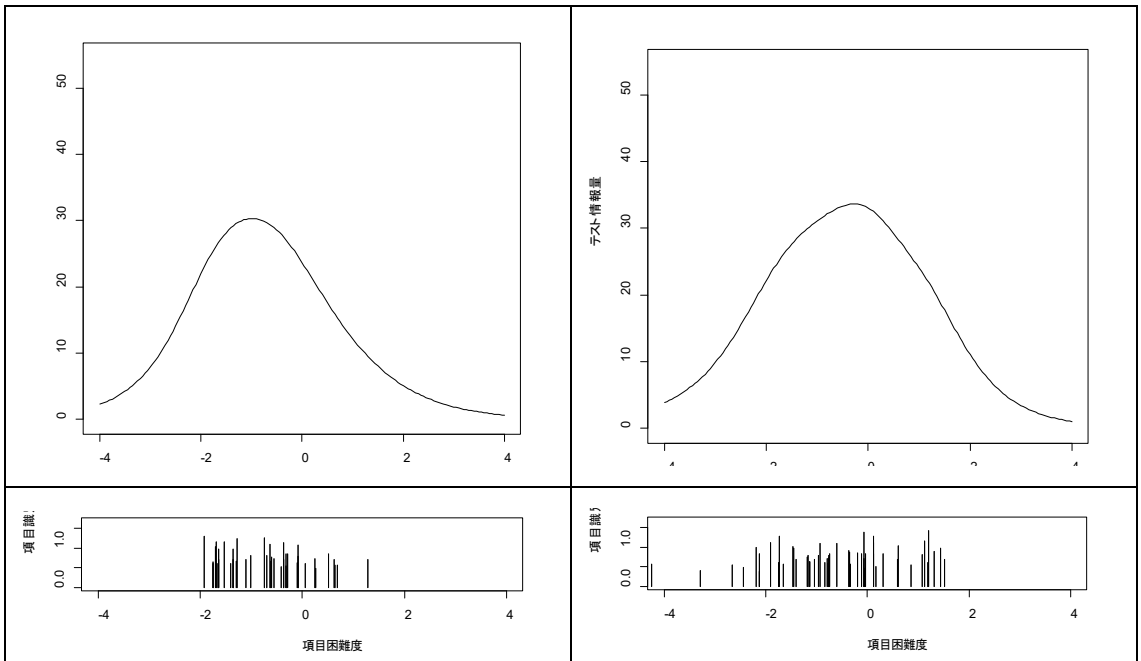
図 27：経年変化分析調査と本体調査の精度

（注）経年変化分析調査が青、本体調査が赤

4.5　各学年・各教科のテスト情報量曲線と項目母数

以下に各学年・各教科のテスト情報量曲線のグラフを図 28 に示す。いずれも上の図がテスト情報量曲線のグラフ、下図が項目母数の垂線プロットである。上下図の X 軸は共通で学力推定値と項目困難度が表現される尺度をあらわす。原点は 0、単位は 1 である。

学年・教科の 5 つの図を比較すると中学・数学のテスト情報量が全域にわたって測定精度が高いことがわかる。中学校・英語は受検にも採点にも時間と労力がかかるパフォーマンスアセスメントである「話すこと」調査のため、2 分冊しか準備できず、結果として調査全体で項目数が相対的に少なくなったことにより、テスト情報量自体は中学・数学に比較して相対的に低くなっている。しかし、困難度自体は原点を中心にある程度広がっているため、全体的に精度良く測定できていることがわかる。

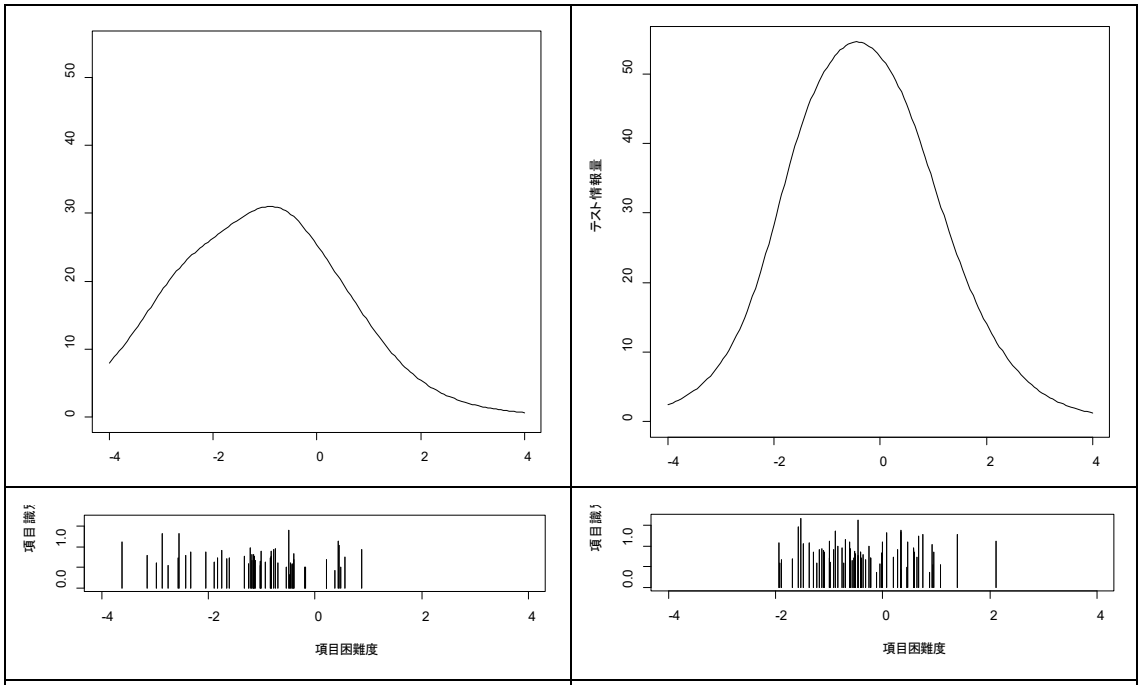


a.小学校・国語　　b.小学校・算数

図 28：テスト情報量と項目母数（小学校：国語・算数）

(注)上図：テスト情報量曲線・下図：垂線プロット・X 軸（共通）：項目困難度

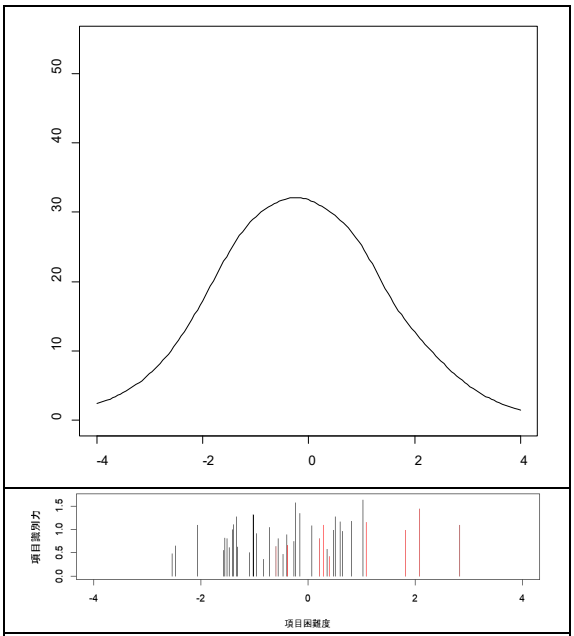
一方、国語に関しては小学校・中学校ともにやや項目困難度が低い項目が多かったためそれに対応する学力層に対してはテスト情報量が高くなっているが、高い学力層のところでテスト情報量が低い傾向にある。国語に限らず、いずれの尺度についても、学力が高い層に対する困難度が高めの項目を少数個増やすことで集団統計量の推測を目的とする経年変化分析調査の精度をさらに向上させることができるであろう（図 29、図 30）。



a.中学校・国語　　b.中学校・数学

図 29：テスト情報量と項目母数（中学校：国語・数学）

(注)上図：テスト情報量曲線・下図：垂線プロット・X 軸（共通）：項目困難度



a.中学校・英語

図 30：テスト情報量と項目母数（中学校：英語）

(注)上図：テスト情報量曲線・下図：垂線プロット・X 軸（共通）：項目困難度・赤：話すこと技能

4.6　学力推定値と学力スコア

上記の手続によりデータを収集し、各項目の特性（困難度や識別力：項目母数）を推定・確定した後、全児童生徒の学力推定値は EasyEstimation2.1.6 を利用し、最尤法によって求めた。求めた最尤推定値はわかりやすいように PISA の尺度にあわせて、基準尺度の原点が 500、単位が 100 となるように線形変換し、これを学力スコアとして報告した。

下の図 31 の中の 3 つのグラフは上から順に中学校英語の学力スコアの相対度数分布、中央の図が学力スコアに対応する学力推定値（最尤推定値）の測定誤差、下の図が各項目の識別力を困難度から伸びる垂線で表した図である。垂線の位置が困難度、長さが識別力を表している。

学力スコアで 650、学力推定値で 1.5 以上の困難度を持つ 3 個の項目が存在する。それぞれの項目の識別力は相対的に高くはなっているが、項目数の少なさ故に、この辺りから測定誤差の値が増加し始めている。上の相対度数分布をみるとやはりその辺りから分布曲線が不安定になっている様子がうかがえる。

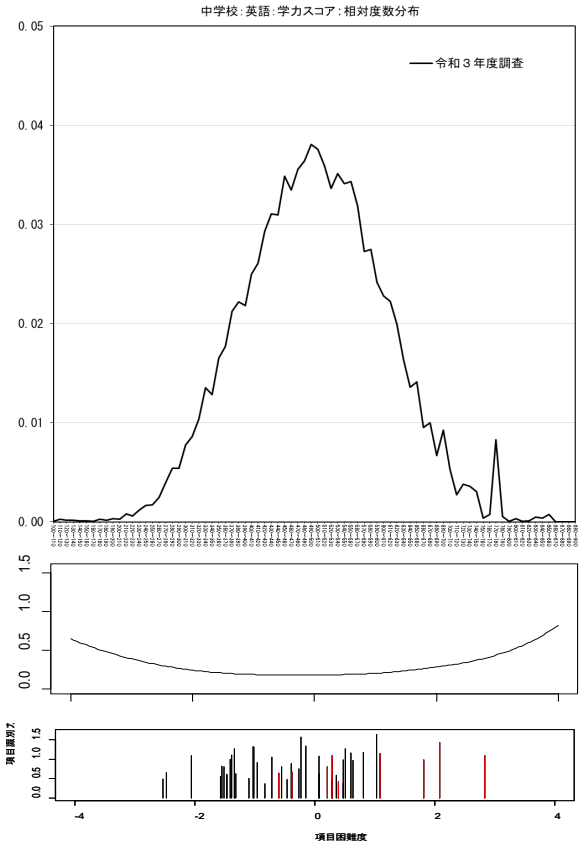


図 31：学力スコアの相対度数分布と測定誤差、項目母数（中学校・英語）赤：話すこと技能

5　令和３年度調査結果

5.1　結果概要

国語及び算数・数学の平成 25 年度、平成 28 年度、令和 3 年度調査の結果等は次ページ以降に示すとおりである。ただし、平成 25 年度は重複テスト分冊法ではない分冊デザインで実施したため参考程度にとどめ、平成 28 年度分布を基準に令和 3 年度の変化の有無をみる。また、PISA や NAEP 等の国際学力調査の知見を踏まえ、本調査においても全国の学力分布の状況の変化の有無は中長期的な観点から継続して分析を行っていく必要がある。

平成 28 年度と令和３年度の調査結果を分析すると、小学校・中学校ともに国語に関しては、学力スコア分布（累積相対度数分布）の状況は両年度間でほとんど変化は観察されなかった。換言すれば、国全体としてみれば、児童生徒の学力の低下や向上といった変化は認められなかった。

算数・数学については、令和３年度の学力スコア分布（累積相対度数分布）は基準である平成 28 年度の学力スコア分布の右側に（全体的にみて学力スコアが高い方へ）若干移動していることが観察できる。これについては、国全体でみれば、算数・数学について若干学力が向上しているとも解釈しうるが、上記のとおり、全国の学力分布の状況の変化の有無は中長期的に継続して分析する必要があることを踏まえ、次回（令和６年度予定）以降の結果もあわせて引き続き分析していくこととする。

なお、中学校英語については、令和３年度が初めての調査実施であるため、今回は経年比較はできない。

5.2　学力スコアの分布

5.2.1 小学校国語

表 11：学力スコアの標本統計量（小学校・国語）

実施年度	国語（学力スコアの標本統計量）					
	児童数	平均	標準偏差	25 パーセンタイル	中央値	75 パーセンタイル
平成２５年度	5,984	500.3	110.3	427.9	498.6	567.5
平成２８年度	11,122	505.8	123.7	426.1	503.4	582.0
令和３年度	16,321	505.8	120.8	428.4	504.3	581.8

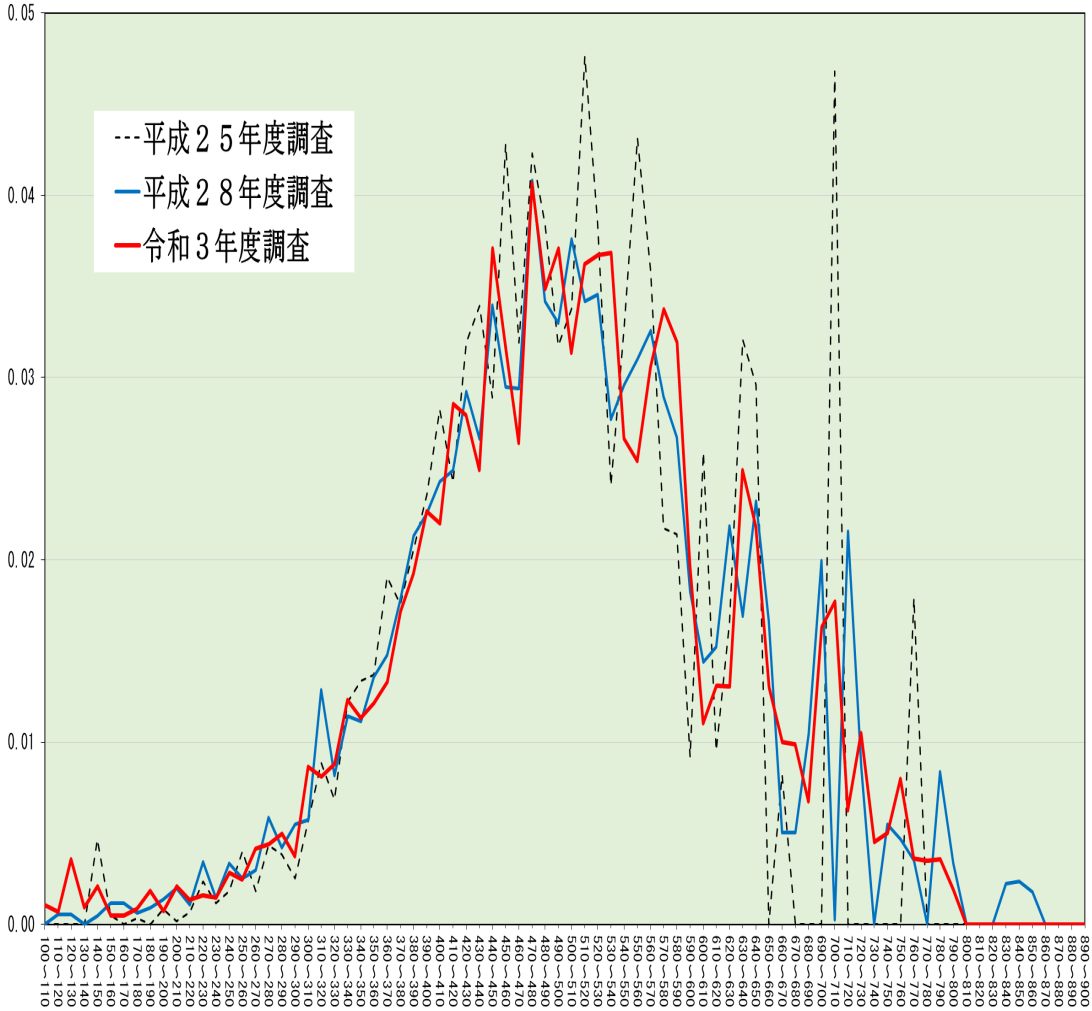


図 32：小学校：国語：学力スコア：相対度数分布

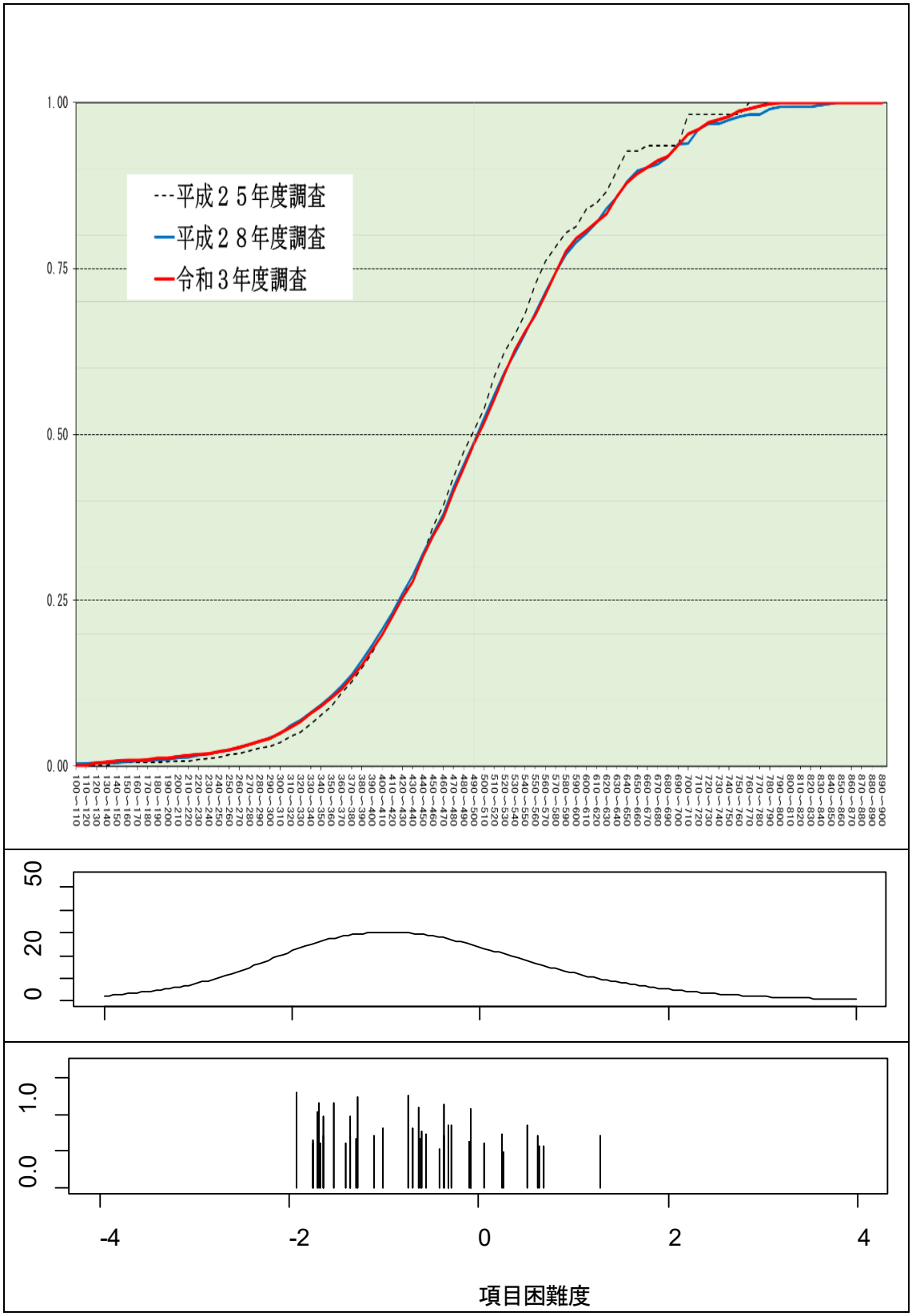


図 33：小学校：国語：学力スコア：累積相対度数分布とテスト情報量、項目母数

5.2.2 小学校算数

表 12：学力スコアの標本統計量（小学校・算数）

実施年度	算数（学力スコアの標本統計量）					
	児童数	平均	標準偏差	25 パーセンタイル	中央値	75 パーセンタイル
平成２５年度	5,952	512.0	117.3	436.3	506.4	586.0
平成２８年度	11,009	502.0	122.3	425.7	501.7	577.6
令和３年度	16,078	507.2	126.0	429.6	508.1	587.5

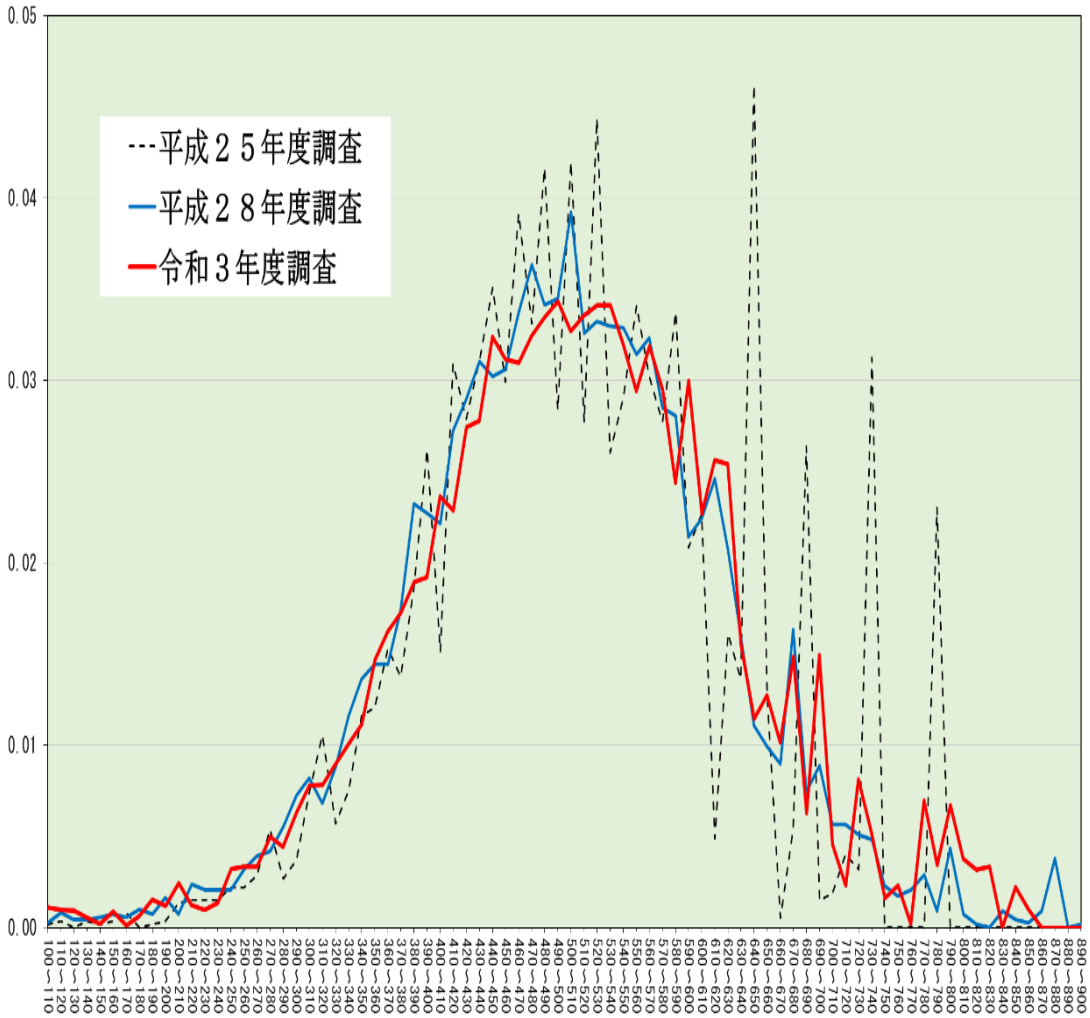


図 34：小学校：算数：学力スコア：相対度数分布

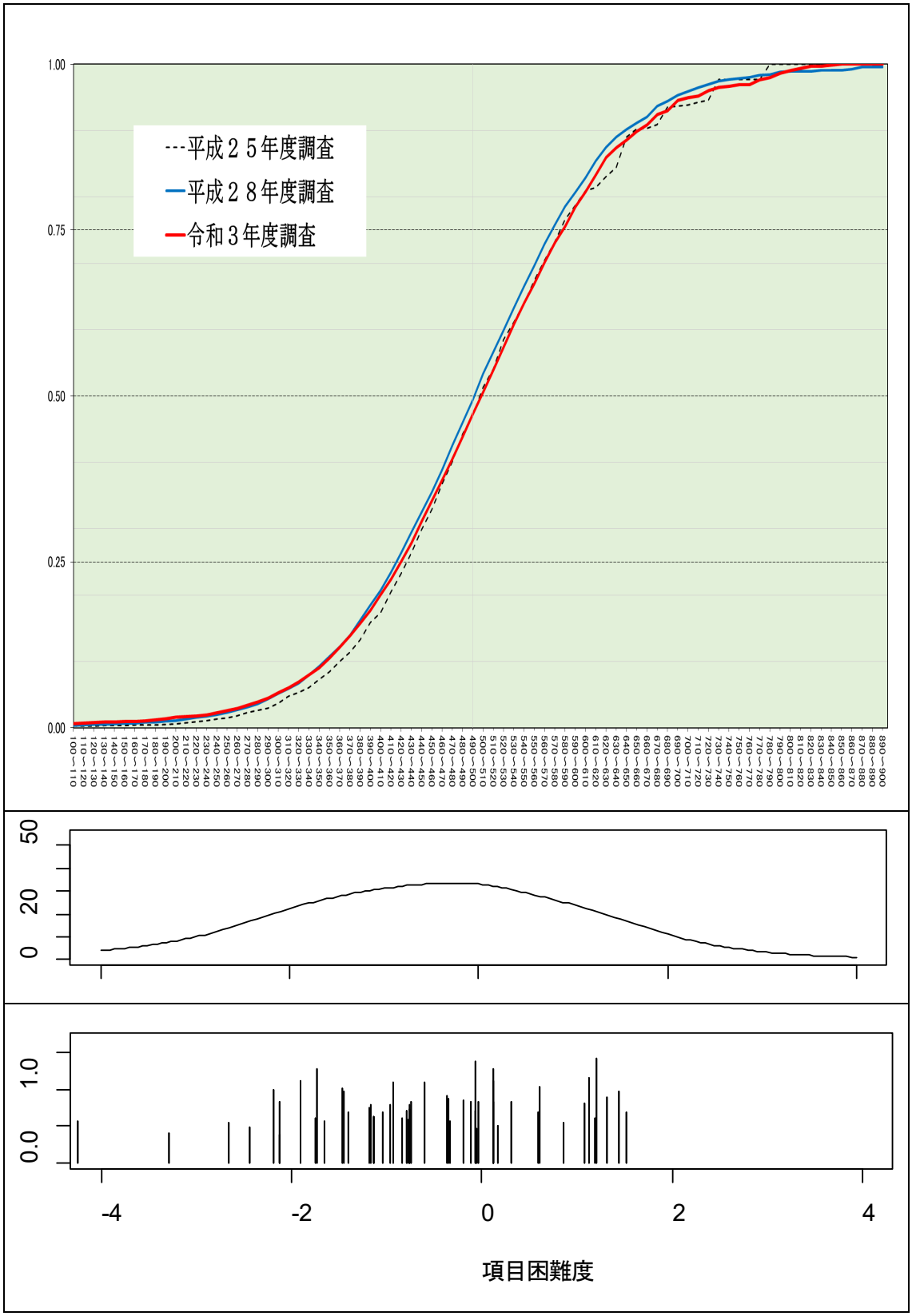


図 35：小学校：算数：学力スコア：累積相対度数分布とテスト情報量、項目母数

5.2.3 中学校国語

表 13：学力スコアの標本統計量（中学校・国語）

実施年度	国語（学力スコアの標本統計量）					
	生徒数	平均	標準偏差	25 パーセンタイル	中央値	75 パーセンタイル
平成２５年度	12,491	496.1	124.0	418.0	490.7	565.2
平成２８年度	27,029	508.6	128.0	429.0	503.5	580.7
令和３年度	25,206	511.7	121.3	434.6	510.0	585.3

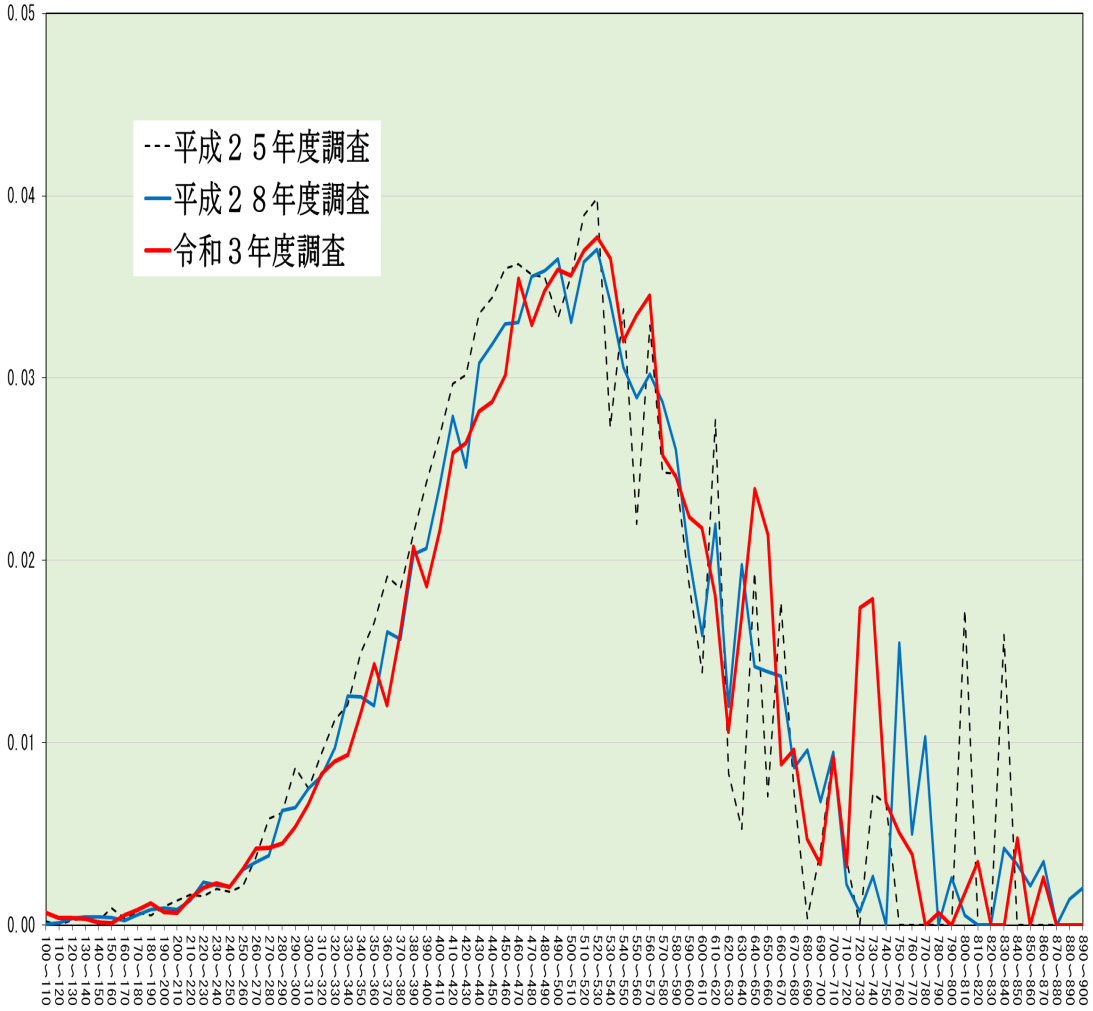


図 36：中学校：国語：学力スコア：相対度数分布

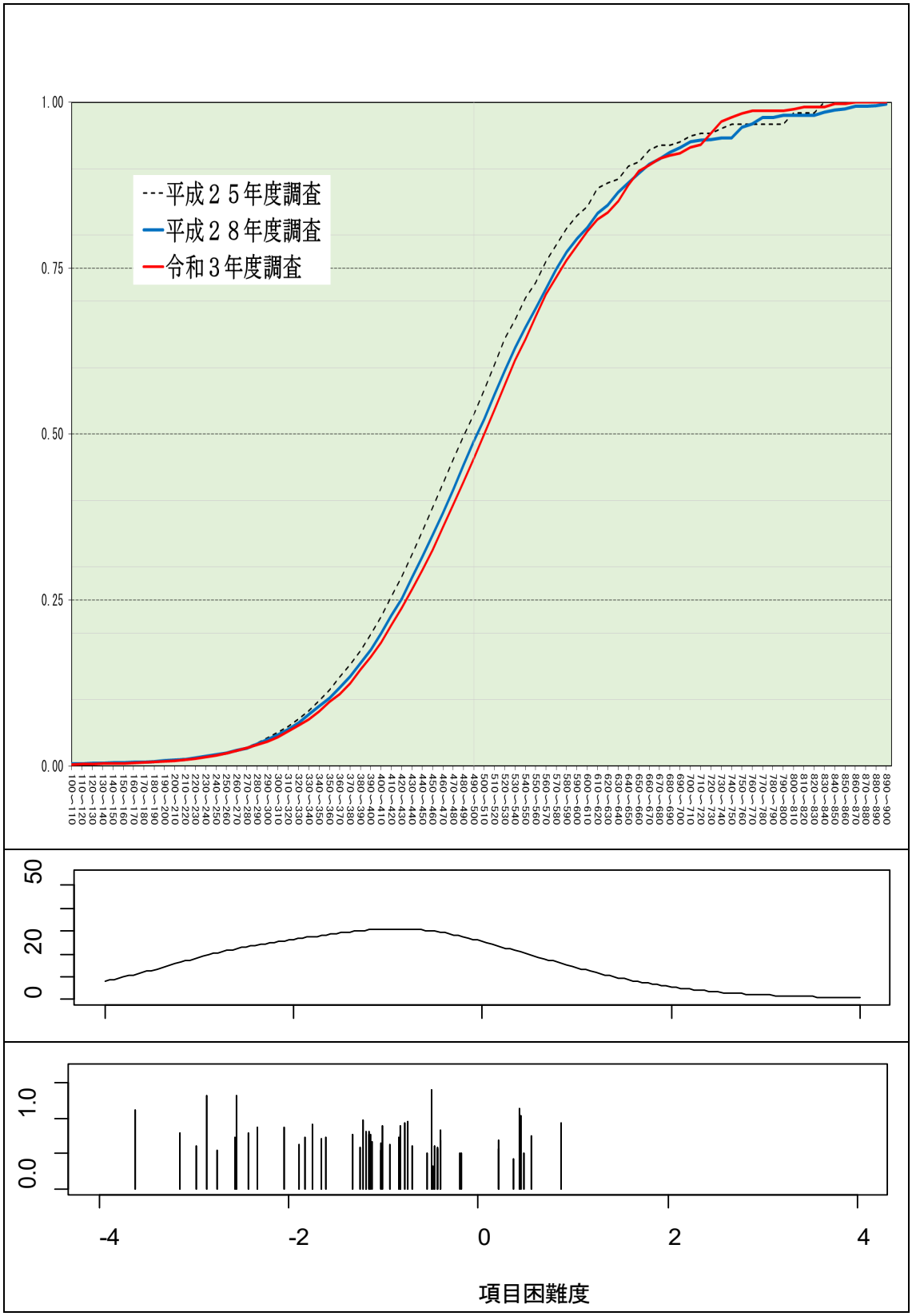


図 37：中学校：国語：学力スコア：累積相対度数分布とテスト情報量、項目母数

5.2.4 中学校数学

表 14：学力スコアの標本統計量（中学校・数学）

実施年度	数学（学力スコアの標本統計量）					
	生徒数	平均	標準偏差	25 パーセンタイル	中央値	75 パーセンタイル
平成２５年度	13,059	503.1	131.7	418.5	494.5	571.5
平成２８年度	26,493	502.0	116.6	425.6	500.3	576.3
令和３年度	25,145	511.0	118.2	431.8	512.1	588.9

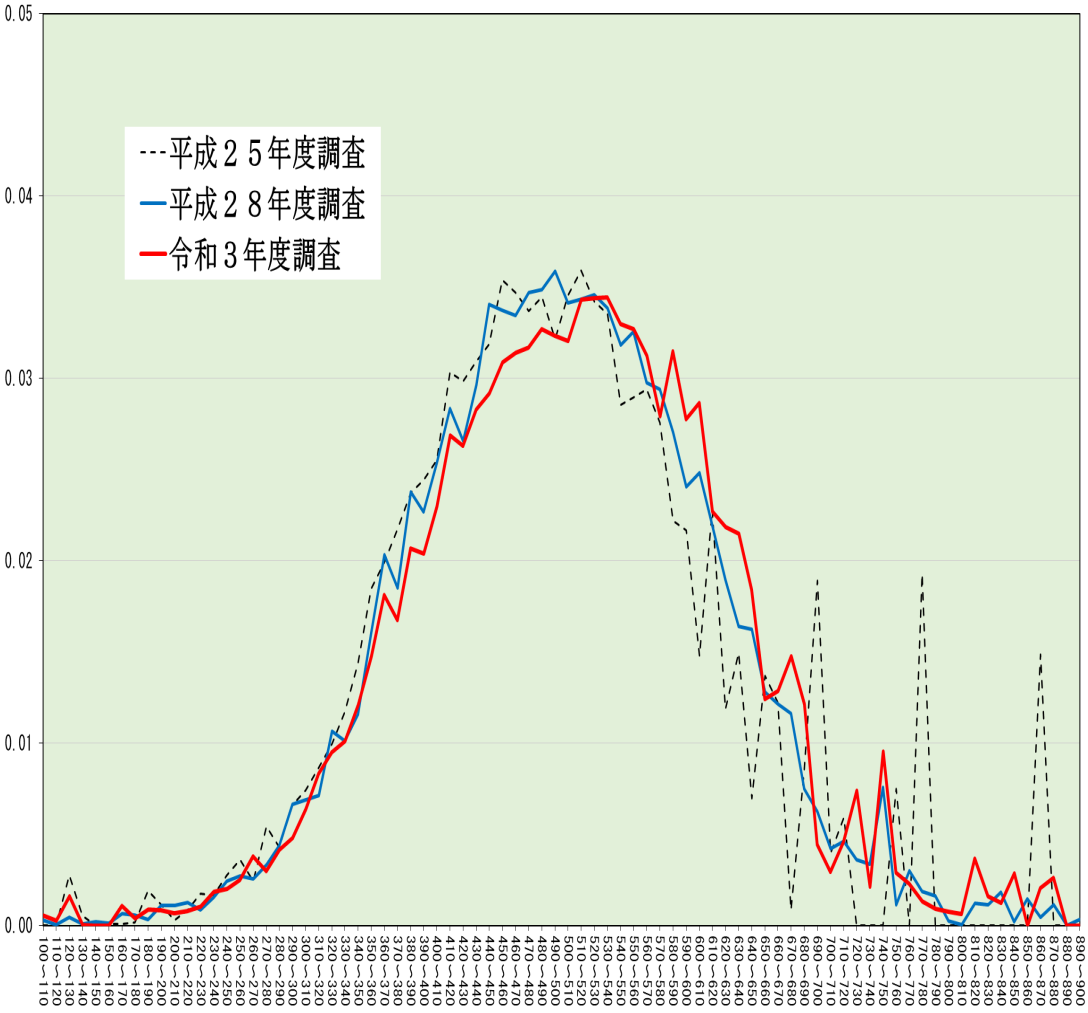


図 38：中学校：数学：学力スコア：相対度数分布

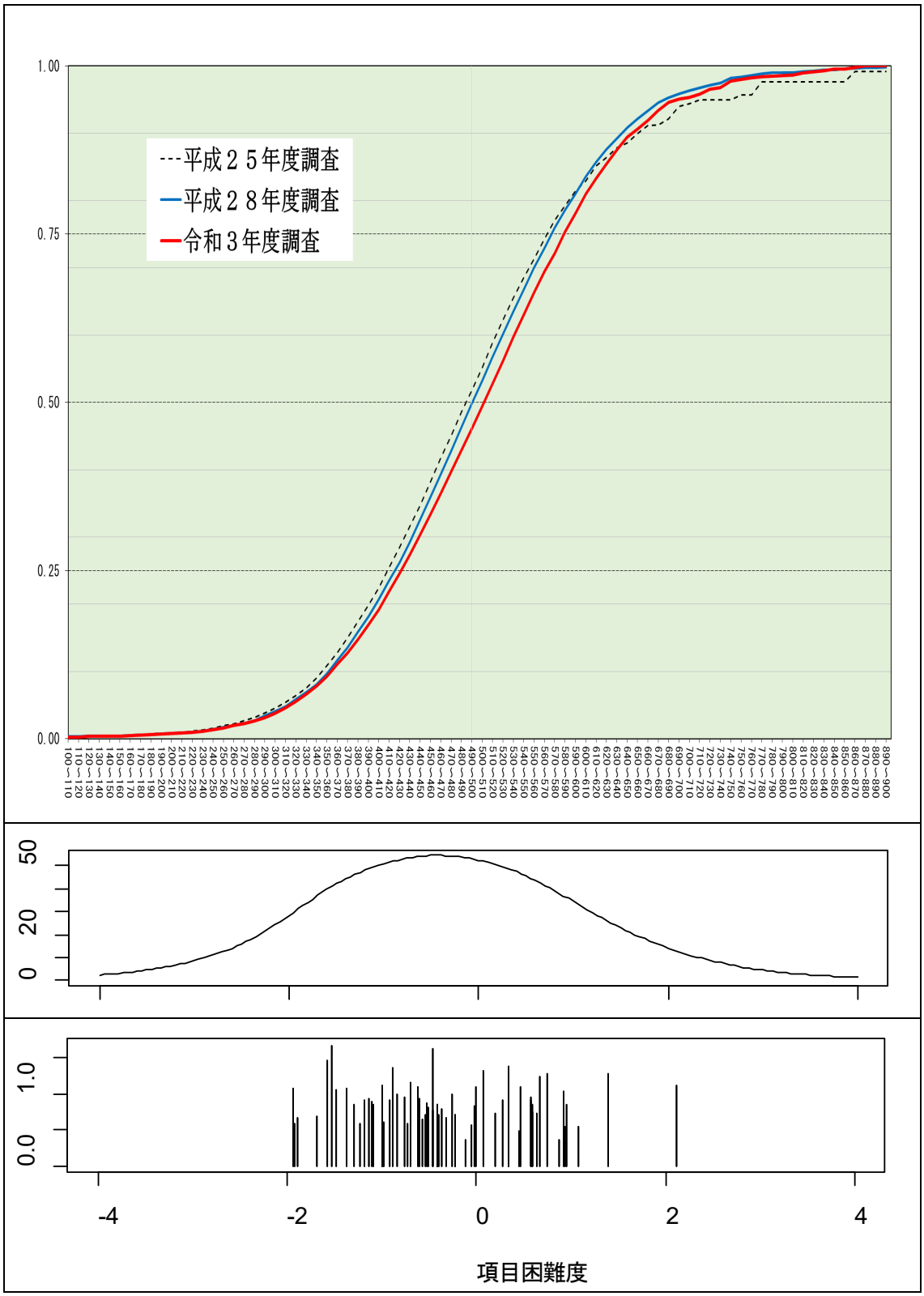


図 39：中学校：数学：学力スコア：累積相対度数分布とテスト情報量、項目母数

5.2.5 中学校英語

表 15：学力スコアの標本統計量（中学校・英語）

実施年度	英語（学力スコアの標本統計量）					
	生徒数	平均	標準偏差	25 パーセンタイル	中央値	75 パーセンタイル
令和３年度	22,946	501.1	110.2	426.9	500.0	572.7

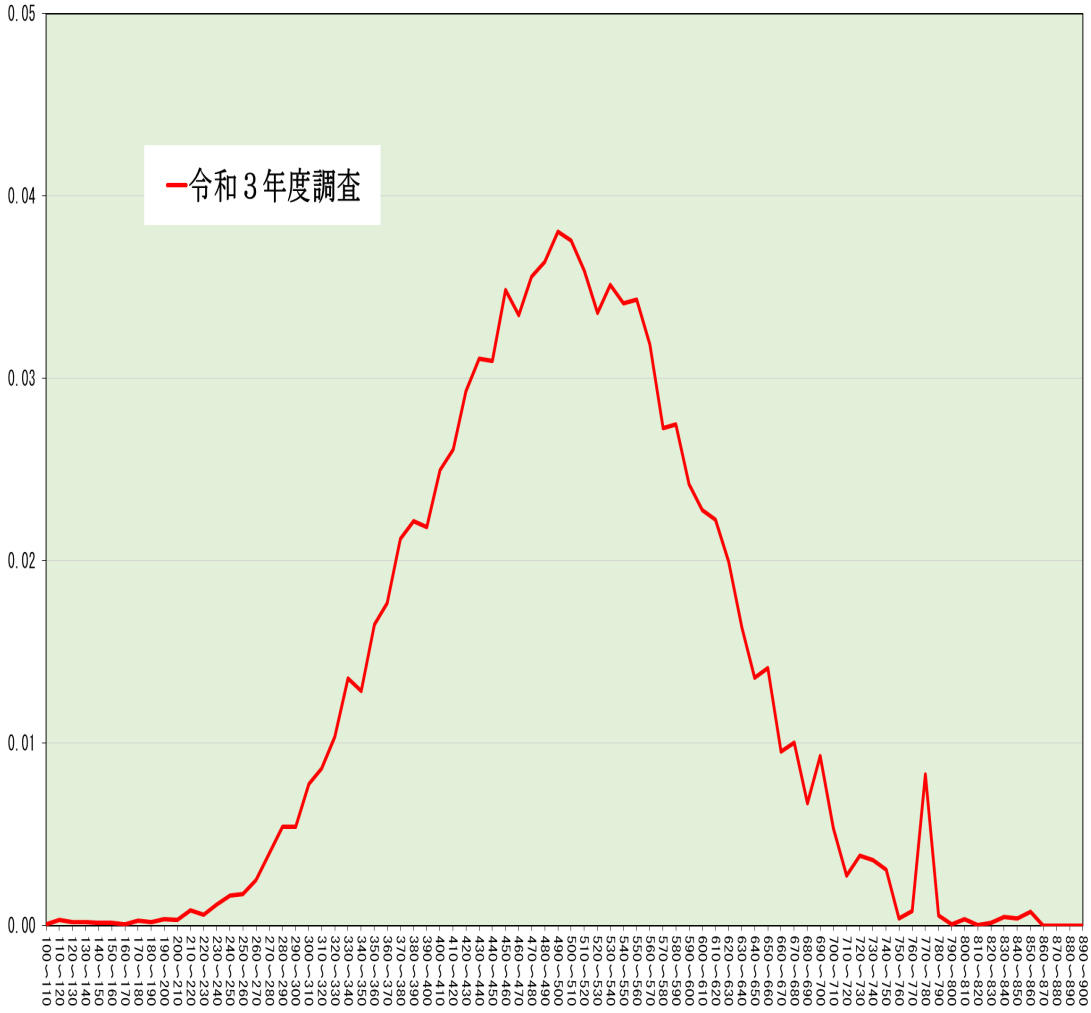


図 40：中学校：英語：学力スコア：相対度数分布

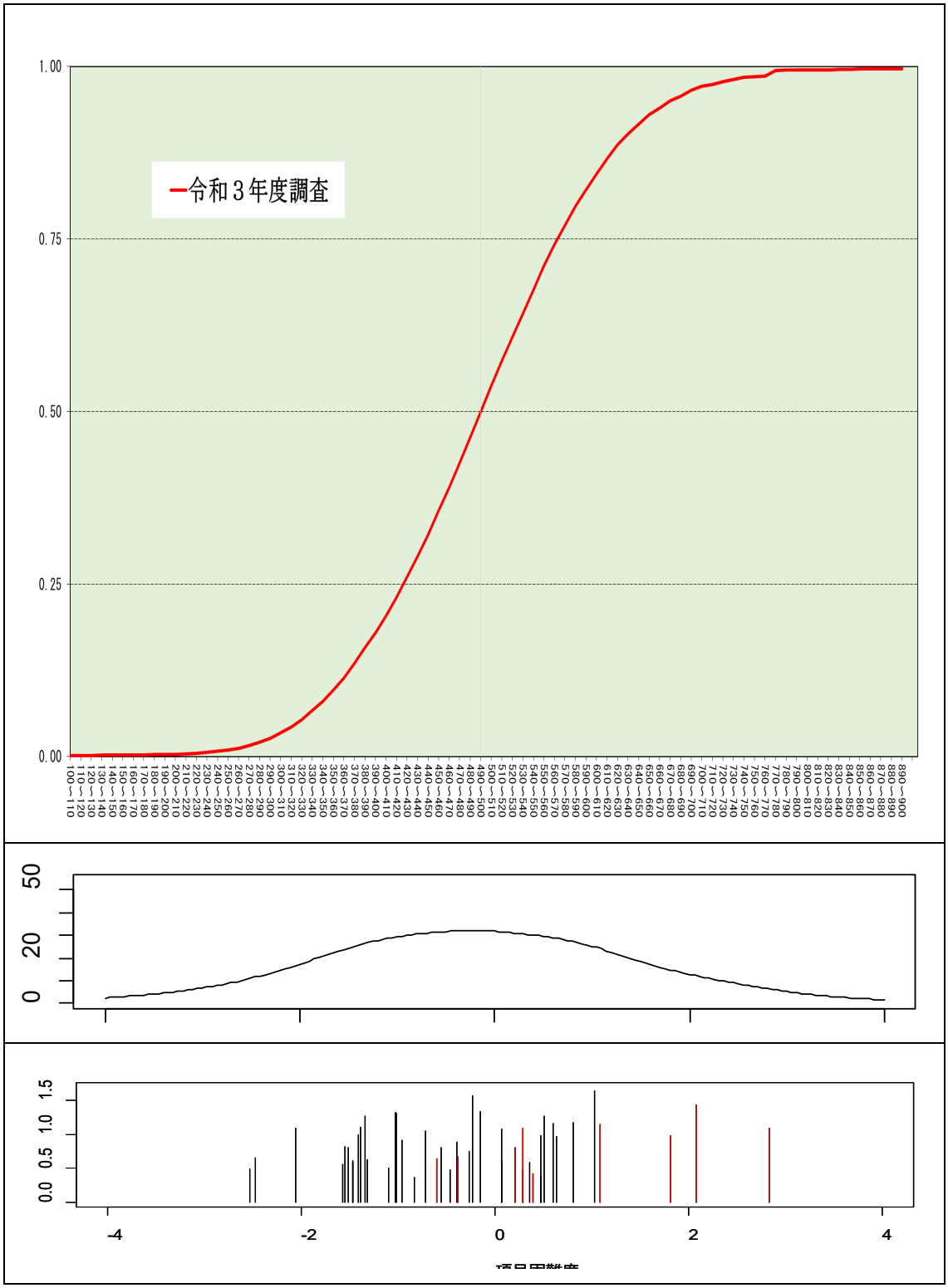


図 41：中学校：英語：学力スコア：累積相対度数分布とテスト情報量、項目母数

参考文献

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.

Baker, F. B., & Kim, S. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Marcel Dekker.

Beaton, A. E. (1987). Implementing the New Design: The NAEP 1983-84 Technical Report. National Assessment of Educational Progress, Educational Testing Service, Rosedale Road, Princeton, NJ.

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.

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability : Implications for data aggregation and analysis. In K.J. Klein & S. W. J. Kozlowski（Eds.）, *Multilevel theory, research, and methods in organizations : Foundations, extentions, and new directions*. San Francisco, CA : Jossey-Bass, pp.349-381

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29-51.

Bock, R. D. (1997). The nominal categories model. in W. J. van der Linden & R. K. Hambleton（Eds.), *Handbook of modern item response theory*, New York: Springer. pp. 22-49

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.

Child,R. A. & Jaciw, A.P. (2003). Matrix sampling of Items in Large-Scal Assessments, *Practica Assessment, Research & Evaluation*, *8*

(Retrieved from http://PAREonline.net/getvn.asp?v=8&n=16)

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.

Drasgow, F., Levine, M. V., & Williams, M. E. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, **38**, 67-86.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement*, **11**, 59-79.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. London: Lawrence Erlbaum.

Frey A., Hartig, J. & Rupp, A.A. (2009). An NCME Instructional Moduke on Booklet Designs in Large-Scal Assessments of Student Achievement: Theory and Practice, *Educational Measurement: Issues and Practice*, **28**, 3, 39-53.

Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel / Hierarchical Models*. Cambridge University Press.

Goldstein, H. (2011). *Multilevel Statistical Models. 4th ed.* London: Wiley.

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.

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.

Hox, J. J. (2010). *Multilevel Analysis：Techniques and applications(second ed.)*, Hove, UK：Routledge Academic.

伏見正則 (1989). 乱数. 東京大学出版会.

池田央 (1972). 心理学研究法第 7 巻テスト I. 東京大学出版会.

池田央 (1994). 現代テスト理論. 朝倉書店.

池田央 (2010). 重複テスト分冊法と学力調査　指導と評価,1 月号,44-47,図書文化.

石井吾郎 (1972a). 実験計画法の基礎　サイエンス社

石井吾郎 (1972b). 実験計画/配置の理論　培風館

Kendall, M. G., & Stuart, A. (1979). *The advanced theory of statistics* (4th ed., Vol. 2). New York: Oxford University Press.

Kolen, J. Michael., & Brennan, R. L. (2014). *Test Equating, Scaling, and Linking*：*Methods and Practices(3rd ed.)*. New York: Springer.

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**, 107-118.

Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data (2nd Edition)*, New York: Wiley.

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.

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.

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）．受検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）．共通受検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.

Olson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient, *Psychometrika*, **44**, 443-460.

Pophan, W.J. (1993) Circumventig thehigh costs of authentic assessment. *Phi Delta Kappan*,**7,**470-473.

Reckase, M., D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*(3), 207-320.

芝祐順（編）(1991)．項目反応理論―基礎と応用―．東京大学出版会.

Steele, F., & Harvey Goldstein, H. (2007) Multilevel Models in Psychometrics. In C. R. Rao., & S. Sinharay.（Eds.）, *Handbook of Statistics, Vol.26.* North Holland. pp.4 01-420

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, **7**, 201-210.

竹内啓編 (1989)．統計学辞典　東洋経済新報社

津田孝夫 (1995)．モンテカルロ法とシミュレーション―電子計算機の確率論的応用―（三訂版）培風館.

豊田秀樹 (2005)．項目反応理論［理論編］－テストの数理－． 朝倉書店.

Upton,G. & I. Cook (2010). A Dictionary of Statistics, Oxford University Press.

van der Linden,W.J.(Eds.) (2016). *Handbook of Item Response Theory*. Chapman and Hall/CRC.

van der Linden, W. J., & Hambleton, R. K. (Eds.) (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful? *IERI Monograph Series Vol. 2*, pp.9-36

Wu, M. (2004). Plausible values. *Rasch Measurement Transactions*, **18**(2), 976-978.

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.

関係報告書一覧：全国的な学力調査（全国学力・学習状況調査等）追加分析調査

平成 22 年度　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（追加分析報告書）

平成 23 年度　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（追加分析報告書）

平成 24 年度　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（追加分析報告書）

平成 25 年度　東日本大震災の学力への影響～IRT 推算値による経年比較分析～（追加分析報告書）

平成 27 年度　全国学力・学習状況調査における経年変化分析調査の年度間等化に関する調査研究（追加分析報告書）

平成 29 年度　経年変化分析調査との対応づけによる本体調査の年度間比較の試み（追加分析報告書）

付録 A：EasyEstimation オプション指定関係

A.1 項目母数の推定・等化

（ア）Estimation Item parameters： 2 PL モデル

（イ）デフォルトから変更する Option：

①Const D：1.702

②Crit. M-step：0.001

（エ）中学校英語の 4 技能は分けずに処理（R3 基準年）

（オ）等化

①初期推定 ：項目固定法による等化を行わずに、単独で項目母数を推定

②等化方法：R3 は H28 項目をアンカーとして項目固定法による等化

③注意：いくつかの方法を試み専門的判断の上、最適な結果を採用すること

A.2 能力母数（受検者母数）の推定

（ア）能力母数の推定

①能力母数の推定：MLE／MAP／EAP／POP

②デフォルト設定

③PV の数：10

（イ）詳細指定

①MLE　　デフォルト

②MAP　　事前分布デフォルト：標準正規分布

③EAP　　事前分布デフォルト：標準正規分布 ／求積点 41

④POP　　デフォルト：範囲（-4< θ <4）／求積点 41

（※POP：Mislevy & Bock(1982)の estimation of the latent distribution を実行）

⑤Plausible Value（PV）　事前分布デフォルト：標準正規分布

（ウ）分冊における全問正答／誤答の場合の MLE の処理

①全問正答／誤答の回答項目のうち項目識別力が最低の項目に 0.5 を与える方式

②option：「全問正答誤答処理(1 PL or 2 PL only)」にチェック

（エ）EasyEstimation の機能等留意点

①入力データは BIBD のフィッシャー表現による大行列のままで可能

（不完全データとなっていても EasyEstimation 側で観測値のあるところを読み取り、その中で項目識別力最小の項目に対応する部分を自動的に 0.5 と設定する）

②大行列ではなく分冊ごとのデータ行列でも推定可能

③PV：乱数のシーズ指定が可能

④項目母数等はすでに計算済みのものを使用

付録 B：EasyEstimation の仕様

B.1　EasyEstimation シリーズの概要

経年変化分析調査においてはテスト仕様の設計の際に、測定モデルとして項目反応理論（IRT）を用いている。そのため分析には専用の計算ソフトウェアが必須である。そのような専用ソフトには商用ソフトとしては Scientific Software International 社の BILOG-MG や IRTPRO、PARSCALE、Australian Council for Educational Research（ACER）の ConQuest などが、またフリーソフトとしては R の ltm パッケージなどがある。PARSCALE の発生バージョンは NAEP で、ACER ConQuest は PISA などで採用されている。しかし、いずれも購入価格の問題やかなり特殊なプログラミングスキルが必要なこと、また ConQuest では扱える項目反応モデルに制限があることなどから、一般には利用しにくい。そこで、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の展開可能性について」で正式に採用し、BILOG-MG や IRTPRO の推定結果ともクロスバリデーションを行いながら結果の正確さを担保できた、熊谷(2009)による EasyEstimation を経年変化分析調査では採用している。

EasyEsimation（熊谷 2009、2012）[5]には、ユーザーインターフェイスを初め、以下に示すような優れた特徴と使い勝手の良さがあるため、我が国ではすでに多くの研究機関等へダウンロードされ新しいテスト開発などに利用されている。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、さらに広く使われることになるであろう。

B.2 EasyEstimation の特徴

1）研究目的に限り無料で利用できる国産のフリーソフトウェアである。
2）マウス操作のみで分析可能な GUI（グラフィカル・ユーザ・インターフェース）により、テスト分析の入門者においても容易に分析を実行できる。
3）多母集団分析や項目母数を一部固定しての分析など、実用上必要な分析オプションが豊富に準備されている。
4）計算結果の精度は他のソフトウェアを使ったクロスバリデーションで確認済み。
5）EasyEstimation は 2 値型データ以外に順序付多値型データ、名義尺度データが可能。
6）DIF（特異項目機能）分析用の EasyDIF（熊谷 2012）も公開されている。

主な機能としては、

[5] http://irtanalysis.main.jp/

1. テトラコリック（tetrachoric：四部）相関係数行列からの固有値を利用した 1 次元性の仮定の確認、
2. 項目母数の推定、
3. 受検者母数の推定、
4. 項目特性曲線、テスト情報量曲線の出力、

の 4 つを持っている（図 B.1）。

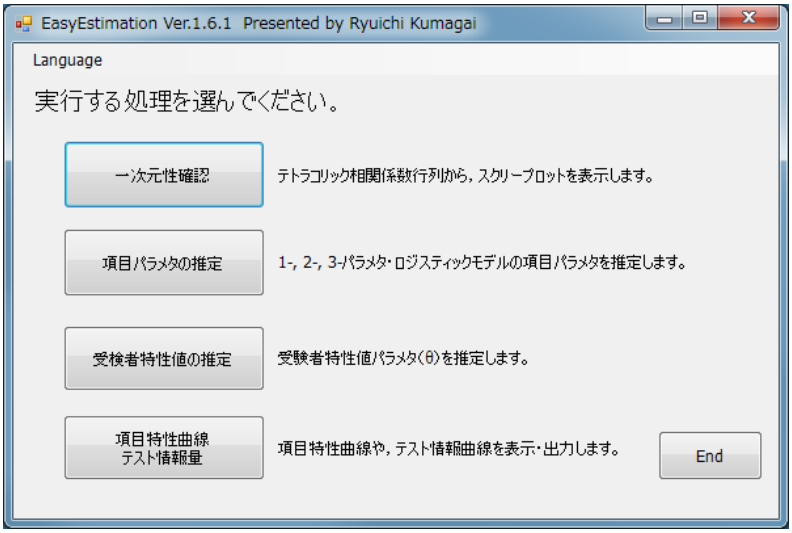


図 B.1　EasyEstimation 実行画面

B.2　テストの 1 次元性確認

通常の IRT 分析ではテストが 1 次元性を有している（一つの構成概念を測定している）という仮定を置いている（2 次元以上を仮定した多次元 IRT モデルも提唱されている）。EasyEstimation では、テストが 1 次元性を有しているかを確認するための一つの指標として、テトラコリック相関係数行列から算出される固有値をグラフにしたもの（スクリープロット）を出力する。通常、因子分析の枠組みで固有値を計算する場合には積率相関係数行列が用いられるが、カテゴリー数が少ない場合のカテゴリカルデータ（例えば 0－1 の 2 値型データ）に対して分析を実行すると、「困難度」に対応した因子が抽出されるなど、望ましくない面があるため、テスト分析の文脈ではテトラコリック相関係数行列（多値型データではポリコリック相関係数行列）が用いられる。

テトラコリック相関係数は、分析的に求めることができないため、数値計算により推定値が求められることとなる。EasyEstimation では、Olson(1979)で述べられている周辺度数を利用した最尤推定法によりテトラコリック相関係数を推定している。

B.3　項目母数の推定

EasyEstimation では、2 値型データに対する項目反応モデル（1、2、3 パラメタ・ロジスティック・モデル）の項目母数を推定することができる。項目母数の推定には、周辺最尤推定法（3 パラメタ・ロジスティック・モデルにおいてのみ、当て推量パラメタにベータ分布を事前分布とするベイズ推定を利用）を採用している。EasyEstimation を用いた項目母数推定においては、一部の項目母数の値を既知として、残りの項目母数を推定する機能（項目固定法）が利用できる。経年変化分析調査における等化分析ではこの機能を利用している。

B.4　受検者母数の推定

EasyEstimation では、受検者母数の推定方法として 1）最尤推定法、2）maximum a posteriori（MAP）法、3）expected a posteriori（EAP）法を利用することができる。最尤推定法の場合、全問正答（誤答）の受検者においては、受検者母数を推定することができなかったが、EasyEstimation では最も識別力の低い項目について 0.5 正答（誤答）という項目反応パターンを与えることで、全問正答（誤答）の反応パターンに対しても、何らかの推定値を与えるオプションが用意されている。受検者母数推定においては、母数推定値のほか、その標準誤差および受検者適合度（person-fit）指標が算出される。この受検者適合度指標については、Drasgow, Levine, & Williamas（1985）および Drasgow, Levine, & McLaughlin（1987）によるZ_L統計量を採用している。Z_L統計量は標準正規分布に従い、この値が負に大きいほど当該受検者の解答傾向がモデルから乖離している（例えば、困難度が高い項目にばかり正答し、困難度が低い項目には誤答している）ことになる。

また EasyEstimation では、受検者母数そのものではなく、母集団分布の推定を行う機能もある。受検者母数を最尤推定法で求めた場合、全受検者母数の標準偏差を求めると、誤差の影響により真の値よりも標準偏差が大きくなる（EAP や MAP を利用した場合は、事前分布の平均方向に推定値が縮小するというベイズ推定値の特徴により、標準偏差は小さくなる）。そこで、各受検者の母数を利用するのではなく、母集団分布を推定することで誤差の影響は抑え、標準偏差の推定値を真の値に近づけることが可能となる（推定オプションの指定は「POP」）。

B.5　項目特性曲線・テスト情報量曲線の表示

項目の特徴は 10.3 で推定された項目母数により決定されるが、数値そのものよりも、項目特性曲線（item characteristic curve: ICC）を俯瞰することで得られる情報が多い。特に多値型デー

タでは複数の項目母数が得られることから、具体的に ICC を眺めることが必須となる。EasyEstimation では、この ICC についても簡易に出力することができる。

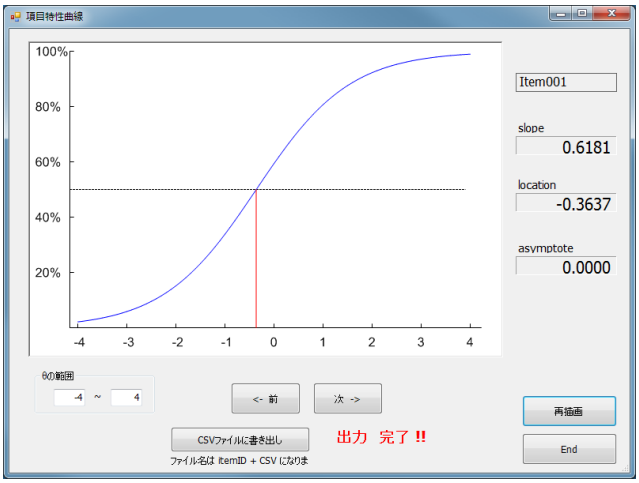


図 B.2　ICC 表示画面

また IRT の利点の一つとして、テスト情報量を用いたテストの精度の表示が挙げられる。古典的テスト理論の枠組みでは、テストの精度は信頼性係数として 0〜1 の数値で表現された。IRT では、テストの精度をテスト情報関数という、潜在特性尺度値の関数として表現する。つまり、どの能力範囲において精度が高いのか（低いのか）を表現することができ、この関数をグラフにしたものをテスト情報量曲線と呼ぶ（図 B.2）。また、信頼性係数がデータセット（標本）に依存して算出される（言い換えれば、同じ問題項目からなるテストでも、受検者が異なれば係数の値が異なる）のに対し、テスト情報量は項目母数により決定されるため、データセットとは独立していることも大きな違いである。EasyEstimation では、このテスト情報量曲線も、ICC と同様に簡便に出力できる（図 B.3）。

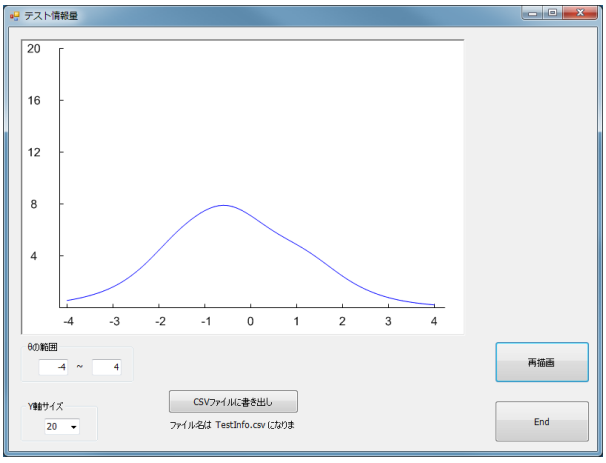


図 B.3　テスト情報量曲線表示画面

付録 C：推算値の計算アルゴリズム

C.1　推算値

推算値（plausible values : PV's）とは多重代入法（multiple imputation）に関する Rubin(1987) の理論的基礎をもとに、Mislevy ら(1991) により大規模アセスメントに適用された手法であり、現在 PISA や TIMSS をはじめとした国際的な学力テストにおいても採用されている。推算値は受検者の能力母数 θ の事後分布からの無作為に取り出した複数の能力値である。推算値を使う利点として、個人の能力の推定結果の不確実性を考慮することができ、また項目数が少ない場合でも、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点が挙げられる。

ここでは大規模学力調査における下位領域ごとの集団比較について考える。通常、PISA をはじめとした国際的な学力調査では、試験のデザインに経年変化分析調査と同様、重複テスト分冊法が採用されている。しかし、経年変化分析調査が一つの分冊には国語なら国語のみの問題から構成されているのに対して、PISA では cluster と呼ばれる項目群にさらにわかれ、ある cluster には科学的リテラシーの問題が、同じ分冊の他の cluster ではリーディング・リテラシーの問題が混合されて一つの分冊が構成されている。そのため、経年変化分析調査や、一般の受検者別の能力を測定することを目的としたテストに比べると各受検者が解答する各クラスターあたりの項目数は少ない。

しかし、経年変化分析調査でも、例えば下位領域ごとの比較を目的とすると利用できる項目数がさらにしぼられることを想定すると同様のことは生じる。そのため、下位領域に分けられた少数個の項目から得られる個々の EAP（expected a posteriori）推定値や MLE（maximum likelihood estimation）推定値を直接用いて集団の能力分布を推定する方法の場合、分散の過大評価・過小評価が生じ、集団間の正確な分散の比較ができない。すなわち、各受検者が解答する項目数が少なく各受検者の能力値から集団統計量を算出する方法の場合、正確に分散を評価できない。もっとも、経年変化分析自体はそのときどきの指導要領に依存する問題設計になっているわけではない。むしろ、指導要領に即して作成される本体調査の CBT 化が実現したときにはそちらで必要になる分析方法であると予想されているものである。

C.2　von Neumann の棄却法（rejection method）

ここでは事後分布からの無作為抽出をおこなうにあたり von Neumann の棄却法（rejection method）を取り上げる。棄却法は特定の領域において乱数を発生させ採択域内であればその値を分布に従う乱数としてとり、それ以外の領域に発生していた場合棄却し任意の個数の乱数を

得るまでそのサブルーチンを繰り返す非常に素朴な手法である。この方法を利用することで一様乱数を用いて受検者の能力母数 θ の事後分布に従う標本を生成することができる。受検者の能力母数 θ の事後分布$h(\theta|\mathrm{x})$は、

$$h(\theta|\mathrm{x}) = \frac{f(\mathrm{x}|\theta)g(\theta)}{\int f(\mathrm{x}|\theta)g(\theta)d\theta} \tag{C.1}$$

で表される。このとき受検者の項目反応パターンをx、能力母数を θ、θ が所与のときの項目反応パターンxの条件付き確率密度関数を$f(\mathrm{x}|\theta)$としている。また、事前分布$g(\theta)$には、通常、標準正規分布が仮定されることが多い。

一様乱数を発生させる領域を設定する際、受検者の能力母数 θ の尺度値である EAP 推定値や事後分布の最大確率密度が既知である必要がある。そのため、棄却法を行うにあたって受検者の能力値に関する推定事後分布を求める必要がある。本研究では効率よく棄却サンプリングを行うため、一様乱数を発生させる領域を、横軸にあたる θ に関しては[θ_{EAP}-4.75、 θ_{EAP}+4.75]、縦軸にあたる確率密度を 0 から最大事後確率に 1.0001 をかけた値として設定する。

棄却法による推算値の算出の一連の過程を整理すると、

① θ_{EAP} ±4.75 で一様乱数を発生させ θ の仮の値とする
② 0 から最大事後確率に 1.0001 をかけた値までの間で一様乱数を発生させる
③ 発生した乱数が採択域である密度関数に①の値を代入したものよりも小さければ推算値として採用し、それ以外の場合棄却する
④ 任意の数（例えば各受検者に対し 10 個）の推算値が得られるまで繰り返す

となる。棄却法の概念図は図 C.1 で示したようになる。このとき赤で示した点が採択された乱数で、青で示した点が棄却された乱数を示している。

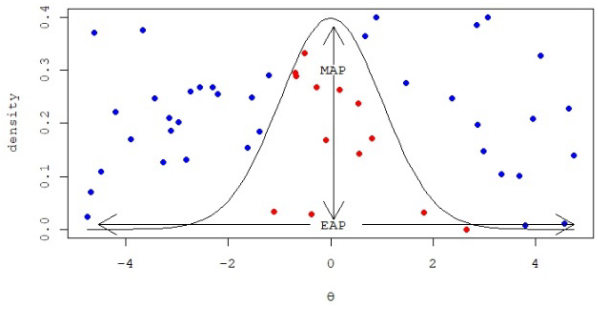


図 C.1 棄却法の概念図

C.3　推定事後分布

C.3.1　EAP 推定値と周辺分布

上述したように、推算値を得るために推定事後分布を求める必要がある。はじめに EAP 推定値と周辺分布を求める。EAP 推定値は θ の事後分布の期待値として、

$$\theta_{EAP} = \int_{-\infty}^{\infty} \theta h(\theta|x)\, d\theta \qquad (C.2)$$

と定義される。ただしこの積分計算を解析的に行うことはできないため、数値計算により近似的に解を求める。具体的には $L(X_i)$ を $f(\mathrm{x}|\theta)$ の対数尤度関数とし、ノード（分点）をx_i、ノード数を N として、エルミート・ガウス求積法（Gauss-Hermite quadrature）を用いて算出する。

村木(2011) は分点について等間隔で区切るよりもエルミート・ガウス求積法を採用したほうが効率的に精度の高い値が得られる利点があることが報告している。ガウス型公式は、

$$\int_b^a w(x)f(x)dx = \sum_{i=1}^{N} A_i f(x_i) \qquad (C.3)$$

であり、エルミート・ガウス求積法における区間$[a,b]$は $[-\infty,\infty]$、重み関数$w(x)$は、

$$w(x) = \exp(-x^2) \qquad (C.4)$$

とあらわされ、係数A_iは、

$$A_i = \frac{2^{n+1}n!\sqrt{\pi}}{[H_n'(x_i)]^2} \qquad (C.5)$$

となる。求積点X_iにおける事後分布を表す係数G_iは、

$$G_i = \frac{L(X_i)A_i}{\sum_{i=1}^{n} L(X_i)\, A_i}, \quad (i = 1,2,\dots,n) \qquad (C.6)$$

となる。このとき$\sum_{i=1}^{n} L(X_i)\, A_i$は周辺分布である。以上より EAP 推定値は、

$$\theta_{EAP} = \sum_{i=1}^{N} X_i G_i \tag{C.7}$$

で計算できる。このとき得られた EAP 推定値を基準に±4.75 したものを θ に関して発生させる乱数の上限と下限とする。

C.3.2　MAP 推定値

次に MAP 推定値を求める。具体的には加藤(2014) を参考に算出することとする。ただし計算するにあたりに解析的に求めることが困難であることから適当な初期値から解が収束するまで繰り返し推定値を更新する数値計算により近似的に解を求める。数値計算の手法として、ここでは Fisher のスコア法および反復回数が 100 回を超えるものに関しては二分法を利用することで MAP 推定値を得る。得られたθ_{MAP}に 1.0001 をかけたものを採択域である分布関数の確率密度における上限とする。このとき下限は 0 とする。

以上の手続により推算値を算出するにあたり必要な乱数の発生領域の設定ができる(図 C.1 参照)。この発生領域に一様乱数を発生させ棄却法のアルゴリズムを利用することで推算値を求めることができる。

付録 D：多値項目反応モデル

D.1　多値モデル

短答形式や論述形式には、受検者が解答を実際に記入／記述するという多肢選択形式にはない特徴がある。その反面、例えば、正答が複数ある場合や解答のつづり字に間違いがある場合などには正答の判断が主観的になることがあり、採点の客観性が低くなることが少なくない。とくに複数の採点者がいる場合、あらかじめ採点者間で採点基準をそろえるといった手続きが必要になってくる。また採点結果も単純な正誤ではなく、部分点の付与も必要となる。ＣＢＴ化が進めば自然言語処理などの技術をつかって、この問題はある程度解決されると見込まれるが、いずれにしても、手採点あるいは機械採点の結果えられた多値データを尺度値 θ に変換する必要がある。そのために有用な IRT モデルの一つが多値項目反応モデル（多値モデル、順序モデルなどとも総称される）である。

D.2　多値モデルの必要性

多肢選択形式の項目の場合、採点結果は基本的に 2 値データ（0: 誤答、1: 正答）として表現できる。もしテストがこの形式の項目から構成されていれば、図D.1（左）に示すように、テスト結果は行を受検者、列を項目とした 2 値行列として表現できる。このような行列データは、項目反応データ（item response data）あるいは項目反応パターン（item response pattern）などと呼ばれる。2 値の項目反応データを取り扱うのに適した項目反応モデル（item response model）には、Rasch モデル（Rasch model）（Rasch 1960）、1 母数ロジスティックモデル（one parameter logistic model、 1 PL モデル）、2 母数ロジスティックモデル（two parameter logistic model、２PL モデル）、３母数ロジスティックモデル（three parameter logistic model、 3 PL モデル）（Birnbaum、 1968; Lord 1952; Lord & Novick 1968）がある。このうち、Rasch モデルと 1 PL モデルは数学的には同一であるものの、その発展過程が異なることから、それぞれ Rasch 系モデルと Thurstone 系モデルに属するモデルとして区別される（村木 2011）。

短答形式（論述形式）の項目の場合、採点基準によっては部分点（partial credit）を与えて多段階に採点することがある。このとき、テスト結果は 2 値データではなく多値データとして表現される。もし各項目が（0: 誤答、1: 部分点、2: 正答）と 3 段階の順序データとして採点されるなら、テスト結果として図 D.1（右）に示すような多値の項目反応データが得られる。項目反応モデルの中には、このような段階反応（graded response）を取り扱うことのできる多値モデルが考案されている。その代表例として、段階反応モデル（graded response model: GRM）（Samejima 1969）、部分採点モデル（partial credit model : PCM）（Masters 1982），一般化部分採点モデル（generalized partial credit model : GPCM）（Muraki 1992）があげられる。

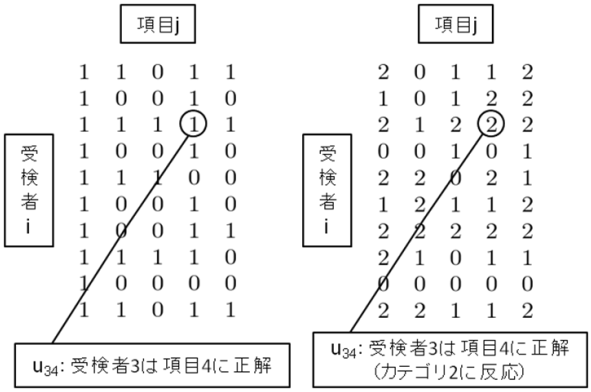


図 D.1　2 値の項目反応データ（左）と多値の項目反応データ（右）の例

D.3　モデル選択の問題

IRT を利用する際には、各項目にどの項目反応モデルを適用するかを決定する必要がある。その判断基準には、項目形式、受検者数、項目数、モデルフィットの問題、結果の解釈（説明）のしやすさの問題、さらには判断する人の哲学や好みなど、様々な要素が考えられる。そのため、モデルを選択する際には、テストの専門家などによる総合的な判断が求められる。

全米学力調査 NAEP の場合、多肢選択形式の項目には 3 母数ロジスティックモデル、採点結果が正答・誤答の 2 値で表現される短答形式の問題には 2 母数ロジスティックモデル、採点結果が段階反応となる論述形式の項目には GPC モデルが採用されている。PISA では、選択肢形式、論述形式などの項目に 1 母数ロジスティックモデルが利用されている。また、医療系大学間共用試験医学系 CBT では、多肢選択形式の項目に 2 母数ロジスティックモデルが利用されている。

経年変化分析調査の場合、多肢選択形式の項目には 2 母数ロジスティックモデルを採用することにした。多肢選択形式の項目の場合、当て推量母数がモデルに含まれる母数ロジスティックモデルを採用することも考えられる。しかしながら、受検者数を踏まえた推定結果の安定性、結果の解釈のしやすさの問題、これまでの実証的研究の結果から 2P 母数ロジスティックモデル採用が妥当と判断した。

D.4　段階反応モデル

Samejima(1969) が考案した GR モデルは、多段階で採点されるテスト結果の分析だけでなく、多くの心理検査や社会調査で用いられるリッカートタイプの質問項目（例：質問に対し、5 段階で当てはまる程度を答えさせる）などにも適用される。

項目jは多段階に採点される項目であり、その段階反応（採点結果）はK個$(0,1,\cdots,K-1)$のカテゴリーに分類されるとする。GR モデルでは、潜在特性値θをもつ受検者が項目jにカテゴリーkと反応する確率（採点される確率）$P_{jk}(\theta)$は、

$$P(u_j = k|\theta) = P_{jk}(\theta) = P^*_{jk}(\theta) - P^*_{jk+1}(\theta) \tag{D.1}$$

と定義される。u_jは受検者の項目jj への反応であり、右辺の$P^*_{jk}(\theta)$は潜在特性値θをもつ受検者が項目jに$u_j \geqq k$と反応する確率を表している。すなわち、$P_{jk}(\theta)$は潜在特性値θをもつ受検者がk以上のカテゴリーに反応する確率と$k+1$以上のカテゴリーに反応する確率の差として定義される。θを変数として見たとき、$P_{jk}(\theta)$は項目反応カテゴリー曲線（item response category characteristic curve: IRCCC）、$P^*_{jk}(\theta)$は境界特性曲線（boundary characteristic curve: BCC）と呼ばれる。

GR モデルでは、BCC を正規累積モデルあるいはロジスティックモデルによって表現する。導関数を計算しやすいなど、数学的な取り扱いが容易であるという理由から、BCC として 2 PL モデルがよく利用される。カテゴリー$k = 1,2,\cdots,K-1$における$P^*_{jk}(\theta)$を 2 PL モデルによって表現すると、

$$P^*_{jk}(\theta) = \frac{1}{1 + \exp\left[-Da_j(\theta - b^*_{jk})\right]} \tag{D.2}$$

となる。ただし、受検者は必ずK個（$0,1,\cdots,K-1$）のカテゴリーのいずれかに反応するものとし、

$$P^*_{j0}(\theta) = 1 \tag{D.3}$$

$$P^*_{jK}(\theta) = 0 \tag{D.4}$$

とする。

(D.2)式の BCC に含まれる母数のうち、項目の特性を記述するための母数は、識別力母数（discrimination parameter）a_jと BCC の位置母数（location parameter）b^*_{jk}である。GR モデルでは、段階的な反応を記述するため、項目内の識別力母数はすべて等しく、BCC の位置母数の値はカテゴリーkの昇順に大きいと仮定する。それゆえ、a_jの添え字にカテゴリーkは含まれず、BCC の位置母数には、

$$b^*_{j1} < b^*_{j2} < \cdots < b^*_{jk} < \cdots < b^*_{jK-1} \tag{D.5}$$

という大小関係がある。(D.2)式のDは尺度因子（定数）であり、通常、$D=1$や$D=1.7$(より正確には、1.702)が利用される。$D=1.7$とすれば、(D.2)式は正規累積モデルの非常によい近似となる。

(D.1)式の IRCCC は、識別力母数a_jと IRCCC の位置母数b_{jk}によって記述される。カテゴリー0 とカテゴリー$K-1$については、それぞれ$P_{j0}(\theta)=0.5$と$P_{jK-1}(\theta)=0.5$となる潜在特性値を IRCCC の位置母数として利用する。(D.2)式の 2 PL モデルでは$\theta=b_{jk}^{*}$のとき$P_{jk}^{*}(\theta)=0.5$となることに注意すれば、(D.1)式、(D.3)式、(D.4)式から IRCCC の位置母数b_{j0}、 b_{jK-1}は、

$$b_{j0}=b_{j1}^{*} \tag{D.6}$$

$$b_{jK-1}=b_{jK-1}^{*} \tag{D.7}$$

と表現される。また、カテゴリー$k=1,2,\cdots,K-2$については、IRCCC の位置母数b_{jk}として、そのカテゴリーをとる確率がもっとも高くなる潜在特性値を利用する。(D.1)式が(D.2)式の$k,k+1$との差で定義されることから、カテゴリーkの IRCCC には識別力母数a_jと BCC の位置母数b_{jk}^{*},b_{jk+1}^{*}が含まれる。このとき、IRCCC の位置母数b_{jk}と BCC の位置母数b_{jk}^{*},b_{jk+1}^{*}との関係は、

$$b_{jk}=\frac{b_{jk}^{*}+b_{jk+1}^{*}}{2} \tag{D.8}$$

と表現される。

図 D.2（左）～図 D.4（左）に、3 つのカテゴリー$k=0,1,2$をもつ項目の IRCCC の例を示す。各項目の識別力母数の値、IRCCC の位置母数の値は図下（注）に示すとおりである。(D.2)式の尺度因子は、いずれの項目も$D=1$を利用している。

図 D.2（左）の IRCCC をみると、潜在特性値 θ に対する項目 1 の特性を把握することができる。すなわち、$\theta=0$付近の潜在特性値をもつ受検者はカテゴリー1 にもっとも反応しやすく、それより小さい θ をもつ受検者はカテゴリー$k=0,1,2$の順に反応しやすく、それより大きい θ をもつ受検者はカテゴリー$k=2,1,0$の順に反応しやすいことがわかる。また、最下位のカテゴリー0 に反応する確率は潜在特性値 θ に対して右下がりの曲線で表現され、 $\theta=b_{10}=-1$のときにその反応確率が 0.5 になっている。同様に、最上位のカテゴリー2 の場合、反応確率は右上がりの曲線で表現され、$\theta=b_{12}=1$のときに反応確率がちょうど 0.5 になっている。一方、中間のカテゴリー1 の場合、そのカテゴリーに反応する確率は単峰形の左右対称な曲線で表現され、$\theta=b_{11}=0$のときに最大値をとっている。

図 D.2（左）と図 D.3（左）を比較すると、識別力の違いが IRCCC にどのような影響を与えるかを理解することができる。項目 1 と項目 2 の違いは識別力母数の値だけであり、IRCCC の位置母数は同一の値である。両者の IRCCC をみると、項目の識別力が高くなると IRCCC の傾斜が急になる傾向があり、受検者の反応したカテゴリーに応じて受検者の潜在特性値を識別しやすくなることがわかる。

図 D.4 は、IRCCC の位置母数の間隔が極端に狭く、中間のカテゴリーへの反応確率が非常に低い場合の例である。(0：誤答、1：部分点、2：正答）と採点する項目ならば、項目 3 は部分点をとる受検者が極端に少なく、受検者の解答が正答か誤答かのどちらかにはっきりと分かれる特性をもつ。数学の問題で例をあげるなら、2×3+5 を解くにあたり、正しく解けなかった人は 0 点、2×3 まで解けた人は 1 点、正しく解けた人は 2 点と採点することにする。通常、たし算はかけ算よりかなり易しいので、2×3 まで解ける人は最終的な正解まで答えられる可能性が高い。このような場合、部分点をとる受検者は極端に少なくなり、受検者の解答は正答か誤答かにはっきりと分かれる。

ところで、BCC に 2 PL モデルを利用する場合、2 つのカテゴリー$k = 0,1$をもつ GR モデルは 2 PL モデルに一致する。(D.1)式および$P_{j0}^{*}(\theta) = 1$、 $P_{j2}^{*}(\theta) = 0$より、

$$P_{j0}(\theta) = P_{j0}^{*}(\theta) - P_{j1}^{*}(\theta) = 1 - P_{j1}^{*}(\theta) \tag{D.9}$$

$$P_{j1}(\theta) = P_{j1}^{*}(\theta) - P_{j2}^{*}(\theta) = P_{j1}^{*}(\theta) \tag{D.10}$$

となる。2 つのカテゴリーを（0: 誤答、1: 正答）と考えれば、(D.9)式と(D.10)式は、それぞれ 2 PL モデルの誤答確率と正答確率を表現している。

2 PL モデルなどと同様に、GR モデルは潜在特性値 θ が 1 次元（スカラー）の項目反応モデルに属する。それゆえ、GR モデルを適用するには、対象となるテストが局所独立の仮定と 1 次元性の仮定を満たす必要がある。局所独立の仮定とは、1 つのテストにおいて、ある潜在特性値をもつ受検者がある項目に正答する確率は他の項目に正答する確率の影響を受けないという仮定である。確率論的には、ある受検者が各項目に正答するのは互いに独立な事象であるということを意味する。1 次元性の仮定とは、1 つのテストを構成する項目はただ 1 つの構成概念を測定するものでなければならないという仮定である。実際には、スクリープロットや各種の統計量を用いてモデルに必要な仮定やモデルのデータへのあてはまり具合などを確認する。

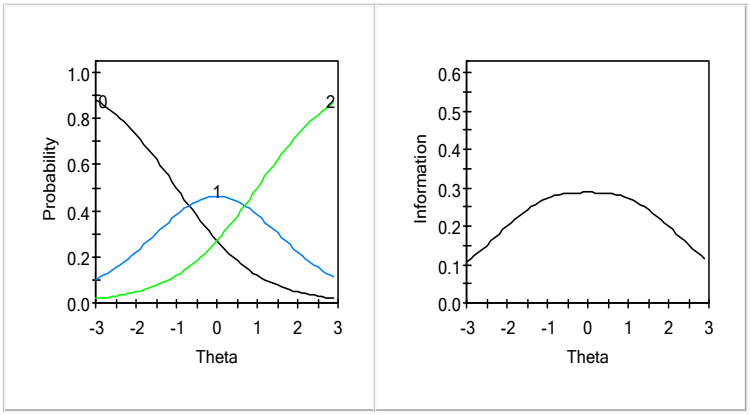


図 D.2　項目 1 の IRCCC（左）と項目情報量（右）

（注）a_1=1, b_{10}=-1, b_{11}=0, b_{12}=1, (D=1)

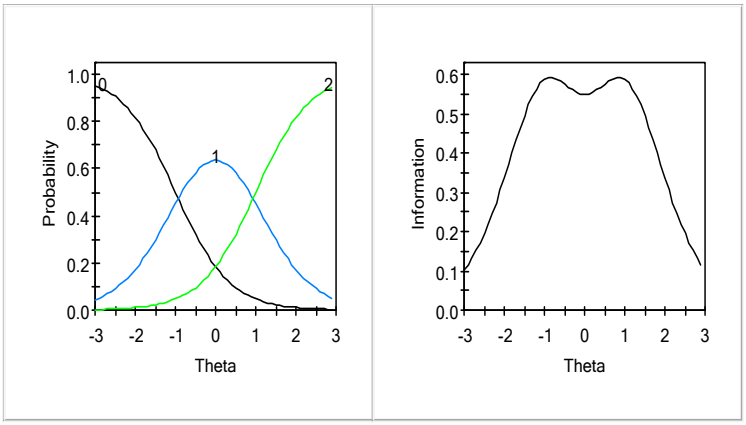


図 D.3　項目 2 の IRCCC（左）と項目情報量（右）

（注）a_2=1.5, b_{20}=-1, b_{21}=0, b_{22}=1, (D=1)

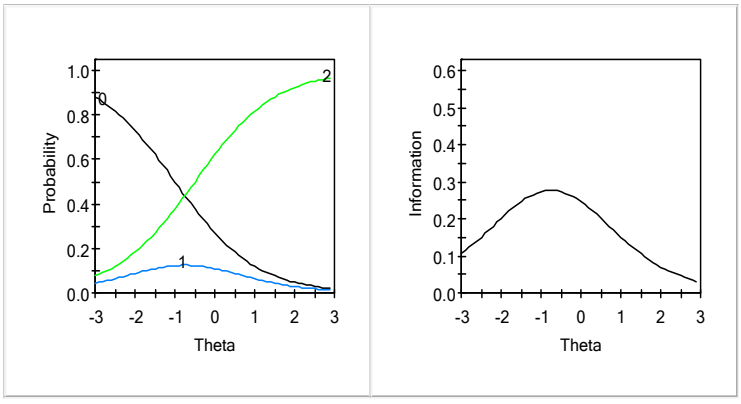


図 D.4　項目 3 の IRCCC（左）と項目情報量（右）

（注）a_3=1, b_{30}=-1, b_{31}=-0.75, b_{32}=-0.5, (D=1)

D.5　項目母数の推定

項目反応モデルに含まれる母数のうち、項目の特性を記述する母数をまとめて項目母数と呼ぶことがある。GR モデルの場合、識別力母数a_j 、BCC の位置母数b^*_{jk}、IRCCC の位置母数b_{jk}が項目母数に相当する。本節では、EM アルゴリズムによる周辺最尤推定法（marginal maximum likelihood estimation method: MMLE）を用いて GR モデルに含まれる項目母数を推定するための一般的な方法論を紹介する。なお、項目母数の推定についても、Baker and Kim (2004)が非常に詳しく参考になる。

いま、I人の受検者がそれぞれK個のカテゴリーをもつJ項目のテストを受検したとする。このとき、潜在特性値 θ をもつ受検者が反応パターンmをとる確率は、局所独立の仮定に注意すれば、

$$P(V_m|\theta) = \prod_{j=1}^{J}\prod_{k=1}^{K} P_{jk}(\theta)^{v_{mjk}} \tag{D.11}$$

と表現される。ここで、反応パターン行列V_mは、観測された反応パターンmを表現するためのJ行K列の大きさをもつ 2 値行列である。その要素はv_{mjk}であり、反応パターンmにおいて項目jのカテゴリーkが観測されたなら$v_{mjk} = 1$、それ以外は$v_{mjk} = 0$である。例えば、図 D.1（右）1 行目の段階反応の場合、反応パターン行列V_mは図 D.5 のように表現される。なお、J項目がそれぞれK個のカテゴリーをもつ状況では、反応パターンの組み合わせの数は$M = K^J$個になる。

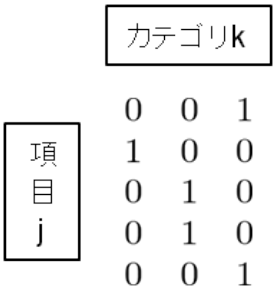


図 D.5　反応パターン行列V_mの例

項目母数を推定する際に局外母数となる θ を消去するため、潜在特性値の確率密度関数$g(\theta)$を用いて(D.11)式から θ を積分消去すると、反応パターンmが観測される（周辺）確率は、

$$P(V_m) = \int_{-\infty}^{\infty} P(V_m|\theta)g(\theta)d\theta \tag{D.12}$$

となる。通常、潜在特性値の確率密度関数$g(\theta)$には標準正規分布が仮定される。

項目母数を周辺最尤推定するため、テスト結果（データ）が得られたもとでの周辺尤度関数を記述し、その尤度関数を最大化するときの項目母数の値を周辺最尤推定値とする。反応パターンmをとる受検者の数をI_mとすると、多項分布$(I, P(V_m))$を用いて項目母数の尤度関数は、

$$L = \frac{I!}{\prod_{m=1}^{M} I_m!} \prod_{m=1}^{M} [P(V_m)]^{I_m} \tag{D.13}$$

となる。尤度関数の微分の操作を容易にするため、(D.13)式の両辺の対数をとると、項目母数の対数周辺尤度関数は、

$$\log L = \log I! - \log \sum_{m=1}^{M} I_m! + \sum_{m=1}^{M} I_m \log P(V_m) \tag{D.14}$$

となる。項目母数の周辺最尤推定値は、(D.14)式において項目母数についての 1 次偏導関数を 0 とおいた非線形連立方程式（周辺尤度方程式）を数値的に解くことによって得られる。

項目母数を周辺最尤推定する際、(D.12)式の積分を計算する必要がある。通常、区分求積法の 1 つである Gauss-Hermite 求積法などを用いて近似計算する。連続値である θ 上において、H 個の離散的な求積点$X_h (h = 1,2,\cdots,H)$とそれらの求積点に対応する重み$A(X_h)$を用いて、

$$\tilde{P}(V_m) = \sum_{h=1}^{H} \prod_{j=1}^{J} \prod_{k=1}^{K} \left[P_{jk}(X_h)\right]^{v_{mjk}} A(X_h) \tag{D.15}$$

と近似できる。

周辺尤度方程式は、EM アルゴリズムを用いて数値的に解くことができる。周辺尤度方程式は、求積点X_hにおいて項目jにカテゴリーkと反応する期待頻度r_{jkh}と求積点X_hにおける期待人数f_hを含む形に変形できる。EM アルゴリズムでは、項目母数の更新量が基準値未満になるなどの収束条件を満たすまで E ステップと M ステップが繰り返される。E ステップにおいて仮の項目母数を定め、

$$r_{jkh} = \sum_{m=1}^{M} \frac{I_m \sum_{j=1}^{J} \sum_{k=1}^{K} \left[P_{jk}(X_h)\right]^{v_{mjk} A(X_h) v_{mjk}}}{\tilde{P}(V_m)} \tag{D.16}$$

$$f_h = \sum_{m=1}^{M} \frac{I_m \sum_{j=1}^{J} \sum_{k=1}^{K} \left[P_{jk}(X_h)\right]^{v_{mjk}} A(X_h)}{\tilde{P}(V_m)} \tag{D. 17}$$

を計算する。M ステップでは、Newton-Raphson 法や Fisher のスコアリング法を用いて周辺尤度関数を最大化する。収束条件を満たさないならば、M ステップで更新された項目母数の推定値を仮の項目母数として E ステップに戻る。収束条件を満たしたときの計算結果が最終的な項目母数の推定値となる。なお、推定された項目母数の標準誤差は、推定が終了した時点での Fisher 情報行列の逆行列における対角要素の平方根として求められる。

ここまで、テストを構成するJ個の項目がそれぞれK個という同数のカテゴリーをもつ場合について記述してきた。同様の方針により、各項目のカテゴリー数が異なる場合や 2 PL モデルが混在する場合も項目母数の周辺最尤推定が可能である。各項目のカテゴリー数が異なる場合は、カテゴリー数Kを項目jに依存する変数K_jとして扱えばよい。2 PL モデルが混在する場合でも、カテゴリー数が 2 つの場合の GR モデルは 2 PL モデルと同等なので、$K_j = 2$と考えることによって同様の取り扱いが可能である。このとき、図 D.5 に示した反応パターン行列V_mの各行の列数は項目jによって異なることになる。

GR モデルを含む多値 IRT モデルの母数を推定可能なソフトウェアが無料あるいは有料で提供されている。代表的なソフトウェアとして、SSI 社（Scientific Software International, Inc.）の PARSCALE 4（Muraki & Bock 2003）がある。国内で開発されたフリーソフトウェアとしては、EasyEstimation（熊谷 2009）がある。また、SSI 社からは IRTPRO（Cai, Thissen, & du Toit 2011）がリリースされている。

D.6　GR モデルの情報関数

テストを実施するということは、テストを構成する項目を利用して受検者の潜在特性値θについての情報を得る行為であると解釈できる。その際、各項目を通して得られるθに関する Fisher 情報量を項目情報関数（item information function）と呼び、項目jの項目情報関数$I_j(\theta)$は、

$$I_j(\theta) = -E\left[\frac{\partial^2 log\, P_{jk}(\theta)}{\partial \theta^2}\right] = \sum_{k=0}^{K-1} \left\{ -\frac{\partial^2 log\, P_{jk}(\theta)}{\partial \theta^2} \right\} P_{jk}(\theta) = \sum_{k=0}^{K-1} I_{jk}(\theta) P_{jk}(\theta) \tag{D. 18}$$

と表現できる。Samejima(1969) は、(D.18)式の$I_{jk}(\theta)$を項目カテゴリーの情報関数、$I_{jk}(\theta)P_{jk}(\theta)$をカテゴリー$k$の情報量占有率と呼んだ。GR モデルでは、

$$I_{jk}(\theta) = -\frac{\partial^2 \log P_{jk}(\theta)}{\partial \theta^2} = \frac{\left[P'_{jk}(\theta)\right]^2 - P_{jk}(\theta)P''_{jk}(\theta)}{\left[P_{jk}(\theta)\right]^2} \tag{D.19}$$

$$I_{jk}(\theta)P_{jk}(\theta) = \frac{\left[P'_{jk}(\theta)\right]^2 - P_{jk}(\theta)P''_{jk}(\theta)}{\left[P_{jk}(\theta)\right]^2} P_{jk}(\theta) = \frac{\left[P'_{jk}(\theta)\right]^2}{P_{jk}(\theta)} - P''_{jk}(\theta) \tag{D.20}$$

となる。ただし、$P'_{jk}(\theta) = \partial P_{jk}(\theta)/\partial\theta$、 $P''_{jk}(\theta) = \partial^2 P_{jk}(\theta)/\partial\theta^2$である。(D.20)式を(D.18)式に代入して整理すれば、GR モデルの項目情報関数$I_j(\theta)$は、

$$I_j(\theta) = \sum_{k=0}^{K-1} \left\{ \frac{\left[P'_{jk}(\theta)\right]^2}{P_{jk}(\theta)} - P''_{jk}(\theta) \right\} \tag{D.21}$$

$$= \sum_{k=0}^{K-1} \left\{ \frac{\left[P'_{jk}(\theta)\right]^2}{P_{jk}(\theta)} \right\} - \sum_{k=0}^{K-1} P''_{jk}(\theta) \tag{D.22}$$

$$= \sum_{k=0}^{K-1} \frac{\left[P^{*\prime}_{jk}(\theta) - P^{*\prime}_{jk+1}(\theta)\right]^2}{P^*_{jk}(\theta) - P^*_{jk+1}(\theta)} - \sum_{k=0}^{K-1} \left[P^{*\prime\prime}_{jk}(\theta) - P^{*\prime\prime}_{jk+1}(\theta)\right] \tag{D.23}$$

$$= \sum_{k=0}^{K-1} \frac{\left[P^{*\prime}_{jk}(\theta) - P^{*\prime}_{jk+1}(\theta)\right]^2}{P^*_{jk}(\theta) - P^*_{jk+1}(\theta)} \tag{D.24}$$

となる。なお、(D.22)式から(D.23)式への計算には(D.1)式を利用している。また、(D.23)式の第 2 項は、 $k = 1,2,\cdots,K-1$の項は消去されること及び(D.3)式と(D.4)式から 0 である。

GR モデルの BCC として(D.2)式の 2 PL モデルを利用すれば、

$$P'_{jk}(\theta) = P^{*\prime}_{jk}(\theta) - P^{*\prime}_{jk+1}(\theta) = D_{aj}P^*_{jk}(\theta)Q^*_{jk}(\theta) - D_{aj}P^*_{jk+1}(\theta)Q^*_{jk+1}(\theta) \tag{D.25}$$

と計算できる。ただし、 $Q^*_{jk}(\theta) = 1 - P^*_{jk}(\theta)$である。このとき、項目情報関数$I_j(\theta)$は、

$$I_j(\theta) = D^2 a_j^2 \sum_{k=0}^{K-1} \frac{\left[P^*_{jk}(\theta)Q^*_{jk}(\theta) - P^*_{jk+1}(\theta)Q^*_{jk+1}(\theta)\right]^2}{P_{jk}(\theta)} \tag{D.26}$$

となる。

図 D.2（右）～図 D.4（右）に、(D.26)式によって描いた項目情報関数の例を示す。例示した 3 つの項目は、前節で利用した項目と同一である。図をみると、IRCCC の位置母数の近辺で情報量が大きいことや、その近辺での情報量は識別力母数の値が大きい項目のほうが大きいことがわかる。情報量の逆数が測定誤差の大きさと関係することから、テストを作成するときには、目的とする測定レベルに見合った困難度（位置母数：location parameter ともいう）の項目を用意することや、なるべく識別力の高い項目を用意することが大切である。

項目が 2 つのカテゴリーだけをもつ場合、(D.26)式の項目情報関数は 2 PL モデルのそれと一致する。項目情報量の点からも、2 PL モデルは 2 つのカテゴリーをもつ GR モデルと同等であることが確認できる。さらに、Samejima(1969) によれば、ある項目にカテゴリーを追加した場合、追加する以前と比べて同等かそれ以上の項目情報量が得られるということが証明されている。

Fisher 情報量の加法性から、局所独立の仮定を満たすテストの項目情報量をすべて加算するとテスト全体の情報量に相当する。これはテスト情報量と呼ばれ、その潜在特性値 θ の関数はテスト情報関数（test information function）と呼ばれる。$P_{jk}^{*}(\theta)$として(D.2)式の 2 PL モデルを利用すれば、テスト情報関数$I(\theta)$は(D.26)式の項目についての和となり、

$$I(\theta) = \sum_{j=1}^{J} I_j(\theta) \sum_{j=1}^{J} \sum_{k=0}^{K-1} D^2 a_j^2 \frac{\left[P_{jk}^{*}(\theta)Q_{jk}^{*}(\theta) - P_{jk+1}^{*}(\theta)Q_{jk+1}^{*}(\theta)\right]^2}{P_{jk}(\theta)} \qquad \text{(D. 27)}$$

と表現できる。

テスト情報量は、テストの測定精度と密接な関連がある。テスト情報量を用いると、潜在特性値の最尤推定量 $\hat{\theta}$ の標準誤差を $1/\sqrt{I(\hat{\theta})}$ として見積もることができる。それゆえ、テスト情報関数をみれば、テストがどの付近の潜在特性値をどのくらい正確に測定できるのかが具体的にわかる。テスト情報量の大きい尺度値レベルが潜在特性値をより正確に測定できる部分であり、テスト情報量の小さい尺度値レベルが潜在特性値の測定精度が低くなる部分である。

図 D.6 に、図 D.2～図 D.4 の 3 項目からテストが構成されている場合のテスト情報関数$I(\hat{\theta})$と最尤推定量 $\hat{\theta}$の標準誤差に相当する $1/\sqrt{I(\hat{\theta})}$ の曲線を示す。当該テストは、その受検者集団において平均的な尺度値レベルをもつ受検者の潜在特性値を他の尺度値レベルより正確に測定できることが読み取れる。

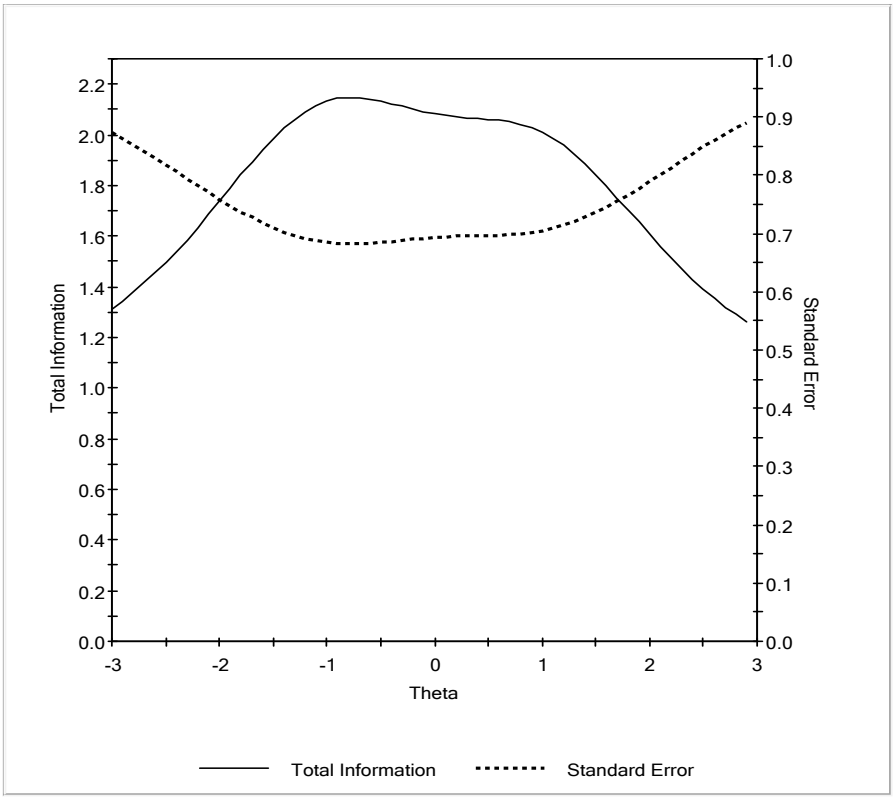


図 D.6 テスト情報量と標準誤差

付録Ｅ：令和３年度 経年変化分析調査 テクニカルレポート 執筆編集委員会

技術助言

柴山直（東北大学大学院教育学研究科教授、編集責任）

土屋隆裕（横浜市立大学データサイエンス学部教授）

佐藤喜一（九州大学アドミッションセンター教授）

熊谷龍一（東北大学大学院教育学研究科准教授）

宇佐美慧（東京大学大学院教育学研究科准教授）

事務局

浅原寛子（文部科学省総合教育政策局調査企画課学力調査室室長）

浦田晴香（文部科学省総合教育政策局調査企画課学力調査室専門官）

早野富美（文部科学省総合教育政策局調査企画課学力調査室専門職）

溝口朗央（文部科学省総合教育政策局調査企画課学力調査室係員）

令和３年度『全国学力・学習状況調査』

経年変化分析調査 テクニカルレポート

――

公開日：令和４年３月 28 日

編集：文部科学省総合教育政策局調査企画課学力調査室

※著作権について：

本テクニカルレポートを引用する際には出典を記載してください。出典の記載方法は以下の通りです。

（出典記載例）

令和３年度『全国学力・学習状況調査』経年変化分析調査 テクニカルレポート（文部科学省）