

## 測定精度と問題数の関係

—IRT を活用した学力調査の在り方について—

The Relationship between Item Size and Measurement Accuracy  
in Ability Assessment using IRT

褓岩 晶\*

HOROIWA Akira

## Abstract

This paper discusses the relationship between item size (test length) and the measurement accuracy of individual scores using the Item Response Theory in ability assessment. Further, it examines the effect of a multiple-form design, leading to the extension of item size, on measurement accuracy compared to a single-form design. Two simulation studies in this paper are conducted by modifying the script of the simulation studies in the research project “Study on a framework for academic ability assessment (FY2021 - FY2023)” by the National Institute for Educational Policy Research.

First, I provide an overview of chapter 4 in the report of the research project “Study on a framework for academic ability assessment,” especially the explanation of the relationship between item size and measurement accuracy. Next, previous studies on this topic are reviewed. Most of these studies did not examine the relationship between item size and measurement accuracy of individual scores but explored that between sample size and measurement accuracy of item parameters. Distinguishing between test length and total item size, two simulation studies show the relationship between test length, total item size, and the measurement accuracy of individual scores.

---

\* 教育データサイエンスセンター 総括研究官

## 1 はじめに

国際的な学力調査である「国際数学・理科教育動向調査 (TIMSS)」(国立教育政策研究所 2021) や「OECD 生徒の学習到達度調査 (PISA)」(国立教育政策研究所 2024a) では、受検者の能力を項目反応理論 (Item Response Theory、以下 IRT と略す) を用いて測定しており、また文部科学省が行っている「全国学力・学習状況調査」の「経年変化分析調査」(文部科学省・国立教育政策研究所 2022) においても IRT で受検者の得点が算出されている。学力調査における能力の推定や得点の算出において、IRT は必要不可欠な測定技術となっており、「全国学力・学習状況調査」の本体調査でもその導入が検討されている (全国的な学力調査の CBT 化検討ワーキンググループ 2021: 7)。

本稿は、国立教育政策研究所が行ったプロジェクト研究「学力アセスメントの在り方に関する調査研究 (令和 3~5 年度)」(国立教育政策研究所 2024、以下「学力アセスメント」と略す) で行われたシミュレーション研究を応用して、学力調査で得点の推定に IRT を使う場合、問題数と得点の測定精度にはいかなる関係があるか、調査問題をどの程度用意すれば測定精度がよくなるかを議論する。受検者が解答する問題数は、多ければ多いほど得点の測定精度は高くなるが、解答時間が長いと受検者は疲労し、能力を発揮できない可能性があり、さらに調査を学校で実施する場合、スケジュールの関係等で調査のための解答時間を十分にとれない場合がある。限られた時間の範囲内で、少ない問題数の違いがどの程度の測定精度の違いにつながるのか、これを把握することが学力調査の調査設計や作問をする上で重要になる。

また、受検者全員で同じ問題冊子を 1 種類だけ使うのではなく、問題が一部共通する複数の冊子を使うことで問題数を拡張し、1 回の調査で様々な問題を出題することが TIMSS、PISA、そして経年変化分析調査で行われている。このような出題方法は、「item-sampling」(Lord 1962) や「重複テスト分冊法」(日本テスト学会 2010: 45) と呼ばれるが、このような問題数の拡張が得点の測定精度にどのような影響を及ぼすのかも検証する。

本稿では最初に、プロジェクト研究「学力アセスメント」の報告書第 4 部の内容、すなわち同研究の測定技術班が行った議論やシミュレーション研究について概観し、同研究の報告書で測定精度と問題数の関係がどのように述べられているのかを説明する。次に、IRT を使用する際の問題数について、これまでの先行研究でどのような議論がなされてきたか概観する。先行研究のほとんどは、得点算出の前提となる項目パラメータの測定精度と問題数、受検者数の関係を問うものだが、学力調査に関わる調査主体や作問者が求める得点の測定精度といった視点から、本稿で行うシミュレーション研究の必要性を明らかにする。そして、学力調査における問題数を受検者個人への「出題数」と調査全体で使用する「総問題数」とに分けて捉え、それぞれに対するシミュレーション研究を通して、得点の測定精度と問題数の関係を検証する。

本稿のシミュレーション研究では、IRT の数理モデル (項目反応モデル) として、Birnbaum (1968: 400) の 2 パラメータ・ロジスティックモデル (two-parameter logistic model) を使用する。このモデルを de Ayala (2009: 100) は次のように表している。

$$P(x_j = 1 | \theta, \alpha_j, \delta_j) = \frac{e^{\alpha_j(\theta - \delta_j)}}{1 + e^{\alpha_j(\theta - \delta_j)}}$$

$x_j$  は  $j$  という問題 (項目) に対する反応を意味し、1 ならば正答、0 ならば誤答を表しており、 $P$

は  $\theta$ 、 $\alpha_j$ 、 $\delta_j$  が決まっているときに  $x_j$  が 1、つまり問題  $j$  に正答するときの正答確率を表している。 $P$  の条件であり、右辺にも使われている  $\theta$ 、 $\alpha_j$ 、 $\delta_j$  は、 $\theta$  が受検者の能力の値であり、IRT で算出される得点、 $\alpha_j$  が識別力パラメータ (discrimination parameter) と呼ばれる問題  $j$  が  $\theta$  の違いをどれくらい判別するのか、 $\theta$  とどのくらい関係性が強いのかを表す値、 $\delta_j$  が困難度パラメータ (difficulty parameter) と呼ばれる問題  $j$  の難易度を示す値である。2 パラメータ・ロジスティックモデルでは、 $\alpha_j$  と  $\delta_j$  の 2 つのパラメータで問題  $j$  が特徴づけられる。この他、右辺には  $e$  が使われているが、これはネイピア数と呼ばれる定数 ( $e=2.718\dots$ ) である。IRT では、受検者の解答データ (受検者と問題の組合せごとに 1: 正答、0: 誤答の値が入る) に上記の式を当てはめ、各問題の項目パラメータの値を推定し、それから各受検者の能力値、つまり得点を算出する。IRT には 2 パラメータ・ロジスティックモデル以外にも様々なモデルがあるが、巖倉ら (2019: 284-295) は、1 パラメータ・ロジスティックモデル、3 パラメータ・ロジスティックモデルとの比較をシミュレーションデータで行い、学力調査で IRT を使用するのであれば、2 パラメータ・ロジスティックモデルを推奨している。

本稿で行ったシミュレーション研究は、プロジェクト研究「学力アセスメント」の報告書第 4 部付録 (国立教育政策研究所 2024c) で使われたものを一部変更して使っている。シミュレーションに必要な統計ソフト「R」(R core team 2024) のスクリプトも同付録の中で公開されており、それらは本稿と同様の研究だけでなく、簡単な変更だけで IRT を学力調査に導入する際の様々な検証に利用可能なものである。

## 2 プロジェクト研究と測定技術班

プロジェクト研究「学力アセスメント」の研究組織は、研究代表者の下で事務局、総括班、現況分析班、作問・結果分析班、測定技術班、データサイエンス班に分かれており、そこでの測定技術班の役割は、全国学力・学習状況調査のコンピュータ使用型調査 (CBT) への移行とそれに伴う IRT の導入という条件のもとで、「問題バンクの構築・運用の在り方」、IRT の視点を加えた「結果分析の枠組み」、「作問の枠組み」を検討することであった (国立教育政策研究所 2024b: 100)。測定技術班は、国立教育政策研究所内外のテスト理論の専門家から構成されており、学力測定の技術的側面に特化した形で、IRT を活用した学力調査の在り方を議論し、その成果は同研究の報告書第 4 部にまとめられている。

同報告書の第 4 部第 1 章では、全国学力・学習状況調査の目的を「指導改善」と「政策の効果の検証」に分け、それに応じた測定すべき内容を整理し、第 2 章では、全国学力・学習状況調査の毎年悉皆で行われる「本体調査」と、3 年に 1 度、標本調査として行われる「経年変化分析調査」とを調査目的と測定内容のどの部分を担うものと位置づけるのか、両調査を経年比較可能にするためには調査問題の作問、公開・非公開をどのように行っていくべきか、つまり問題バンクの構築・運用の在り方が検討され、実現可能性や長所短所を加味して複数の案を提示している。第 3 章は、結果分析の枠組みとして、調査後のデータ処理、古典的テスト理論 (IRT 以前のテスト理論) による分析、IRT による分析、調査結果のフィードバックで行うべき事項と流れが解説されている。この章では、教科別の得点以外に、それぞれの教科で下位分野の得点化が可能かどうか検討するため、第 4 部付録にある「付録 4 下位尺度の推定に関するシミュレーション」(国立教育政策研究所 2024c: 165-170) が行われた。このシミュレーション研究で使われたスクリプトを変更して、本稿の

シミュレーション研究を行う。第4章は、IRTを用いる場合の作問の枠組みについて議論しており、特に測定技術の側面から、問題フォーム（受検者に調査問題を出題する際の問題の組合せ方、後述する問題冊子とほぼ同じ意味）とその配信方法（同じ問題フォームを同じ日に配信するのか、学校単位で配信するか、それとも受検者ごとにランダムに配信するか）が検討されている。第4章の「1-3 問題の難易度に関わる作問上の注意点」では、「付録5 調査問題の困難度に関するシミュレーション」（国立教育政策研究所 2024c: 171-176）が行われ、問題の難易度によって受検者の得点の測定精度がどのように変わるか検証しており、「第2節 問題フォームの組み方」では、「付録6 複数フォームの影響に関するシミュレーション」（国立教育政策研究所 2024c: 177-182）が行われ、複数の問題フォームを使用した場合でもIRTで推定した調査結果に違いが見られないことを明らかにした。ただし、同節の「付録7 フォーム組みの影響に関するシミュレーション」（国立教育政策研究所 2024c: 183-191）では、フォームの組み方として後述する「釣合い型不完備ブロック計画」とそれ以外とが比較され、前者の推定結果が僅かではあるが良好であることが示された。

では、プロジェクト研究「学力アセスメント」では、本稿のテーマである測定精度と問題数の関係についてどのように述べられているのか。1つ目は、「付録4 下位尺度の推定に関するシミュレーション」に関連するところで、教科の下位分野を尺度化する際、下位分野が多くなるほど1つの尺度に関わる問題数が少なくなるため、下位分野の得点の測定誤差が大きくなるという点である（国立教育政策研究所 2024b: 130, 137）。後述する先行研究にもあるとおり、測定に使われる問題数が少なくなると、得点の測定精度が下がることを意味する。付録4では、「問題数の影響の検証」として4問、8問、16問、32問、64問の場合について調べているが、シミュレーションデータである真の得点とIRTで推定した得点とのRMSE（root mean square error、一人一人の真の得点と推定得点との差を2乗した値を平均し、その平方根を求めた値）を使っており、受検者集団全体に対して1つの値で評価している。本稿では、このシミュレーションを応用し、全体とは別に、受検者を得点で区分し、それぞれの区分ごとに問題数が測定精度に与える影響を明らかにする。2つ目は、「付録6 複数フォームの影響に関するシミュレーション」に関連するところで、複数の問題フォーム（問題冊子）を使った場合に単独の問題フォームと結果が異なるのかを検証している。その結論は、受検者の得点、その平均等の精度に大きな違いは見られないというものであった（国立教育政策研究所 2024b: 144）。このシミュレーションでは、1受検者の解く問題数が16問のときのみを扱い、総問題数がそれと同じになるフォーム組みと、1受検者の解く問題数は同じだが総問題数が1.25倍、1.75倍、3.25倍になるフォーム組みとを比較している。本稿では1受検者の解く問題数も変更しながら検証する。

受検者個人の得点の測定精度と問題数との関係は、上記のようにプロジェクト研究「学力アセスメント」では2か所しか扱われていない。測定技術班の目的が、IRTの導入に特化したものであったため、シミュレーション研究で測定精度を扱うときでも、どの方法が学力調査に適しているのか調べるために用いられており、受検者個人の得点の測定精度を見る基準も全体でのRMSEの値のみになっている。既に述べたように、1人の受検者が解く問題数が多ければ多いほど、推定される得点の精度は高くなるが、学力調査を行う場合、調査時間の制約があることは避けて通れない問題であり、限られた時間の中で受検者に出題できる問題数には限界がある。精度を上げるために調査時間を長くできても、後になるほど受検者の疲労がたまり、結果が不正確になると考えられる。そのため、目標となる測定精度を設定してから問題数を決めることは現実的ではなく、できるだけ多くの問題を限られた時間の中で出題するしかない。ただし複雑な問題、時間をかけて取り組む問題を

出題したいとき、どの程度の問題数になると、どの程度の精度になるのか知ることは、学力調査を計画する上で重要である。プロジェクト研究の報告書の中では、それに答えられなかったが、本稿で行うシミュレーション研究はそれを補うものになっている。

### 3 問題数と受検者数に関するこれまでの研究

ここでは IRT を使用する際の問題数と測定精度の関係に触れている先行研究を整理する。多くの研究は、問題数だけでなく、受検者数も合わせて検討している。この種の研究は、受検者の能力の測定精度よりも、能力測定に先行する項目パラメータの推定が正確にできる条件を検証しており、その条件の 1 つとして問題数や受検者数が検証されている。

例えば最も古いものとしては、Lord (1968) による Birnbaum の 3 パラメータ・ロジスティックモデルを実際の大学入試データに当てはめた研究がある。この研究では、テストの情報量 (精度) がテストの長さ (問題数) と比例することや (Lord 1968: 1009)、無答を誤答とすることが何らかの不正確さを、特にテストを終えられなかった受検者に対して生んでいることなどが指摘されており (Lord 1968: 1009)、最初期の IRT 研究ではあるが、現在でも検討すべき内容を含んでいる。その中で Lord は、問題数が 50 問以上、受検者が 1000 人以上でないと識別力パラメータの誤差が大きくなると指摘している (Lord 1968: 1016)。50 問という問題数は、日本の全国学力・学習状況調査の国語や算数・数学では出題するのが難しい量であるが、当時は現在の IRT で使われているような周辺最尤推定法 (Bock and Lieberman 1970) や EM アルゴリズム (Bock and Aitken 1981) がなかったこと、IRT の数理モデルが調査問題を 3 つのパラメータで特徴づけるもの (困難度パラメータ、識別力パラメータとともに、選択問題で偶然正答する可能性を考慮した当て推量パラメータを含む) であったことが関係していると考えられる。

Hulin ら (1982) による最小のサンプルサイズ (受検者数) とテストの長さ (問題数) を検証した研究では、15 問、30 問、60 問の問題に対して 200 人、500 人、1000 人、2000 人の受検者からなるシミュレーションデータを作り、項目パラメータと能力値 (得点) の推定精度を調べている。Hulin らによれば、2 パラメータ・ロジスティックモデルであれば 30 問、500 人が必要、3 パラメータ・ロジスティックモデルでは 60 問、1000 人が必要とされている (Hulin et al. 1982: 249)。ただし 2 パラメータ・ロジスティックモデルについては、60 問、200 人でも正確な推定が行えるとしており (Hulin et al. 1982: 259)、受検者数と問題数は一方を増やせばもう一方を減らせるとされている。Hulin らの論文では述べられていないが、Lord と同様、周辺最尤推定法や EM アルゴリズムが使われていない可能性があり、現在の IRT よりは最小の受検者数と問題数が大きいように思われる。

Drasgow (1989) は、周辺最尤推定法とそれ以前から用いられていた同時最尤推定法とを比較したシミュレーション研究を行っており、そこでは 5 問、10 問、15 問、25 問の問題に対して、200 人、300 人、500 人、1000 人からなるシミュレーションデータが用いられている。Drasgow は、周辺最尤推定法が同時最尤推定法よりも、かなり少ない問題数、小さいサンプルサイズでも正確な推定を行えるとして、2 パラメータ・ロジスティックモデルの場合、少なくともサンプルサイズ 500 人、問題数 10 問は必要であるとしている (Drasgow 1989: 85)。先述の Hulin らの結果と比べて必要な問題数がかなり少なくなっているが、項目パラメータが極端な値でなければ、200 人、5 問でも偏りのない、誤差の小さい項目パラメータの推定ができるとも述べている (Drasgow 1989: 88)。また、Drasgow は、受検者の能力分布と困難度パラメータの分布が近いほど測定の精度が向上するとしており

(Drasgow 1989: 81)、本稿のシミュレーション研究でもこの指摘に沿った困難度パラメータの設定を行う。

最小の受検者数と問題数を検証したものではないが、周辺最尤推定法と EM アルゴリズムを使う際の能力値の事前分布がその事後分布に与える影響を検証した Seong (1990) の研究では、45 問で 100 人と 1000 人からなるシミュレーションデータが用いられ、能力 (得点) の推定値は受検者数を変えても劇的には変わらないこと (Seong 1990: 308)、一方、受検者数が増えると困難度パラメータと識別力パラメータの測定精度が向上すること (Seong 1990: 310) が指摘されている。同じように事前分布の影響を検証した Stone (1992) の研究では、10 問、20 問、40 問の問題に対して、250 人、500 人、1000 人からなるシミュレーションを行い、2 パラメータ・ロジスティックモデルの場合、特に 10 問から 20 問にかけては識別力パラメータの推定に問題数が影響を与えているが、困難度パラメータにはこの傾向が見られず (Stone 1992: 5)、困難度パラメータの推定値は 250 人、10 問でも正確かつ安定しているが、識別力パラメータの推定値はそれ以上の規模でないと正確かつ安定しないとされている (Stone 1992: 12)。そして、推定された能力分布のバイアス (偏り) は、サンプルサイズではなく問題数が増えるほど小さくなり、Seong と同様、能力の推定値にはサンプルサイズが影響しておらず (Stone 1992: 11-12)、問題数が能力推定に最も有意な影響要因であるとされている (Stone 1992: 15)。

シミュレーション研究ではないが、Şahin と Anıl (2017) は、50 問、6288 人が受けた実際の言語テストからデータを抽出し、10 問、20 問、30 問の問題に対して 150 人、250 人、350 人、500 人、750 人、1000 人、2000 人、3000 人、5000 人からなるサンプルを作成し、1 パラメータ・ロジスティックモデル、2 パラメータ・ロジスティックモデル、3 パラメータ・ロジスティックモデルを用いた際の項目パラメータの推定精度を調べている。いずれの場合も周辺最尤推定法が使われており (明言されていないが EM アルゴリズムも使われていると思われる)、その結果から、1 パラメータ・ロジスティックモデルは問題数に関係なく 150 人、2 パラメータ・ロジスティックモデルは 10 問のときは 750 人、20 問の時は 500 人、30 問で 250 人、3 パラメータ・ロジスティックモデルは 10 問と 20 問で 750 人、30 問で 350 人の規模が必要としている (Şahin and Anıl 2017: 329)。Drasgow や Stone とは異なり、実際のデータを使っているため、必要最小限度とされている問題数、受検者数が大きくなっているが、実際の学力調査で IRT を用いる際は、この程度の問題数と受検者数が必要になると考えられる。

なお、問題数と測定精度の関係について触れているものではないが、大友 (1996: 98) は、1 パラメータ・ロジスティックモデルは 100 から 200 人、2 パラメータ・ロジスティックモデルは 200 から 400 人、3 パラメータ・ロジスティックモデルは 1000 から 2000 人の受検者がいれば適切な計算が可能であるとしており、豊田 (2012: 35) も、1 パラメータ・ロジスティックモデルは 100 人以上、2 パラメータ・ロジスティックモデルは 300 人以上、3 パラメータ・ロジスティックモデルは 1000 人以上の被験者が必要であるとしている。

以上、先行研究でどのような議論がなされてきたのかを概観したが、そのほとんどは、能力測定や得点算出の前提となる項目パラメータの推定精度と問題数、受検者数の関係を問うものであった。プロジェクト研究「学力アセスメント」や本稿が対象としているような全国規模の学力調査の場合、1 問当たりの受検者数が 1000 人を下回ることはなく、受検者 1 人が解答する問題数も国語、算数・数学、理科、英語等の教科であれば 10 問を下回ることはないため、項目パラメータの推定に支障を来すことはほとんどない。例えば、令和 6 年度全国学力・学習状況調査の本体調査では、小学校の

集計対象児童数は 960,389 人、中学校の集計対象児童数は 904,048 人であり、調査問題数も小学校の国語が 14 問、算数が 16 問、中学校の国語が 15 問、数学が 16 問であり（文部科学省・国立教育政策研究所 2024: 2-3）、項目パラメータの推定に必要とされる受検者数と問題数を上回っている。ただし、出題者の側には、長文の記述問題、コンピュータ使用型調査でシミュレーションを行うような問題、さらには学習指導要領で言うような「主体的・対話的で深い学び」を問う問題といった解答に時間のかかる問題を出題したいというニーズが一定程度存在するため、特に少ない問題数と測定精度との関係を明らかにすることは、解答に時間のかかる問題をどれくらい出題するか検討する上で重要な情報と考えられる。また、重複テスト分冊法のような複数の受検者に異なる問題を出題する方法を使えば、時間のかかる問題を 1 人の受検者に出題する数が少なくても、調査全体としては多く試すことができる。その際に、重複テスト分冊法による問題数の拡張が測定精度にどのような影響を及ぼすかを検証することも重要になる。

#### 4 2つのシミュレーション研究

本稿で行うシミュレーション研究は、学力調査の調査主体や作問者にとって重要な「受検者の得点の測定精度」を対象として、問題数が少ない場合の測定精度と、重複テスト分冊法を用いて調査で使われる総問題数を拡張した場合の測定精度を検証する。前節の最後で述べたように、日本の学力調査の文脈では、項目パラメータの推定が困難になるような受検者数は考えられず、本稿のシミュレーションでも受検者数を全国学力・学習状況調査の全体調査に合わせて 100 万人とするため、先行研究が行ってきたような項目パラメータの検証は行わない。また、先行研究では受検者全体の測定精度を RMSE で評価していたが、本稿のシミュレーションでは、全体だけでなく、受検者の得点によって測定精度が異なる可能性も加味して、受検者を「真の得点」で 7 つに分け、その区分ごとの測定精度も明らかにする。これまでの先行研究で行われなかった区分ごとの測定精度を示すことで、IRT や統計学に余り詳しくない調査主体や作問者でも、問題数と測定精度の関係や、受検者の得点と測定精度の関係を容易に理解でき、学力調査で使用する問題数から測定精度を大まかに把握できるようになる。

シミュレーションの説明の前に、用語の整理を行う。まず問題数について、受検者個人への「出題数」と調査全体で使用する「総問題数」を区別して用いる。1 つ目のシミュレーション研究では、受検者個人への「出題数」の違いを検証するが、2 つ目のシミュレーション研究では受検者個人への「出題数」を一定にし、「総問題数」を変えながら検証する。次に「真の得点」について、先行研究の多くは標準正規分布（平均 0、標準偏差 1）から無作為抽出した値を用いているが、本稿のシミュレーションでは標準正規分布により近いものとするため、標準正規分布から 100 万 2 個のパーセンタイル値（パーセンタイル値の前提となるパーセントは等間隔）を求め、最初の値（0 パーセンタイルで、値は $-\infty$ になる）と最後の値（100 パーセンタイルで、値は $\infty$ にある）を除いた 100 万人分の値を標準化（平均 0、標準偏差 1 に変換）して受検者の真の得点とする。最小値は $-4.75\dots$ 、最大値は $4.75\dots$ であり、真の得点の分布として、受検者を $-2.5$ 未満、 $-2.5$ 以上 $-1.5$ 未満、 $-1.5$ 以上 $-0.5$ 未満、 $-0.5$ 以上 $0.5$ 未満、 $0.5$ 以上 $1.5$ 未満、 $1.5$ 以上 $2.5$ 未満、 $2.5$ 以上の 7 区分に分けたときの受検者割合を図 1 に示す。

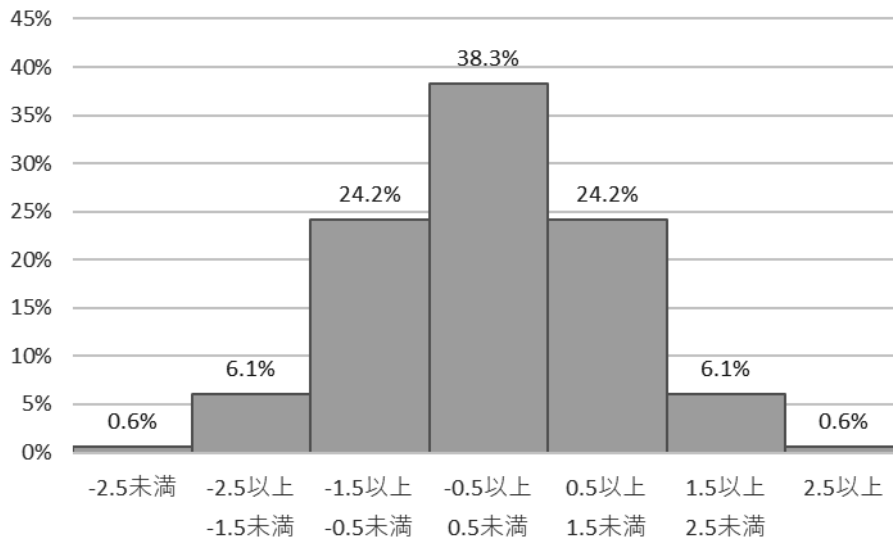


図1 真の得点の分布（7区分ごとの受検者割合）

「受検者の得点の測定精度」については、バイアスと誤差に分けて調べる。バイアス（偏り）は、項目反応理論で算出された受検者の得点の推定値が真値（真の得点）に対してどの方向に（正の方向又は負の方向）、どの程度ずれているかを示す値であり、受検者それぞれについては、以下の式で求められる。

$$\text{バイアス} = \frac{\sum \text{シミュレーション回数}(\text{推定値} - \text{真値})}{\text{シミュレーション回数}}$$

誤差（標準誤差とも呼ばれる）は、項目反応理論で算出された受検者の得点の推定値が真値（真の得点）から平均してどの程度離れているかを示す値であり、受検者それぞれについては、以下の式で求められる。バイアスは「推定値-真値」の平均、誤差は「推定値-真値」の標準偏差に対応している。

$$\text{誤差} = \sqrt{\frac{\sum \text{シミュレーション回数}(\text{推定値} - \text{真値})^2}{\text{シミュレーション回数}}}$$

2つのシミュレーション研究に共通することとして、シミュレーションデータのつくり方と真の項目パラメータの決め方を説明する。シミュレーションは、各条件の下で100回行う。それぞれのシミュレーションでは、受検者の真の得点と問題の真の項目パラメータの値に基づいて架空の解答データ（正答なら1、誤答なら0が入った受検者数×問題数の行列）を作成し、この解答データに対して、第1節で述べたように2パラメータ・ロジスティックモデルを当てはめ、各問題の項目パラメータを推定してから、各受検者の得点を算出する。そして、真の得点とシミュレーションから求めた100回分の推定値を使って、各受検者のバイアスと誤差を計算する。問題の真の困難度パラメータについては、先述した Drasgow（1989）の指摘に従って、標準正規分布から問題数プラス2個

のパーセンタイル値（パーセンタイル値の前提となるパーセントは等間隔）を求め、最初の値と最後の値（ $-\infty$ と $\infty$ ）を除く問題数分の値を使用する。例えば、4問の場合は、 $-0.84\dots$ 、 $-0.25\dots$ 、 $0.25\dots$ 、 $0.84\dots$ 、16問の場合は、 $-1.59\dots$ 、 $-1.19\dots$ 、 $-0.93\dots$ 、 $-0.72\dots$ 、 $-0.54\dots$ 、 $-0.38\dots$ 、 $-0.22\dots$ 、 $-0.07\dots$ 、 $0.07\dots$ 、 $0.22\dots$ 、 $0.38\dots$ 、 $0.54\dots$ 、 $0.72\dots$ 、 $0.93\dots$ 、 $1.19\dots$ 、 $1.56\dots$ になる。真の識別力パラメータの値は、検証を単純にするため、すべて1とする。つまり架空の解答データは実質的には1パラメータ・ロジスティックモデル（識別力パラメータの値がすべて同じ）に従っているが、測定自体は2パラメータ・ロジスティックモデルで行う。得点の推定値にはEAP（事後期待値、*expected a posteriori*）を使用する（Bock and Aitken 1981: 448）。IRTにおける得点の推定方法には、MLE（最尤推定値、*maximum likelihood estimate*）やWLE（重み付き最尤推定値、*weighted likelihood estimate*）のような最尤推定法を用いるものと、MAP（事後確率最大推定値、*maximum a posteriori*）やEAPのようなベイズ推定法（ベイズの定理を使って、事前分布と尤度関数から事後分布を導き出す）を用いるものがあるが、BockとMislevy（1982: 443）は、母集団全体でのRMSEでEAPよりも小さくなる推定値はないとしている。また、袈岩ら（2019: 257-261）は、受検者4000人、出題数40問と200問のシミュレーション研究を行い、MLE、WLE、EAP、PV（*plausible value*、PISAやTIMSSで母集団特性を推定する際に用いる得点）のRMSEを比較し、EAPが最小になること、つまり測定精度が高いことを明らかにしており、プロジェクト研究「学力アセスメント」（国立教育政策研究所 2024b: 130）でも、WLE、EAP、PVのRMSEを比較するシミュレーションを行い、EAPが最小になることを検証している。本稿のシミュレーション研究は、EAPを推定する際の事前分布に標準正規分布を用いており、これは受検者の真の得点と同じ分布であり、かつ困難度パラメータの真値も標準正規分布に従っているため、理想的な条件下で測定精度が7つの区分ごとにどのようになるのか調べている。現実の調査では今回のシミュレーション研究よりもバイアスや誤差は大きくなることが予想される。

### （1）受検者個人への出題数が少ない場合の測定精度

最初のシミュレーション研究では、出題数を4問、8問、12問、16問、20問、40問、60問と変えながら、受検者個人の得点の測定精度を検証する。プロジェクト研究「学力アセスメント」で使われた「付録4 下位尺度の推定に関するシミュレーション」（国立教育政策研究所 2024c: 165-170）のスク립トを一部変更して使用する。架空の解答データの作成にはRのパッケージ「*mirt*」（Chalmers 2012）の関数「*simdata()*」を、2パラメータ・ロジスティックモデルによる推定には同パッケージの関数「*mirt()*」を使用する。

図2では、全受検者100万人について、4問、16問、60問のときの真値とバイアスとの関係を散布図で示す。また、4問、8問、12問、16問、20問、40問、60問の各条件について、全体のバイアスの平均値と、図1で示した7区分のバイアスの平均値を表1に示す。

図2を見ると、いずれの問題数でも平均である0に近づくほどバイアスの絶対値が小さくなっているが、4問でのバイアスの最大（最小）値が3.66...（-3.66...）なのに対し、16問では2.61...（-2.61...）、60問では1.65...（-1.65...）となっている。今回のシミュレーション研究では、調査問題の数によって、困難度パラメータの最大値と最小値が決まり、測定可能な最低得点と最高得点もそこから決まってくる。真の得点の低い方では測定可能な最低得点との差が正の方向に大きく、真の得点の高い方では測定可能な最高得点との差が負の方向で大きくなるため、バイアスの絶対値がかなり大きくなっている。なお、得点の推定値にMLEを使用した場合、全問誤答の場合は得点が $-\infty$ 、全問正答の場合は得点が $\infty$ になるため、図2のEAPの場合とは異なり、真の得点の低い方ではバイア

スが負の値（最小で $-\infty$ ）になり、真の得点の高い方では正の値（最大で $\infty$ ）になり、誤差が大きくなるだけでなく、今回のシミュレーション研究で用いるバイアスと誤差が計算できない場合もある。

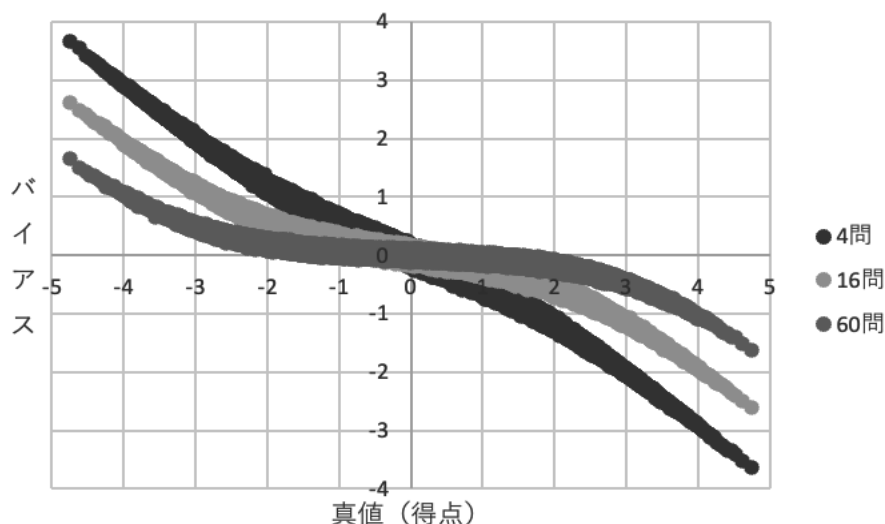


図2 真値とバイアスの関係（4問、16問、60問）

表1を見ると、全体のバイアスの平均値は0であり、母集団の平均得点を推定するのにどの出題数でも支障がないとわかるが、得点の7区分で見ると両端は60問であってもバイアスが0.43と-0.43となっている。この値をPISAやTIMSSのような平均500、標準偏差100の得点に換算すると43点と-43点、偏差値に換算すると4点と-4点のバイアスがあることになる。得点の7区分の-1.5以上-0.5未満（0.5以上1.5未満）に注目すると、4問ではバイアスが0.48（-0.48）であるが、問題数が増えると16問で0.19（-0.19）、40問で0.09（-0.09）となっており、PISA換算で48点、19点、9点、偏差値換算で5点、2点、1点とバイアスが小さくなっていることがわかる。

表1 各出題数でのバイアスの平均値（全体、7区分、小数点以下3桁で四捨五入）

	4問	8問	12問	16問	20問	40問	60問
全体	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-2.5未満	1.85	1.44	1.20	1.03	0.91	0.58	0.43
-2.5以上-1.5未満	1.08	0.77	0.61	0.50	0.42	0.24	0.17
-1.5以上-0.5未満	0.48	0.32	0.24	0.19	0.16	0.09	0.06
-0.5以上0.5未満	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.5以上1.5未満	-0.48	-0.32	-0.24	-0.19	-0.16	-0.09	-0.06
1.5以上2.5未満	-1.08	-0.78	-0.61	-0.50	-0.42	-0.24	-0.17
2.5以上	-1.85	-1.44	-1.20	-1.03	-0.91	-0.58	-0.43

図3では、全受検者100万人について、4問、16問、60問の時の真値と誤差との関係を散布図で示す。表2では、4問、8問、12問、16問、20問、40問、60問の各条件について、全体の誤差の平均値と、7区分の誤差の平均値を示す。

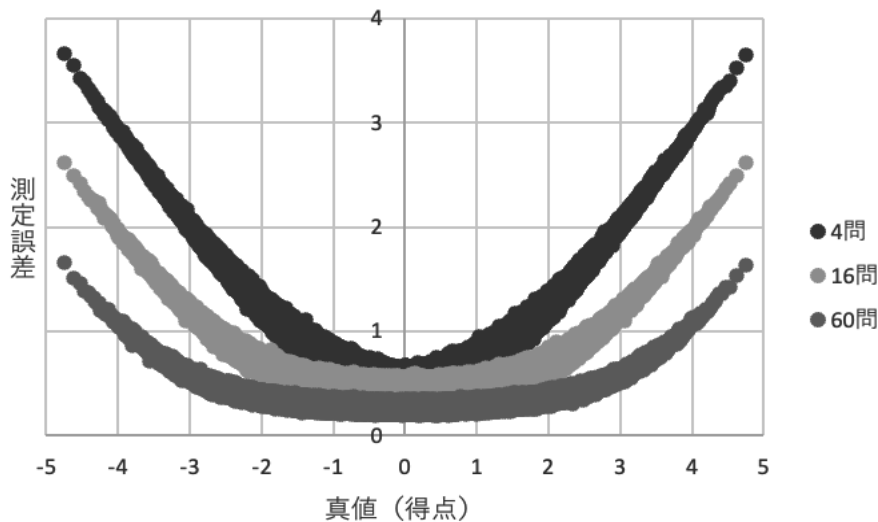


図3 真値と誤差の関係（4問、16問、60問）

図3を見ると、いずれの出題数でも平均である0に近づくほど誤差が小さくなっているが、バイアスとは異なり完全に0にはならない。差の2乗平均の平方根であるため、推定値と真値に違いがあれば、誤差が0になることはない。また、バイアスの絶対値が大きいところでは、推定値と真値の差が大きいため、誤差も大きくなっている。

表2 各出題数での誤差の平均値（全体、7区分、小数点以下3桁で四捨五入）

	4問	8問	12問	16問	20問	40問	60問
全体	0.71	0.61	0.54	0.49	0.46	0.35	0.29
-2.5未満	1.87	1.48	1.25	1.09	0.97	0.67	0.53
-2.5以上-1.5未満	1.15	0.88	0.74	0.65	0.58	0.42	0.35
-1.5以上-0.5未満	0.70	0.59	0.53	0.48	0.44	0.34	0.29
-0.5以上0.5未満	0.54	0.52	0.48	0.44	0.41	0.32	0.27
0.5以上1.5未満	0.70	0.59	0.53	0.48	0.44	0.34	0.29
1.5以上2.5未満	1.15	0.88	0.74	0.65	0.58	0.42	0.35
2.5以上	1.87	1.48	1.24	1.09	0.97	0.67	0.53

表2にある全体の誤差の平均値は4問のときに0.71、それを出題数間で比較すると4問から8問で0.1、8問から12問で0.07、12問から16問で0.05、16問から20問で0.04というように誤差が小さくなっており、出題数が少ないときほど問題が増えることによる誤差への影響が強いことがわかる。受検者の平均得点である0を含む-0.5以上0.5未満の区分に注目すると、4問増えるごとに0.03から0.04と一定の大きさを誤差が小さくなっており、40問から60問でも問題が増えることによる誤差の減少が認められる。60問実施することは、学力調査として学校や受検者にかなりの負担を与えるが、項目反応理論を用いた測定としては、60問以下の出題数であれば問題が増えることで測定精度が大きく変わるといえる。なお、誤差を1.96倍して得点の推定値から引いた値を下限とし、足した値を上限とする範囲は「95%信頼区間」と呼ばれており、この範囲に95%の確率で真の値が含まれるとされている。例えば、ここでのシミュレーション研究は100万人の受検者それぞれについて100回のシミュレーションを行っており、「100回それぞれの推定値±誤差×1.96」で

求められる 95%信頼区間に真の値が 100 回中何回含まれるか調べたところ、出題数が 4 問の場合、100 万人の平均は約 95.7 回であり、ほぼ 95%と一致する。若干多くなっているのは、真の値が-2.5 未満や 2.5 以上といった極端な場合に誤差が大きくなり、95%信頼区間に真の値が 100 回すべて含まれることなどが影響している。この 100 万人の平均は、出題数が 16 問のときは約 95.4 回、60 問のときは約 95.2 回と出題数が増えるほど 95 回に近づいていく。本稿のシミュレーション研究では受検者の得点分布を標準正規分布（平均 0、標準偏差 1）としており、その平均 0 は PISA 換算で 500 点、偏差値換算で 50 点であり、出題数 4 問で平均得点の誤差を 0.54（これは表 2 の-0.5 以上 0.5 未満の区分に対応する）とすると、その 95%信頼区間は-1.07 から 1.07 の幅 2.14 の範囲、PISA 換算で 393 点から 607 点の幅 214 の範囲、偏差値換算で 39 点から 61 点の幅 21 の範囲となる。4 問では最も測定精度が高い平均付近であっても、個人の測定値として役に立たないかもしれない。また 60 問であっても、95%信頼区間は-0.53 から 0.53 の幅 1.05 の範囲、PISA 換算で 447 点から 553 点の幅 105 の範囲、偏差値換算で 45 点から 55 点の幅 11 の範囲となり、学力調査としては問題ないかもしれないが、入学試験や資格試験のような結果が受検者の利害に大きく関わる試験としては、誤差が大きいかもしい。いずれにしろ、受検者の測定精度という面だけで言えば、調査時間の範囲内で、できるだけ多くの問題を出題することが重要であるといえる。EAP は事後分布の期待値、つまり平均値になっており、同じ事後分布の標準偏差は表 2 で求めた誤差の推定値（標準誤差）になる（Bock and Aitken 1981:453）。そこで、表 3 では、4 問、8 問、12 問、16 問、20 問、40 問、60 問の各条件について、全体の標準誤差の平均値と、7 区分の標準誤差の平均値を示す。

表 3 各出題数での標準誤差の平均値（全体、7 区分、小数点以下 3 桁で四捨五入）

	4 問	8 問	12 問	16 問	20 問	40 問	60 問
全体	0.75	0.63	0.56	0.50	0.46	0.35	0.29
-2.5 未満	0.77	0.67	0.62	0.58	0.55	0.45	0.40
-2.5 以上-1.5 未満	0.76	0.65	0.59	0.54	0.50	0.39	0.33
-1.5 以上-0.5 未満	0.75	0.63	0.56	0.50	0.46	0.35	0.29
-0.5 以上 0.5 未満	0.75	0.62	0.54	0.49	0.45	0.33	0.28
0.5 以上 1.5 未満	0.75	0.63	0.56	0.50	0.46	0.35	0.29
1.5 以上 2.5 未満	0.76	0.65	0.59	0.54	0.50	0.39	0.33
2.5 以上	0.77	0.67	0.62	0.58	0.55	0.45	0.40

誤差の推定値が正しければ、表 3 は表 2 と一致するはずである。しかし、表 2 と表 3 を比較すると、全体の平均値については出題数が 20 問を超えるとほぼ一致しているが、得点が-2.5 未満や 2.5 以上の区分を見ると表 3 では誤差がかなり過小推定されていることがわかる。逆に平均を含む-0.5 以上 0.5 未満の区分やその前後では、値は小さいが過大推定になっている。IRT 用のプログラムでは、得点の標準誤差を計算してくれるが、出題数が少ないときは、上記のように標準誤差にもバイアスが生じている。

## (2) 総問題数を拡張した場合の測定精度

PISA、TIMSS のような国際的な学力調査や、全国学力・学習状況調査の経年変化分析調査では、全受検者が同じ問題を解くのではなく、調査に使われる全問題を複数の冊子に分割し、受検者個人

はその冊子の中から1つを割り振られ、それに解答するという出題方法がとられている。こうすることで、受検者個人への出題数を変えずに調査全体としての総問題数を拡張できる。プロジェクト研究「学力アセスメント」では、「付録6 複数フォームの影響に関するシミュレーション」で、複数の問題冊子を使った場合に単独の問題冊子と結果が異なるかを検証している（国立教育政策研究所 2024b: 144）。この検証自体は他の先行研究にはなく、本稿ではさらにそれを問題数との関係から捉え直す。

「付録6 複数フォームの影響に関するシミュレーション」と同様、問題冊子の組み方として次の4つの場合（表4から7）を取り上げる。問題群に濃い網掛けを付したところは、総問題数が拡張されている部分と捉えられる。いずれの場合も各冊子は4問題群（1問以上の問題からなる問題の集合）の組合せとして作られており、それぞれの問題群は必ずすべての順番に同じ回数配置され（今回の場合は1番から4番に1回ずつ）に配置され、かつ問題群はいずれかの冊子で必ず他のすべての問題群と同時に出题される（つまり冊子間には必ず共通の問題群が存在する）という条件を満たしている。そして、これらの条件を満たし、総問題数を拡張した場合（表5、6、7）の冊子の組み方は「釣合い型不完備ブロック計画」（balanced incomplete block design）と呼ばれている（OECD 2014: 30）。

表4 4問題群による冊子組み（総問題数の拡張なし）

冊子番号	1番目	2番目	3番目	4番目
冊子1	問題群1	問題群2	問題群3	問題群4
冊子2	問題群2	問題群3	問題群4	問題群1
冊子3	問題群3	問題群4	問題群1	問題群2
冊子4	問題群4	問題群1	問題群2	問題群3

表5 5問題群による冊子組み（総問題数1.25倍）

冊子番号	1番目	2番目	3番目	4番目
冊子1	問題群1	問題群2	問題群3	問題群4
冊子2	問題群2	問題群3	問題群4	問題群5
冊子3	問題群3	問題群4	問題群5	問題群1
冊子4	問題群4	問題群5	問題群1	問題群2
冊子5	問題群5	問題群1	問題群2	問題群3

※濃い網掛けは4問題群から拡張された部分。

表6 7問題群による冊子組み（総問題数1.75倍）

冊子番号	1番目	2番目	3番目	4番目
冊子1	問題群1	問題群7	問題群3	問題群6
冊子2	問題群2	問題群1	問題群4	問題群7
冊子3	問題群3	問題群2	問題群5	問題群1
冊子4	問題群4	問題群3	問題群6	問題群2
冊子5	問題群5	問題群4	問題群7	問題群3
冊子6	問題群6	問題群5	問題群1	問題群4
冊子7	問題群7	問題群6	問題群2	問題群5

※濃い網掛けは4問題群から拡張された部分。

表 7 13 問題群による冊子組み（総問題数 3.25 倍）

冊子番号	1 番目	2 番目	3 番目	4 番目
冊子 1	問題群 1	問題群 13	問題群 3	問題群 9
冊子 2	問題群 2	問題群 1	問題群 4	問題群 10
冊子 3	問題群 3	問題群 2	問題群 5	問題群 11
冊子 4	問題群 4	問題群 3	問題群 6	問題群 12
冊子 5	問題群 5	問題群 4	問題群 7	問題群 13
冊子 6	問題群 6	問題群 5	問題群 8	問題群 1
冊子 7	問題群 7	問題群 6	問題群 9	問題群 2
冊子 8	問題群 8	問題群 7	問題群 10	問題群 3
冊子 9	問題群 9	問題群 8	問題群 11	問題群 4
冊子 10	問題群 10	問題群 9	問題群 12	問題群 5
冊子 11	問題群 11	問題群 10	問題群 13	問題群 6
冊子 12	問題群 12	問題群 11	問題群 1	問題群 7
冊子 13	問題群 13	問題群 12	問題群 2	問題群 8

※濃い網掛けは 4 問題群から拡張された部分。

ここでのシミュレーション研究では、個人への出題数が 4 問と 16 問の場合について、総問題数の拡張なし（つまり総問題数が 4 問と 16 問、表 4 がこれに該当）、総問題数 1.25 倍（5 問と 20 問、表 5 がこれに該当）、総問題数 1.75 倍（7 問と 28 問、表 6 がこれに該当）、総問題数 3.25 倍（13 問と 52 問、表 8 がこれに該当）で、測定精度が変わるかどうか検証する。測定精度には先のシミュレーション研究と同様、バイアスと誤差を用い、全体と得点の 7 区分ごとにそれらを求める。プロジェクト研究「学力アセスメント」で使われた「付録 6 複数フォームの影響に関するシミュレーション」（国立教育政策研究所 2024c: 178-182）のスク립トを一部変更して使用するため、架空の解答データの作成には先のシミュレーション研究と同様、R のパッケージ「mirt」（Chalmers 2012）の関数「simdata()」を使うが、2 パラメータ・ロジスティックモデルによる推定にはパッケージ「TAM」（Robitzsch et al. 2017）の関数「tam.mml.2pl()」を使用する。IRT での測定部分に異なる関数を用いているが、結果は先のシミュレーション研究と一致している（拡張なしの結果と表 1、表 2 を参照）。なお、調査問題は困難度パラメータの小さい方から問題群 1、2、3 と順番に配置しており、問題群によって難易度が異なっている。個人への出題数 4 問、総問題数 1.25 倍の場合、困難度パラメータは問題群 1 が -0.97...、問題群 2 が -0.43...、問題群 3 が 0、問題群 4 が 0.43...、問題群 5 が 0.97... となり、総問題数 1.75 倍の場合は -1.15...、-0.67...、-0.32...、0、0.32...、0.67...、1.15...、総問題数 3.25 倍の場合は -1.47...、-1.07...、-0.79...、-0.57...、-0.37...、-0.18...、0...、0.18...、0.37...、0.57...、0.79...、1.07...、1.47... となっている。個人への出題数 16 問、総問題数 3.25 倍で、問題群 1 の問題の困難度パラメータは -2.08...、-1.78...、-1.58...、-1.44...、問題群 13 は正負が逆の 1.44...、1.58...、1.78...、2.08... となり、このシミュレーション研究では割り当てられる問題冊子によって、問題の困難度パラメータの値に偏りがある。100 万人に 100 回のシミュレーションを行っている点は最初のシミュレーション研究と共通であるが、どの冊子がどの受検者に当たるかはシミュレーションごとにランダムに決めている。

表 8 は受検者個人への出題数が 4 問の場合、表 9 は 16 問の場合で、総問題数を拡張なし、1.25 倍、1.75 倍、3.25 倍にしたときのバイアスの平均値を示している。

表 8 出題数 4 問、総問題数を変えたときのバイアスの平均値  
(全体、7 区分、小数点以下 3 桁で四捨五入)

	拡張なし	1.25 倍	1.75 倍	3.25 倍
全体	0.00	0.00	0.00	0.00
-2.5 未満	1.85	1.85	1.85	1.86
-2.5 以上-1.5 未満	1.08	1.08	1.08	1.09
-1.5 以上-0.5 未満	0.48	0.48	0.49	0.50
-0.5 以上 0.5 未満	0.00	0.00	0.00	0.00
0.5 以上 1.5 未満	-0.48	-0.48	-0.49	-0.50
1.5 以上 2.5 未満	-1.08	-1.08	-1.08	-1.09
2.5 以上	-1.85	-1.85	-1.85	-1.86

表 9 出題数 16 問、総問題数を変えたときのバイアスの平均値  
(全体、7 区分、小数点以下 3 桁で四捨五入)

	拡張なし	1.25 倍	1.75 倍	3.25 倍
全体	0.00	0.00	0.00	0.00
-2.5 未満	1.03	1.04	1.04	1.05
-2.5 以上-1.5 未満	0.50	0.50	0.50	0.52
-1.5 以上-0.5 未満	0.19	0.20	0.20	0.20
-0.5 以上 0.5 未満	0.00	0.00	0.00	0.00
0.5 以上 1.5 未満	-0.19	-0.20	-0.20	-0.20
1.5 以上 2.5 未満	-0.50	-0.50	-0.50	-0.52
2.5 以上	-1.03	-1.04	-1.03	-1.05

表 8 の出題数 4 問の場合は、全体では拡張なし（総問題数 4 問）から 3.25 倍（総問題数 13 問）まで違いが見られず、得点の 7 区分では-0.5 以上 0.5 未満を除く区分で微妙な違いが存在する。ただし、拡張なしと 3.25 倍の差の絶対値が最大である 1.5 以上 2.5 未満の区分であっても、その差は-0.015...であり、バイアスそのもの（-1.08 から-1.09）の 1%程度の大きさでしかない。表 9 の出題数 16 問の場合も、全体では違いが見られず、得点の 7 区分では微妙な違いが見られるが、差の絶対値が最大である 2.5 以上の区分であっても-0.021...で、バイアスそのもの（-1.03 から-1.05）の 2%程度の大きさにすぎない。よって、総問題数の拡張は、得点のバイアスにほとんど影響していないといえる。

表 10 と 11 は、それぞれ受検者個人への出題数が 4 問と 16 問の場合に、総問題数を拡張なし、1.25 倍、1.75 倍、3.25 倍にしたときの誤差の平均値を表している。

表 10 の出題数 4 問の場合、全体では拡張なし（総問題数 4 問）から 3.25 倍（総問題数 13 問）まで違いが見られず、得点の 7 区分では-0.5 以上 0.5 未満を除く区分で違いが存在する。ただし、拡張なしと 3.25 倍の差が最大である 1.5 以上 2.5 未満の区分であっても違いは 0.018...であり、誤差そのもの（1.15 から 1.17）の 2%程度の大きさでしかない。表 11 の出題数 16 問の場合、全体でも得点の 7 区分でも違いが見られる。差が最大である 2.5 以上の区分では、拡張なしと 3.25 倍の差が 0.033...であり、誤差そのもの（1.08 から 1.12）の 3%程度の大きさである。表 8、表 9 のバイアスに比べれば大きいとはいえ、総問題数の拡張は、得点の誤差にほとんど影響していないことがわ

かる。

表 10 出題数 4 問、総問題数を変えたときの誤差の平均値  
(全体、7 区分、小数点以下 3 桁で四捨五入)

	拡張なし	1.25 倍	1.75 倍	3.25 倍
全体	0.71	0.71	0.71	0.71
-2.5 未満	1.87	1.87	1.88	1.88
-2.5 以上-1.5 未満	1.15	1.15	1.16	1.17
-1.5 以上-0.5 未満	0.70	0.70	0.70	0.71
-0.5 以上 0.5 未満	0.54	0.54	0.54	0.54
0.5 以上 1.5 未満	0.70	0.70	0.70	0.70
1.5 以上 2.5 未満	1.15	1.15	1.16	1.17
2.5 以上	1.87	1.87	1.88	1.88

表 11 出題数 16 問、総問題数を変えたときの誤差の平均値  
(全体、7 区分、小数点以下 3 桁で四捨五入)

	拡張なし	1.25 倍	1.75 倍	3.25 倍
全体	0.49	0.49	0.49	0.50
-2.5 未満	1.08	1.09	1.09	1.12
-2.5 以上-1.5 未満	0.64	0.64	0.65	0.66
-1.5 以上-0.5 未満	0.47	0.48	0.48	0.48
-0.5 以上 0.5 未満	0.44	0.44	0.44	0.45
0.5 以上 1.5 未満	0.47	0.48	0.48	0.48
1.5 以上 2.5 未満	0.64	0.64	0.65	0.66
2.5 以上	1.08	1.09	1.09	1.12

なお、今回のシミュレーション研究では、先述したように問題群による難易度の違いが大きくなるように問題を配置している。釣合い型不完備ブロック計画を採用しているため、困難度パラメータが高い（低い）問題のみの冊子ができないようになっているが、実際の調査でもここまで難易度が偏ることはないと予想される。よって、総問題数の拡張による影響は現実にはより少ないと考えられる。

本節の 2 つのシミュレーション研究を比較すると、受検者個人への出題数を増やすことは、個人の得点を測定する際のバイアスと誤差を小さくし、測定精度の向上につながるのに対して、調査全体の総問題数を拡張することは、受検者個人の得点の測定精度とほとんど関係していないといえる。本稿では個人レベルの測定精度のみを扱ってきたが、プロジェクト研究「学力アセスメント」の「付録 6 複数フォームの影響に関するシミュレーション」では、母集団の平均得点や標準偏差といった母集団特性の推定においても、受検者個人への出題数を増やすことが測定精度の向上につながっているのに対して、調査全体の総問題数の拡張が測定精度と関係していないことが示されている(国立教育政策研究所 2024c: 180)。つまり個人であれ、集団であれ、測定精度に影響するのは受検者個人への出題数であり、調査全体の総問題数ではないといえる。

## 5 まとめと今後に向けて

本稿では、プロジェクト研究「学力アセスメント」で使用されたシミュレーション研究を利用して、受検者個人への出題数が少ないときの測定精度に関するシミュレーション研究と、調査全体の総問題数を拡張したときの測定精度に関するシミュレーション研究を行った。IRTにおける問題数の検証を行った研究の中でも、本稿は特に受検者の測定精度を中心に議論しており、得点の区分ごとにバイアスと誤差を見たり、出題数と総問題数を分けて捉えたりしたところに独自性がある。

受検者個人への出題数が少ないときの結果を見て、IRTの使用を躊躇<sup>ちゅうちゆう</sup>するかもしれないが、正答数や正答率を用いたとしても、調査を行うたびに違い＝誤差が生じるし、全問正答や全問誤答、それに近いときは正確な測定ができない＝バイアスが生じている可能性がある。本稿のシミュレーション研究の結果は、IRTを使わない場合でも見過ごせないものである。ただし、本稿では取り上げなかったが、受検者全体の特性、例えば平均得点についてはバイアスが生じないし、他の統計量についても様々な対応策があるため、学力調査という目的であれば少ない出題数でも有益な結果が得られる。また、多次元項目反応モデルを使用したり、質問項目や受検者の属性情報をIRTの推定に使ったりすることで、測定精度を上げることも可能である。さらには、コンピュータ使用型の適応型テスト（computerized adaptive testing）を用いることで、受検者の能力に適した問題を集中的に出題することができ、より少ない出題数で測定精度を高めることも可能である。

調査全体の総問題数を拡張することが測定精度に影響しないということは、受検者個人への出題数を減らさなくても、解答に時間がかかる問題を出題できる可能性を広げてくれる。1人の受検者にそのような問題を多数試さなくても、幾つかの問題群の中にそのような問題を配置することで、受検者全体では多くの問題を試すことができるからだ。

プロジェクト研究「学力アセスメント」の報告書第4部では、多次元項目反応モデルや適応型テスト、問題冊子の組み方やその配布方法といった学力調査を行う上で検討すべき様々な内容を議論している。学力調査に関わっている、若しくはこれから関わっていく方々に本稿とともに参考にしていただければ幸いである。

### 謝辞

本稿の議論は、プロジェクト研究「学力アセスメント」の測定技術班会合で話し合われた内容に着想を得た。同会合に参加された先生方、事務局の皆様方にお力添えをいただいているが、本稿の内容はすべて筆者の責任の下に書かれている。また、先行研究に関する文献研究、シミュレーション研究の一部は、JSPS 科研費 23K2753 の助成を受けている。

### 文献

- Birnbaum, A. (1968). Some latent trait models. In Load, F.M. and Novick, M.R. (eds.) *Statistical Theories of Mental Test Scores*. Addison-Wesley: Massachusetts, pp. 397-424.
- Bock, R. D. and Aitken, M. (1981). Marginal Maximum Likelihood Estimation of Item Parameters Application of an EM Algorithm. *Psychometrika*, 46(4), 443-59.
- Bock, R. D. and Lieberman, M. (1970). Fitting a Response Model for n Dichotomously Scored Items. *Psychometrika*, 35(2), 179-97.
- Bock, R. D. and Mislevy, R. J. (1982). Adaptive EAP Estimation of Ability in a Microcomputer Environment. *Applied*

- Psychological Measurement*, 6(4), 431-444.
- Chalmers, P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, 48(6), 1-29.
- de Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. Guilford Press: New York.
- Dragow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement*, 13(1), 77-90.
- 袈岩晶, 篠原真子, 篠原康正. (2019). PISA 調査の解剖. 東信堂.
- Hulin, C. L., Lissak, R. I., and Dragow, F. (1982). Recovery of Two- and Three-Parameter Logistic Item Characteristic Curves: A Monte Carlo Study. *Applied Psychological Measurement*, 6(3), 249-260.
- 国立教育政策研究所 (編). (2021). TIMSS2019 算数・数学教育/理科教育の国際比較 ——国際数学・理科教育動向調査の2019年調査報告書. 明石書店.
- 国立教育政策研究所 (編). (2024a). 生きるための知識と技能8 OECD生徒の学習到達度調査(PISA) ——2022年調査国際結果報告書. 明石書店.
- 国立教育政策研究所 (編). (2024b). 学力アセスメントの在り方に関する調査研究報告書. 国立教育政策研究所. ([https://www.nier.go.jp/05\\_kenkyu\\_seika/pdf\\_seika/r05/assessment\\_zentai.pdf](https://www.nier.go.jp/05_kenkyu_seika/pdf_seika/r05/assessment_zentai.pdf)).
- 国立教育政策研究所 (編). (2024c). 学力アセスメントの在り方に関する調査研究報告書第4部付録. 国立教育政策研究所. ([https://www.nier.go.jp/05\\_kenkyu\\_seika/pdf\\_seika/r05/assessment\\_huroku.pdf](https://www.nier.go.jp/05_kenkyu_seika/pdf_seika/r05/assessment_huroku.pdf)).
- Lord, F. M. (1962). Estimating Norms by Item-sampling. *Educational and Psychological Measurement*, 22(2), 259-267.
- Lord, F. M. (1968). An Analysis of the Verbal Scholastic Aptitude Test Using Birnbaum's Three Parameter Logistic Model. *Educational and Psychological Measurement*, 28(4), 989-1020.
- 文部科学省・国立教育政策研究所. (2022). 令和3年度全国学力・学習状況調査 経年変化分析調査実施結果報告書. 文部科学省・国立教育政策研究所. ([https://www.nier.go.jp/21chousakekkahoukoku/kannren\\_chousa/pdf/21keinen\\_report.pdf](https://www.nier.go.jp/21chousakekkahoukoku/kannren_chousa/pdf/21keinen_report.pdf)).
- 文部科学省・国立教育政策研究所. (2024). 令和6年度全国学力・学習状況調査の結果(概要). 文部科学省・国立教育政策研究所. (<https://www.nier.go.jp/24chousakekkahoukoku/report/data/24summary.pdf>).
- 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.
- OECD (2014b). *PISA 2012 Technical Report*. OECD. (<https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf>).
- 大友賢二. (1996). 項目応答理論入門: 言語テスト・データの新しい分析法. 大修館書店.
- R Core Team. (2024). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. (<https://www.R-project.org/>).
- Robitzsch, A., Kiefer, T., and Wu, M. (2017). TAM: Test analysis modules. R package version 2.8-21. (<https://CRAN.R-project.org/package=TAM>).
- Şahin, A., and Anil, D. (2017). The effects of test length and sample size on item parameters in item response theory. *Kuram ve Uygulamada Eğitim Bilimleri/Educational Sciences: Theory & Practice*, 17(1), 321-335.
- Seong, T. j. (1990). Sensitivity of Marginal Maximum Likelihood Estimation of Item and Ability Parameters to the Characteristics of the Prior Ability Distributions. *Applied Psychological Measurement*, 14(3), 299-311.
- Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement*, 16(1), 1-16.
- 豊田秀樹. (2012). 項目反応理論 [入門編] 第2版. 朝倉書店.

全国的な学力調査の CBT 化検討ワーキンググループ. (2021). 全国的な学力調査の CBT 化検討ワーキンググループ  
最終まとめ. 文部科学省. ([https://www.mext.go.jp/content/20210719-mxt\\_chousa02-000016768-2.pdf](https://www.mext.go.jp/content/20210719-mxt_chousa02-000016768-2.pdf)).