研究論文

全国学力・学習状況調査「教科に関する調査」の品質検証

―平成 28、30 年度、令和 3、4 年度の比較―

*田　端　健　人

要　旨

本稿の目的は、全国学力・学習状況調査「教科に関する調査」（以下「教科調査」）の品質を検証することにある。第1章では、文部科学省による全国学力・学習状況調査の目的に照らし、調査問題の精度検証が不可欠であることを述べた。第2章では、テスト理論を参照し、テストの品質はその妥当性と信頼性にあることを解説した。第3章では、教科調査の信頼性の検証として、クロンバックのα係数を算出した。同調査の品質検証に関する先行研究を批判的に補完するため、本稿での検証対象年度を、問題数が多かった平成28、30年度と問題数が減少した令和3、4年度に定めた。利用データは、文部科学省から貸与を受けた匿名データ、各年度各学年約10万件である。基準値は先行研究に倣い、国語α=0.75、算数・数学α=0.80とした。計算の結果、平成の教科調査では小6中3の国語と算数・数学すべてで基準値をクリアしたが、令和の教科調査ではすべてで基準値を下回った。第4章では、テスト情報量と測定誤差につき4か年度を比較した。分析にはIRT分析ソフトEasyEstimationを利用し、出力された数値を統計ソフトRにて独自に可視化した。その結果、問題数減少の令和調査では、テスト情報量が低下し、測定誤差も大きくなっていることが分かった。第5章では、EasyEstimationが出力する「項目番号」「難易度」「無回答」の相関係数を独自に計算・可視化し、良いテストの基本原則が守られているかを検証した。その結果、令和3年度小6国語、令和4年度小6算数、令和3、4年度中3国語で、「易しい問題から難しい問題へと配列する」という基本原則が守られていないことが判明した。対して中3数学では4か年度すべてでこの原則が守られていた。また項目難易度と無回答の相関を分析することで、令和3年度小6国語、令和4年度小6算数、平成28、30、令和4年度中3国語で、「難しい問題ほど無回答が多くなる」という自然な現象に反する不自然な結果となった。第6章では、教科調査の妥当性の検証として、本体調査と経年変化分析調査との同一教科間個人スコアの相関の強さを計算した。経年変化分析調査が実施された平成25、28、令和3年度を検証対象とした。相関係数は最大0.86、最小0.70であった。令和3年度で、小6中3の国語、算数・数学のすべてで相関係数が低下している結果となった。ただこの相関係数の評価については、その「基準値」の議論の不足から慎重を要する。そこで本稿では、異教科間の相関係数を算出し、同一教科の異調査間の相関係数を見積もってみた。その結果、同一教科の異調査間の相関係数0.70はどちらかと言えば改善が望ましい値であるとの暫定的評価にいたった。第7章で以上の結論をまとめ、改善を提言した。

Key words：テストの妥当性と信頼性，テスト情報量，項目誤差，経年変化分析調査

1．問題背景

文部科学省が毎年4月、小学6年生（以下「小6」）と中学3年生（以下「中3」）を対象に悉皆で実施している全国学力･学習状況調査の「教科に関する調査（以下「教科調査」）は、A問題とB問題が統合された平成31年度以降、問題数が大幅に減少した。令和4年度は、小6国語14問、小6算数16問、中3国語14問、中3数学14問であった。全14問程度の問題数で、児童生徒の学力を本当に測定できるのだろうか。測定の精度はどれほど信頼できるだろうか。これが本稿の問題意識である。

＊　宮城教育大学教職大学院・教授

文部科学省は、本調査の目的を次のように定めている。

> 義務教育の機会均等とその水準の維持向上の観点から、全国的な児童生徒の学力や学習状況を把握・分析し、教育施策の成果と課題を検証し、その改善を図るとともに、学校における児童生徒への教育指導の充実や学習状況の改善に役立てる。[1]

「教育施策の成果と課題」の検証も、「児童生徒への教育指導(学習指導)の充実」への役立ても、広義の「テスト」である調査問題が、測定したい学力を確かな精度で測定していることを前提としている。だが、教科調査の精度を数量的に検証した先行研究は、問題数が減少した平成31年度以降、テストの品質が低下したと指摘している（cf.,田端，2022，第5,6,7章)。本稿はこの先行研究を批判的に吟味しつつ、新たに令和4年度調査の分析を加え、全国学力・学習状況調査の品質を検証する。

2．テスト品質の検証方法―妥当性と信頼性―

知能や学力であれ、心理状態やパーソナリティであれ、測定のための調査問題は広く「テスト」と呼ばれる。「そのテストで測定したい心理特性を（妥当性の問題)、本当に精度よく（信頼性の問題)、測定しているかどうかを統計的に判断」するための理論が「テスト理論」である（柴山，2020，p.1，括弧内原文)。テスト理論によれば、各テストの品質は、そのテストの「妥当性（validity)」と「信頼性（reliability)」にある。そこで本稿でも、全国学力･学習状況調査「教科調査」の妥当性と信頼性を検証する。

「妥当性」とは、あるテストが測定したい対象を一定の妥当性をもって測定しているか、測定したい対象にそのテストがどれほど妥当しているかの問題である。それゆえ、妥当性の問題には、そのテストが測定したい対象は何か、という哲学的問題も入る。例えば、数学の力を測りたいのに、問題文が長く読解力が試される場合などは、その問題項目が数学力の測定にふさわしいか、妥当性に疑義が生じる。数学力には読解力も入ると主張するとき、私たちは数学力の定義をめぐる哲学的問題圏に入っているわけである。

テスト理論ではこれまで、妥当性は「基準連関妥当性」「構成概念妥当性」「内容的妥当性」の3つに分類されてきた(cf.,柴山，2020, p.4; 村山，2012, p.118)。ところが1980年代以降になると、これら3タイプは構成概念的妥当性に収斂するとの見方が主流になったという（村山，2012，p.118)。構成概念的妥当性とは、「当該のテストで測定しようとしている構成概念（construct）をどの程度うまく測定しているかに関わる妥当性」（柴山,2020,p.7）である。学力であれ非認知能力であれ、測定の対象になるとそれらはすべて、テスト理論のいう「構成概念」、現象学的には「理念的構成物」になる。テストは、この理念的構成物の考案（現象学的には「企投（project)」）からはじまる。特定の理念的構成物、例えば「読解力」とか「やり抜く力」とかを考え出し、それが何かを概念的に明確化し、それを測定するにはどのような問いかけ（「問題」「質問」「項目」等ともいう）がふさわしいか吟味するのが「作題」の作業である。作題した問題の妥当性を検証できるのは、問題への回答が返ってきた後、つまりテスト結果によってである。

妥当性が3タイプに分かれるように、妥当性の検証には複数の方法がある。テスト理論の専門家によれば、「構成概念妥当性を支える証拠を探っていくことは、それ自体クリエイティブなプロセスであり、永続的な作業である」（村山，2012，p.122)。それゆえ、全国学力・学習状況調査の妥当性にしても、創造的で永続的な複数の証拠探しが必要であろう。本稿はその一つとなることをめざしている。

本稿では第6章で、構成概念妥当性を検証するために、教科調査と経年変化分析調査との整合性を検討する。

次に「信頼性」である。これは「そのテストが測りたいものを本当に精度よく測定しているかどうかの問題」（柴山，2020，p.8）である。妥当性の問題にも似ているが、「信頼性は妥当性の必要条件であるが十分

1　文部科学省，2022「令和5年度全国学力・学習状況調査に関する実施要領」.
https://www.mext.go.jp/content/20221207-mxt_chousa02-000026336-1.pdf[2023.09.14最終閲覧]
なお、「教育指導」の文言は、令和6年度には「学習指導」と改められている（文部科学省，2023「令和6年度全国学力・学習状況調査に関する実施要領」参照。https://www.mext.go.jp/content/20231221-mxt_chousa02-000033188-1.pdf [2024.01.30最終閲覧])。

条件ではない」(柴山, 2020, p.9) とされる。

信頼性を検証する方法も複数あるが、なかでもテストの「一貫性 (内的整合性)」の検証は、広く行われる重要な検証の一つである (cf.,宮本ほか編, 2015, pp.66-67)。教科学力とかパーソナリティとかに関する問題項目群が、測定したい理念的構成物をどの程度一貫して測定しているか、逆に一貫性のない項目がどの程度混在しているかが、内的整合性の問題である。

信頼性を検証する方法としてよく用いられるのが「クロンバックのα係数」であり、「信頼性係数」の代表である。本稿もまずはクロンバックのα係数により、教科調査の信頼性を検証する。

3. 信頼性係数 (クロンバックのα) の比較

本稿では、検証対象の年度を、AB問題があった平成28、30年度、問題数が減少した直近 (令和5年9月現在) の令和3、4年度に絞る。先行研究 (田端,2022) では、平成24年度から令和3年度まで追跡されている。先行研究と重複する平成28、30、令和3年度の再検証は、先行研究の批判的吟味になり、新たに加える令和4年度は、先行研究にない直近の動向を経年で検証することになる。

利用データは、文部科学省から貸与を受けた匿名データである[2]。これは母集団 (各学年約100万人) の1割程度の無作為抽出である。

事前のデータ処理を、以下の手順で行った。

1. 貸与匿名データから、国語と算数・数学の正誤情報を取り出す。匿名データでは、欠測値はブランク、無回答は「0」、正答「1」、誤答「2」になっている。
2. α係数は、無回答を誤答に含めると高くなり、含めないと低くなる。今回は無回答を含めない。国が実施する調査であるため、若干厳しい条件でもクリアしてほしいという願いからである。そこで、文部科学省の匿名データの、欠測値 (ブランク) と無回答 (「0」) を含む行 (ID) を削除する。エクセルで、「0」をブランクに置換し、IBM社の統計ソフト SPSS でブランクを「ケースの選択」により全て削除する。
3. 欠測値と無回答を削除したデータセット (正答「1」と誤答「2」だけで構成されたデータセット) を、統計ソフト R[3] の psych パッケージで alpha関数により計算する。データセットの csv ファイルを暫定名「dat」に代入したあと、「alpha(dat[, -1])」のスクリプトで実行できる。[, -1] というスクリプトは、最初の1列目 (ID番号) を含めないという意味である。1列目からデータが始まる場合は、この部分は不要である。

評価のための基準値は、先行研究に倣い、国語は0.75、算数・数学は0.80とする (cf.,田端,2022,p.34, p.36)。先行研究にある通り、これは決して高いハードルではない。非認知能力を測定する質問紙調査でも、よく設計された項目群ならば、α係数は0.80を優に超える。国語とか算数・数学とかの学力の方が、非認知能力のような歴史の浅い理念的構成物よりも、構成概念としていっそう明確であり、測定のための問題づくりの蓄積も豊富であるため、非認知能力の質問紙調査よりいっそう高い一貫性 (内的整合性) があってしかるべきである。

4つの年度の問題数は、表1の通りである。略記号 EL は小学6年 (「小6」と略記) を、JH は中学3年 (「中3」と略記) を意味し、Jpn は国語、Math は算数・数学の略記号である。年度は元号の頭文字で表記した。

表1：問題数一覧

	H28	H30	R3	R4
EL_Jpn	25	20	14	14
EL_Math	29	24	16	16
JH_Jpn	42	41	14	14
JH_Math	51	50	16	14

(単位：問)

表1からわかるように、令和4年度も令和3年度と同様、問題数が少ない。中3数学は令和3年度より2問少なくなっている。

欠測値と無回答を削除した後のデータ人数 (N) は、表2の通りである。

---

2 個票データ等の貸与に関しては、以下のウェブページに記載されている。
https://www.mext.go.jp/a_menu/shotou/gakuryoku-chousa/sonota/1386492.htm [2023.09.14最終閲覧]

3 https://www.r-project.org/ [2023.09.14最終閲覧]

表２：処理したデータ人数（N）一覧

	H28	H30	R3	R4
EL_Jpn	65,530	80,157	75,474	68,438
EL_Math	64,012	64,773	83,033	80,127
JH_Jpn	66,182	56,651	64,055	66,337
JH_Math	35,844	43,472	46,388	44,615

（単位：人）

α係数の計算結果は、表3の通りである。

表３：信頼性係数（クロンバックのα）

	H28	H30	R3	R4
EL_Jpn	0.77	0.78	0.71	0.71
EL_Math	0.85	0.86	0.76	0.75
JH_Jpn	0.80	0.79	0.66	0.60
JH_Math	0.91	0.91	0.74	0.76

平成28、30年度と令和3年度の今回の結果を先行研究（cf.,田端，2022 ,p.36）と照らし合せたところ、α係数はほぼ一致した。

平成28、30年度は、小6中3、国語、算数・数学のすべてで基準値をクリアしている。ところが、令和3、4年度は、すべてで基準値に達していない。深刻なのは、令和3、4年度の中3国語で、α係数が0.70を下回っていることである。信頼性係数の基準値には諸説あるが、「一般的な心理テストの場合、村上（2006：40）は0.8以上という基準を示し」ており、「ホーガン（2010：113）」の見解は「試験問題冊子の信頼性としてはおおむね0.9から0.95が望ましい値であるものの、研究目的（集団の差に関心があるような場合）のためには0.7から0.8でもよいと要約」できるという（光永，2018，p.92,括弧内原文）。こうした見解からすれば、社会的な意義と影響が大きく、データとしても価値の高い全国規模の、かつ文部科学省が実施する全国悉皆調査の信頼性係数が0.70に届かないのは、改善が必要ではないだろうか。

先述のように、無回答を誤答に含めた場合、α係数は高くなる。先行研究によれば、無回答を含む場合、令和3年度小6国語は0.84、算数0.78、中3国語0.76、数学0.86になる（田端，2022，p.36）。この場合、国語は小中共に基準値を超えるが、小6算数は基準値に届かない。もし算数・数学の基準値を0.75に下げるなら、全てでクリアすることになる。

このように信頼性の問題は、計算の仕方や基準値により評価が分かれる。しかし、全国学力・学習状況調査は、国が実施する調査であり、社会的インパクトの大きな調査であることから、若干厳しい条件でもクリアすることが望ましい、と本稿では提案したい。

教科に関する調査は、平成28、30年度は、「知識」関する問題（A問題）に1単位時間、「活用」に関する問題（B問題）に1単位時間、計2単位時間が当てられていた[4]。対してAB問題が統合された平成3、4年度は、1単位時間に半減した[5]。それゆえ問題数も、かつてほど多くは出題できないことは理解できる。また時間数の半減は、調査が児童生徒に与える負担を軽減する点で評価できる。しかし、1単位時間のなかでも、問題数をもう少し増やすことはそれほど難しいことではない。問題を長文で出題することが、問題数を増やせない最大の要因と考えられる。国語であれ算数・数学であれ長文読解の力を問いたいというメッセージはわかるが、テストの信頼性をおろそかにするのは行き過ぎではないだろうか。

次に、テスト情報量や測定誤差の観点から、同じ4つの年度を検証してみる。

4　Cf.，文部科学省「平成28年度全国学力・学習状況調査に関する実施要領」および「平成30年度全国学力・学習状況調査に関する実施要領」.
https://www.mext.go.jp/a_menu/shotou/gakuryoku-chousa/zenkoku/1365022.htm　[2023.09.14最終閲覧]
https://www.mext.go.jp/a_menu/shotou/gakuryoku-chousa/zenkoku/1400166.htm　[2023.09.14最終閲覧]

5　Cf.，文部科学省「令和3年度全国学力・学習状況調査に関する実施要領」および「令和4年度全国学力・学習状況調査に関する実施要領」.以下のウェブページからPDFファイルにアクセスできる。
https://www.mext.go.jp/a_menu/shotou/gakuryoku-chousa/zenkoku/1411707_00004.htm　[2023.09.14最終閲覧]
https://www.mext.go.jp/a_menu/shotou/gakuryoku-chousa/zenkoku/1411707_00009.htm　[2023.09.14最終閲覧]

4．テスト情報量と測定誤差の比較

IRT分析ソフト EasyEstimation[6]を利用すれば、各問題項目の「識別力（slope）」と「難易度（location）」がわかる。これらの値からテスト情報量と測定誤差を可視化できる。

テスト情報量は、図1のような「テスト情報量曲線」で表される。この曲線は横軸に受検者の能力（学力値θ）、縦軸にその能力に関するテストの情報量、つまりその能力を測定する問題が問題冊子全体でどのくらい含まれているかの情報量をとったグラフである（cf., 光永，2018, pp.117-118）。この曲線をみれば、当該問題冊子がどの学力値に対して識別力を持っているかがわかる。

EasyEstimation にかけるためのデータセットの作成方法や、このソフトの扱い方については、先行研究（田端，2022）で詳しく解説されているので本稿では省略する。

作図結果は図1の通りである。EasyEstimation で出力される項目パラメタの slope と location の値から R で独自に、4つの年度を重ね書きした。平成28年度を青、平成30年度を緑、令和3年度を赤、令和4年度を黒にしている。

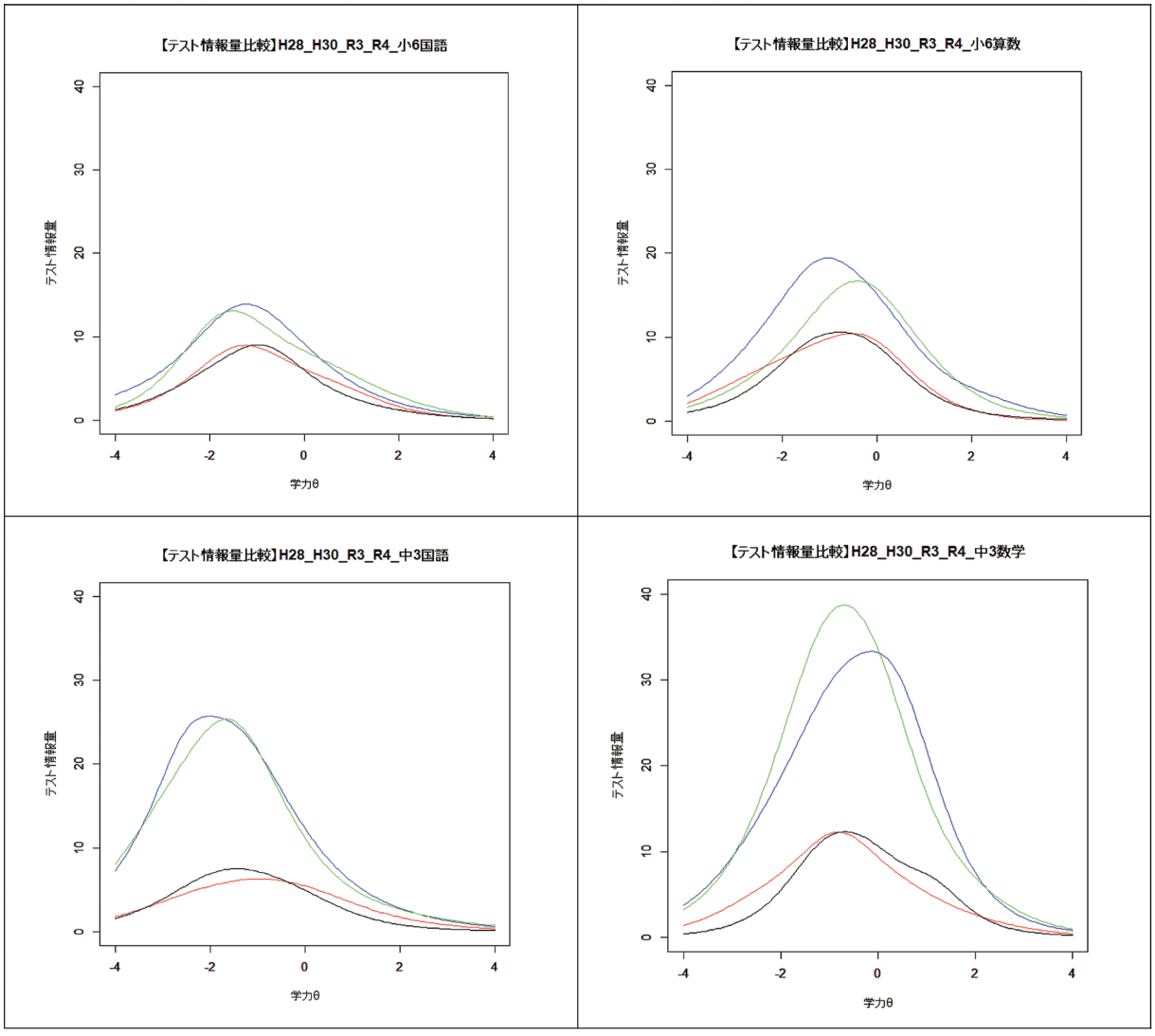

図1：テスト情報量の比較

6　開発者は東北大学の熊谷龍一氏である。以下の URL でサイトにアクセスし、ダウンロードできる。
http://irtanalysis.main.jp/　[2023.09.14最終閲覧]

横軸が学力θ（シータ）で、「0」が平均である。どの年度も曲線の頂点がゼロより小さい位置（マイナス）にあるのは、平均よりやや低い問題の情報量が多い、つまり平均よりやや低い学力を測定する感度が高いことを意味する。曲線の山の高さは情報量の多さを表わす。

平成28、30年度に比べ、令和3、4年度は、いずれの学年いずれの教科でも、テスト情報量が少ない。小6国語の差はさほど大きくないが、中3国語、中3数学の落差は大きい。この関係は、ほぼ問題数と一致する。

測定誤差のカーブを重ね書きすると、図2になる。

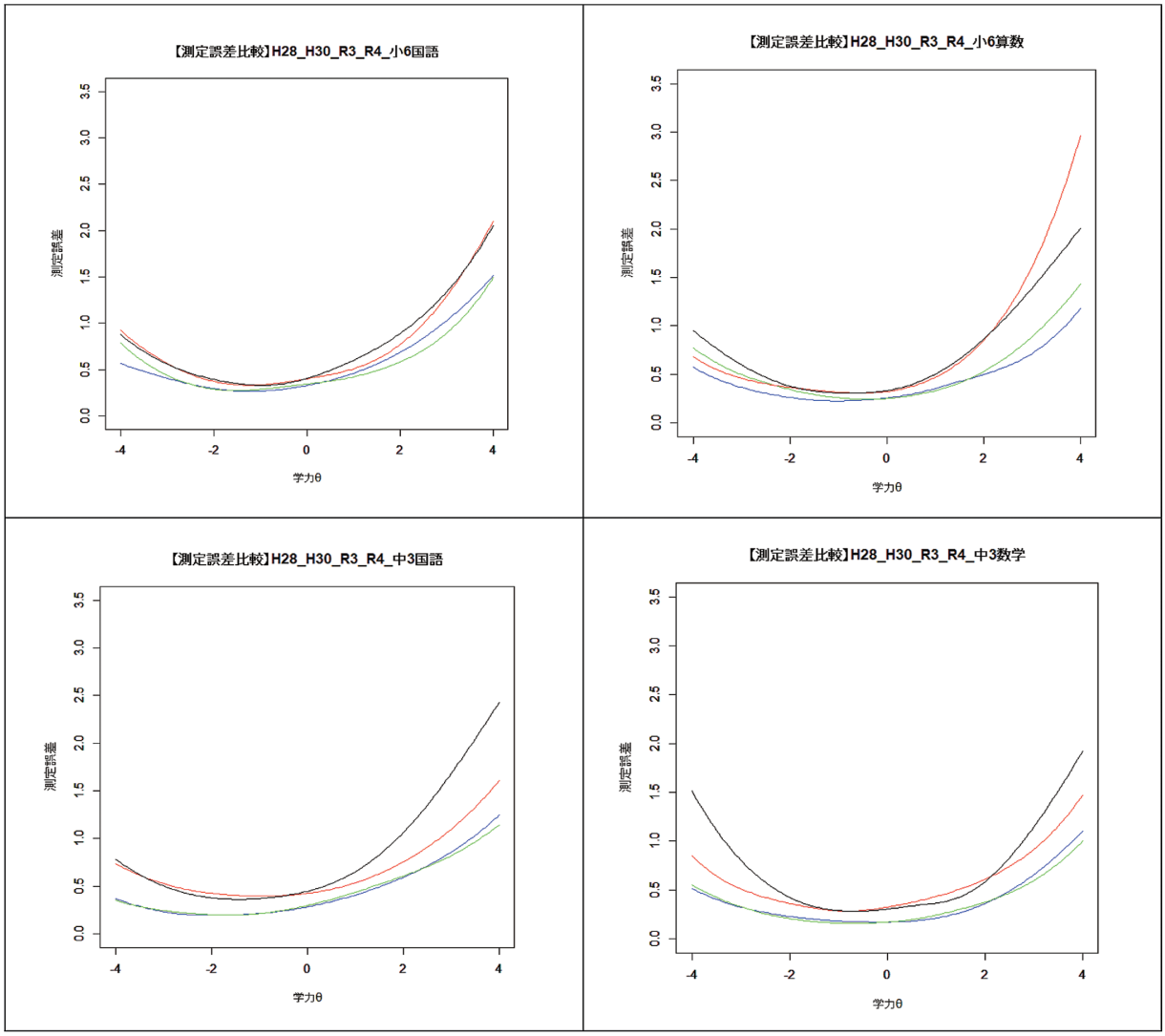

図2：測定誤差の比較

学力の平均（ゼロ）から離れるにつれ、誤差が大きくなっている。特に学力の高い児童生徒の測定で誤差が大きいことが分かる。ところが令和4年度の中3数学は、学力の低い児童生徒の測定誤差も跳ね上がっている。図1のテスト情報量で見ても、令和4年度中3数学は -2あたりで令和3年度と比較しても情報量がぐっと少なくなっている。学力の低い生徒を測定するテスト情報量が小さいため、学力の低い生徒の測定誤差も大きくなっている。令和4年度中3数学は全14問と、中3数学のこれまでの調査で最も問題数が少なかった。その影響と推測できる。

5．項目番号・難易度・無回答の相関関係

先行研究では、良いテストの設計の基本原則を、以下のように提案している7。

7　枠内は、田端, 2022, p.58より引用。

1）問題は易しいものから難しいものへの順に配列する。
2）大問中の小問も易しいものから難しいものへの順に配列する。
3）上記1)2) により無回答が出にくくなる。無回答はできるだけ少ない方がよい。
4）途中で回答行動がストップしても、上記のように並べておけば，学力をより正確に測定できる。
5）見た目の複雑さによる無回答行動を避けるため、問題はなるべくシンプルにし、その難易度に応じた本質的な聞き方にする。

この基本原則からすれば、問題の項目番号と難易度との間に、一定の相関があるのが良いテストの条件になる。また、問題の難易度と無回答数との間にも、一定の相関があるのが自然である。なぜなら、難しい問題の方が易しい問題より無回答数が多いはずだからである。

そこで先行研究では、EasyEstimation の出力から、項目番号（itemID）、難易度（location）、無回答数（omit）のデータを取り出し、データセットにして、それぞれの相関関係を一覧化している。本稿でも同じ要領で、令和４年度を加えた先の４つの年度の相関関係を一覧化してみる。

相関係数（r）の基準値は、r=0.20 を弱い相関の最低ラインとする。

まずは、小６国語の相関関係を図３に示す。各グラフ冒頭に年度を記載した。

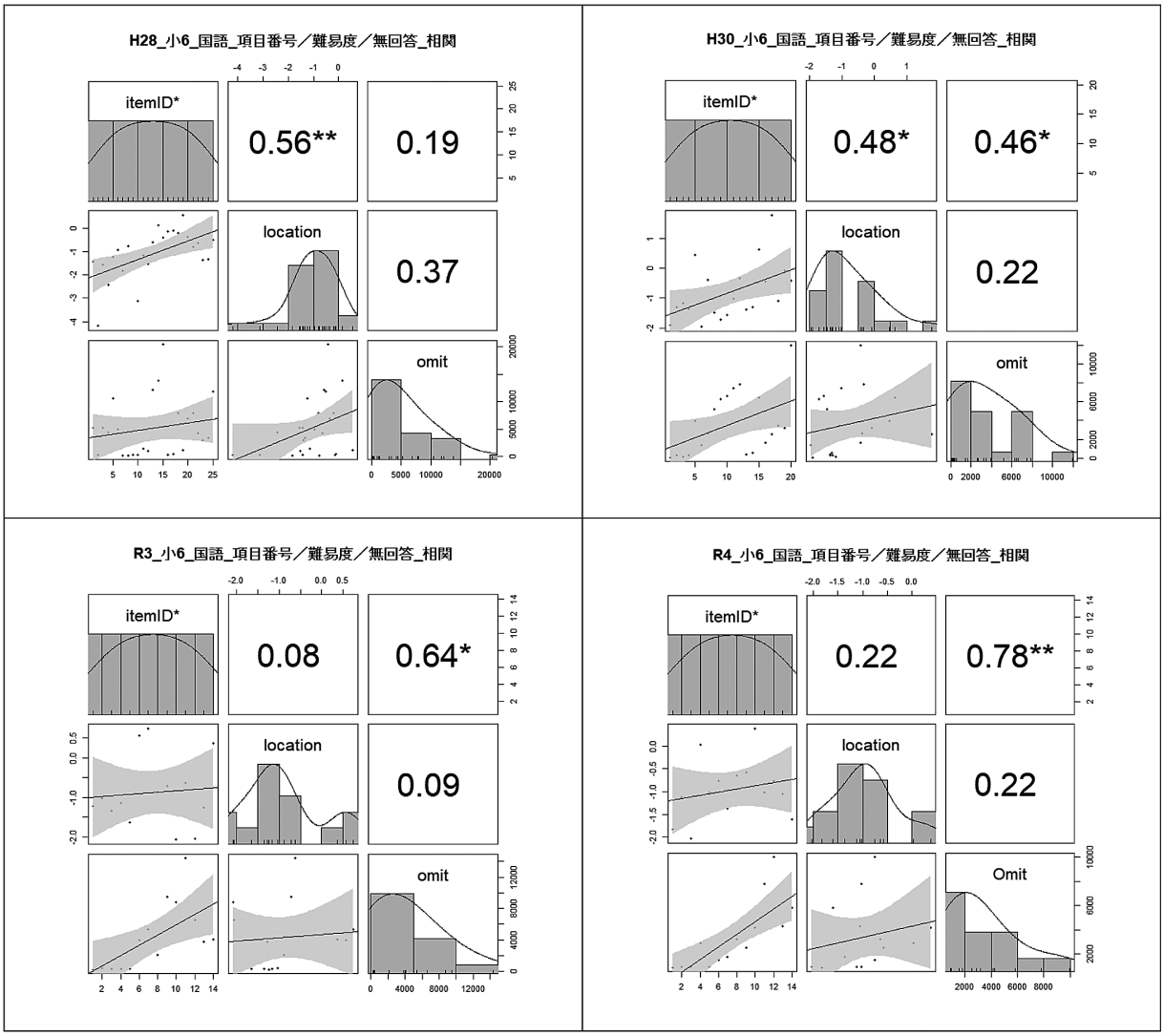

図３：小６国語の項目番号、難易度、無回答の相関関係

令和３年度の小６国語は、項目番号と難易度との相関係数が0.08となり、相互に相関がない。つまり、難しい問題がテストの前半に出たり、易しい問題がテストの後半に出たりするなど、「易しい問題から次第に難しい問題へ」というテストの基本原則が守られていない。しかし、令和４年度の小６国語では、r=0.22

と緩やかな相関が認められ、基本原則に沿ったテストに改善されている。

これと連動して、令和3年度の小6国語では、問題難易度と無回答数との相関係数も0.09となり、「難しい問題ほど無回答数が多くなる」という自然な現象に反する不自然な結果である。これは、難しい問題がテストの前半に出題されることで、そこに時間を取られ、あるいはそこでくじけてしまい、後に出題される簡単な問題に回答できないためと推察される。これに対し、令和4年度は、項目難易度と無回答率の間にr=0.22という緩やかな相関が見られ、「難しい問題ほど無回答が多い」という自然な現象になっている。ここにもテストの改善がうかがわれる。

次に小6算数の相関一覧を図4に示す。

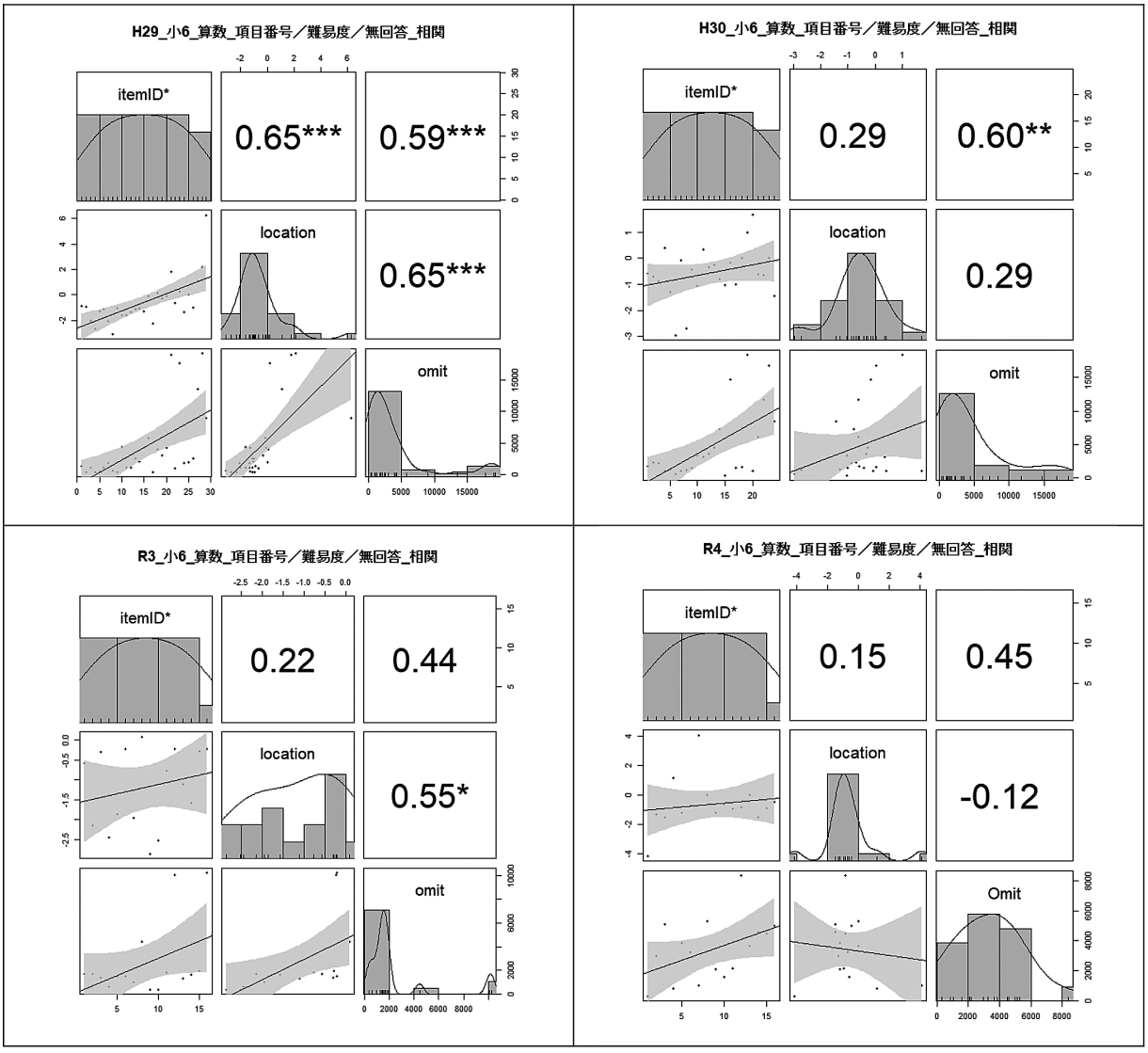

図4：小6算数の項目番号、難易度、無回答の相関関係

令和4年度小6算数では、項目番号と難易度の相関係数が0.15しかない。先行研究が計算した平成27、28、29、30、31年度、令和3年度を合わせると、令和4年度の小6算数ではじめて、相関係数が0.20を下回っている。これと連動して令和4年度は、難易度と無回答数との相関も、-0.12と負の相関になっている。れも改善が必要である。次に中3国語を見てみる（図5）。

中3国語では、令和4年度は令和3年度と同様、項目番号と難易度との間に相関がない。加えて、中3国語は、平成28、30年度でも、問題難易度と無回答数との間に相関がない。連動して、項目番号と無回答数との間にも相関がない年度があったり、負の相関になる年度があったりしている。これらを総合すると、中3国語は、4つの年度いずれも、何か不自然な結果で

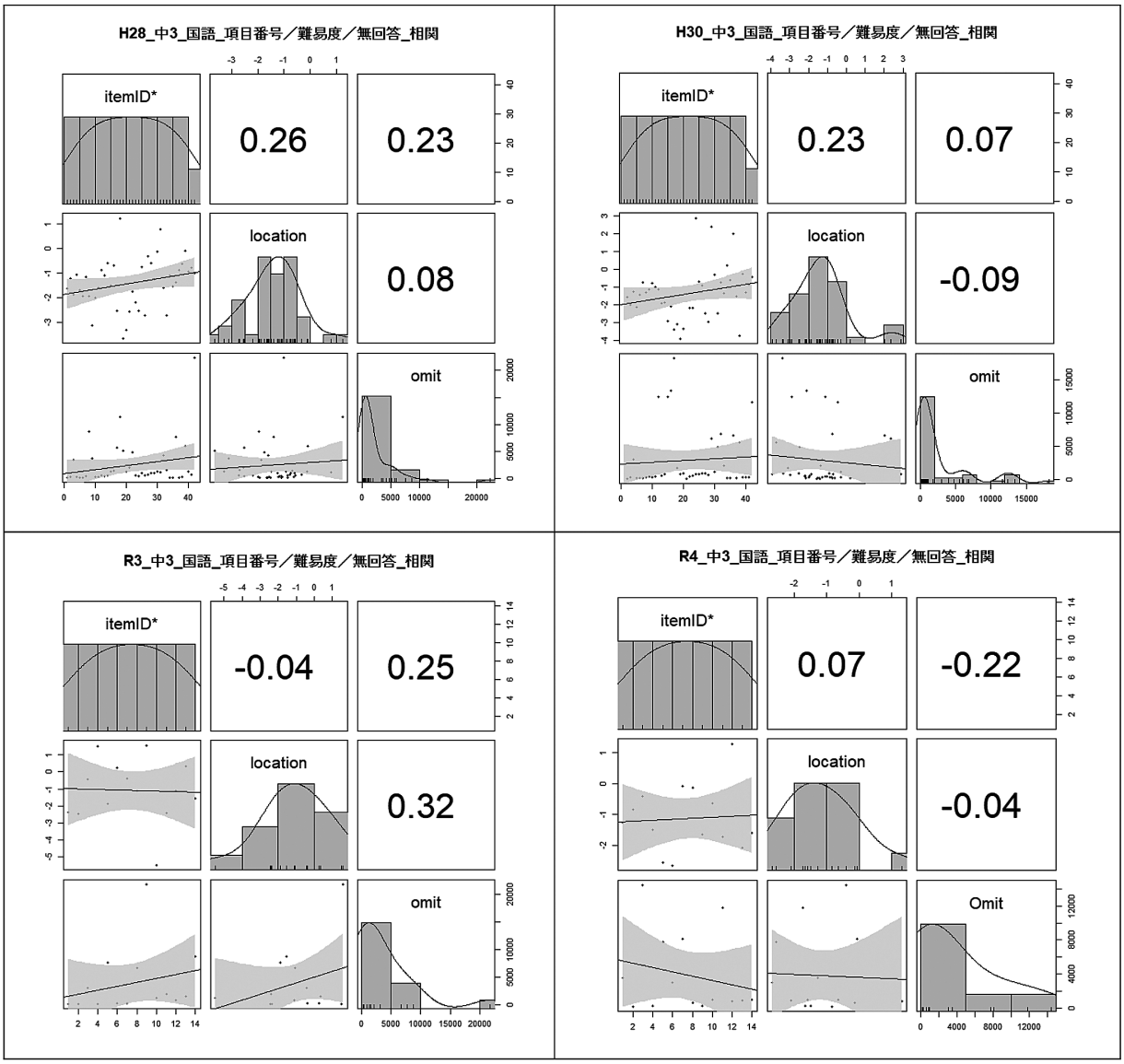

図５：中３国語の項目番号、難易度、無回答の相関関係

ある。

最後に中３数学の相関を見てみよう（図6）。

中３数学は、４つの年度すべてで、項目番号と難易度に一定の相関があり、簡単な問題から難しい問題へという基本原則で出題されている。それと連動して、問題の難易度と無回答数にも一定の相関があり、難しい問題ほど無回答数が多くなる、という自然な結果になっている。さらにどの年度も、項目番号と無回答数との間にも一定の相関があり、テストの後半になるにつれ無回答数が増加するという自然な結果になっている。

６．教科調査の妥当性の検証

最後に教科調査の妥当性を検証する。妥当性とりわけ構成概念妥当性の検証では、同じあるいは類似の構成概念を測定する複数のテスト間のスコアの相関の強さを調べるやり方がある。

> たとえば、新たに開発した心理測定尺度で測定される概念内容（変数x）は、別の概念内容である変数y と理論的に強く関連することが予想される場合、実際に変数x と変数y とを測定したデータをもとに変数x－変数y間に高い相関係数が得られれば、構成概念妥当性を支持する一つの証拠になる、と考えられる。（宮本ほか編, 2015, p.162）

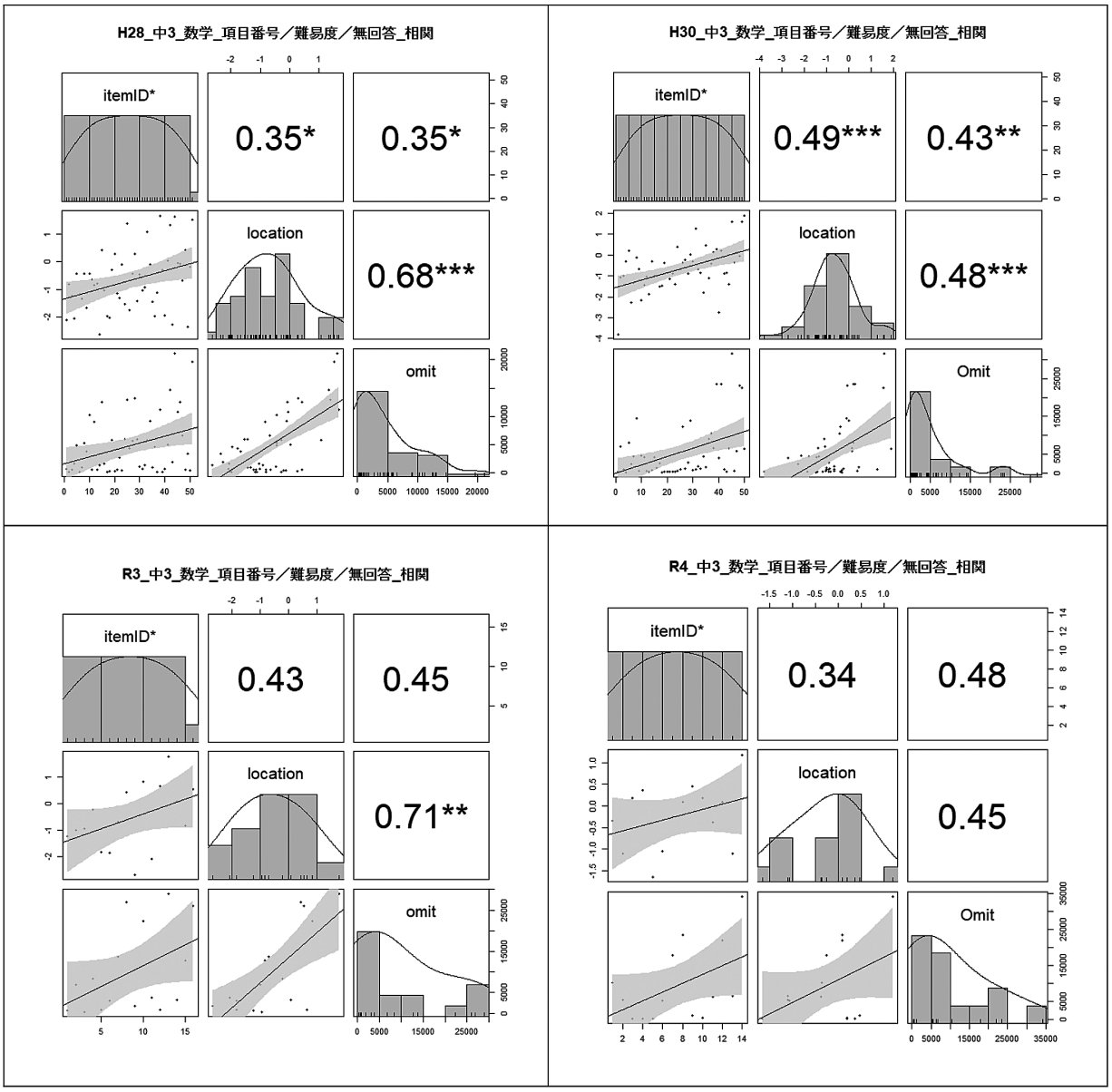

図６：中３数学の項目番号、難易度、無回答の相関関係

全国学力・学習状況調査の場合は、本体調査の教科学力調査に加え、３年に１度ほどの頻度の「経年変化分析調査」でも学力調査が実施されている。実施されたのは、平成25、28年度、令和３年度である。これら３回の経年変化分析調査は、「項目反応理論（item response theory: IRT）」と「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（文部科学省, 2021, p.1）[8]。

そこで、経年変化分析調査で抽出された児童生徒に注目し、これら児童生徒の本体調査教科スコアと、経年変化分析調査教科スコアとの相関係数を計算するならば、本体調査の妥当性を検討できると考えた。

平成25、28年度については、この相関係数がすでに東北大学の調査研究によって算出されている（cf., 東北大学, 2018, pp.115-117）。平成25年度は小６中３の国語と算数・数学すべてで２分冊、平成28年度は全13分冊であった。本体調査（AB問題合算）と各分冊との相関係数の平均をとると表４のH25列およびH28列になる。転記にあたり小数第３位を四捨五入した。

8　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」されている（文部科学省, 2021, p.67）。

表4にはさらに、令和3年度の本体調査と経年変化分析調査の教科スコアの相関係数も独自に算出し併記した。表4のR3列である。令和3年度の経年分析調査貸与個票データには、児童生徒ごとに、本体調査の教科正答数と、経年変化分析調査の教科推定値（本稿の計算ではMLEを利用）とが算出されているため、それらの相関係数を計算した。なおこの相関係数算出にあたり教処理した人数（度数）は、小6国語N=16,198、小6算数N=15,955、中3国語N=24,798、中3数学N=24,733である。

表4：本体調査と経年変化分析調査の教科スコア相関係数

教科	校種	H25	H28	R3
Jpn	EL_Jpn	0.75	0.77	0.70
	JH_Jpn	0.79	0.77	0.70
Math	EL_Math	0.80	0.78	0.76
	JH_Math	0.86	0.86	0.81

全体として、本体調査と経年変化分析調査の教科スコアの相関係数が、令和3年度に低下していることがわかる。ただその評価は難しい。同じ教科の異なるテストの相関係数の「基準値」[9]の研究・議論がほとんどなされていないためである。ここでは今後の議論のための試論を提示しておきたい。

表4の相関係数の評価に関しては、同じ教科の調査でありながら、その相関係数が0.80に届かないことから、本体調査の精度を疑問視することもできる。一方、調査の設計が異なり、テスト精度のかなり高い経年変化分析調査に対し[10]、問題数14問程度の本体調査の相関が0.70以上あることをむしろ評価する見方もできる。

これら相反する評価を天秤にかけるなら、本体調査の妥当性に「大問題」があるとは言えない、という評価が穏当ではないだろうか。このように暫定的に評価したうえで、問題の程度を見積もるために、異なる観点から議論を重ねておきたい。

日本の児童生徒の異なる教科間の相関係数については先行研究がある（田端ほか，2022, p.102; 田端, 2023, p.121）。例えばそこでは、令和4年度全国学力・学習状況調査の小6国語と算数の相関係数が0.70と示されている（田端，2023, p.121）。0.70という値は、表4の令和3年度国語（小中）の相関係数、つまり同一教科での本体調査と経年変化分析調査との相関係数と同じである。異なる教科間の相関係数が、同一教科の異なる調査間の相関係数と同じ、という事実をどう受け止めるのがよいだろう。

先行研究の計算の確かさをチェックするためにも、本稿独自に異教科間の相関係数を、全国学力・学習状況調査匿名データにより算出する。エクセルのデータ分析ツールを用いた。本稿が焦点化する4か年度の結果は表5になる。「Sci」の記号は理科を表す。データがない個所（N.A.）は「—」を入れた。

表5：異教科間の相関係数

校種	教科間	H28	H30	R3	R4
EL	Jpn-Math	0.77	0.76	0.70	0.71
	Jpn-Sci	—	0.72	—	0.72
	Math-Sci	—	0.70	—	0.74
JH	Jpn-Math	0.75	0.72	0.69	0.67
	Jpn-Sci	—	0.74	—	0.65
	Math-Sci	—	0.78	—	0.72

問題数が多かった平成28、30年度の方が、問題数が減少した令和3、4年度よりも、異教科間の相関係数も高い。ちなみに前者の相関係数平均は0.74、後者の平均は0.70である。

構成概念同士の差異は、異教科間の差異の方が、同一教科での異調査間の差異より大きいと考えるのが自然である。国語力と数学力との違いの方が、数学力を測定するA調査とB調査との違いより大きいのが自然である。そうすると、本体調査の妥当性の指標となる相関係数が0.70というのは、好意的に見て最低ライン、どちらかと言えば改善が必要なラインと評価するのが妥当ではないだろうか。

9　基準値とは何か、またそれはどのように設定されるかについては、村上ほか（2019）が参考になる。また田端（2023,pp.96-97）も参照。
10 経年変化分析調査の精度が本体調査よりも格段に高いことについては、文部科学省,2021,p.63を参照されたい。

表6：3つの観点からの品質評価一覧

	平成28				平成30				令和3				令和4			
	小6		中3		小6		中3		小6		中3		小6		中3	
	国語	算数	国語	数学	国語	算数	国語	数学	国語	算数	国語	数学	国語	算数	国語	数学
信頼性係数	＋	＋	＋	＋	＋	＋	＋	＋	－	－	－	－	－	－	－	－
出題順–難易度相関	＋	＋	＋	＋	＋	＋	＋	＋	－	＋	－	＋	＋	－	－	＋
難易度–無回答相関	＋	＋	－	＋	＋	＋	－	＋	－	＋	＋	＋	＋	－	－	＋

7．結論と提言

以上、全国学力・学習状況調査の教科に関する調査につき、平成28、30年度、令和3、4年度の調査項目の品質を、複数の観点から検証してみた。本稿各章での評価は繰り返さない。基準値を定めた3つの観点からの評価をまとめると、表6になる。3つの観点とは、「信頼性係数」「出題順と問題難易度の相関」「問題難易度と無回答の相関」である。基準値を下回った場合を「－」、基準値に届いたか上回った場合を「＋」と評価した。

3観点すべてでマイナス評価となったのは、令和3年度小6国語と令和4年度小6算数、ならびに令和4年度中3国語である。令和3年度小6国語は、本体調査と経年変化分析調査の相関係数も低かった（表4参照）。

令和3年度中3国語は、2観点でマイナス評価である。本体調査と経年変化分析調査の相関係数も低い（表4参照）。

令和4年度小6国語は、2観点でプラス評価だが、それぞれの相関係数は0.22であり、相関があるといえるギリギリである。出題順と問題難易度や、問題難易度と無回答難易度には、もっと強い相関があってよい。

以上から、令和3年度小6国語と中3国語、令和4年度小6国語と算数、ならびに令和4年度中3国語には、改善が必要であると結論づけたい。この結論は、提言でもある。改善のためには、3観点でプラス評価となった調査問題が参考になる。

全国学力・学習状況調査は、毎年小学6年生と中学3年生を対象として悉皆で実施される、社会的意義とインパクトがきわめて大きな調査であり、得られたデータは「国民の宝」と言って過言ではない。全国の自治体や学校で、この調査をもとに毎年、教育施策の成果と課題を検証し、指導や学習の改善を図っている。これほど重要な調査であればこそ、調査問題の品質を今後も継続・発展させ、できる限りの品質向上への取組が必要である。

【付記1】本稿は、科学研究費助成事業、基盤研究B「学力／非認知能力を効果的に育成する教育リーダーのデータサイエンス」（2023-2025年度、課題番号：23H00921、研究代表者：田端健人）の研究成果の一部である。

【付記2】本稿の図表はすべて筆者が独自作成したものである。文部科学省が作成・公表した資料や数値については出典を明記した。

【付記3】文部科学省「個票データ等の貸与利用規約」に則り、本稿は公表以前のしかるべき時期に、文部科学省総合教育政策局による事前確認を受けている。

引用文献

光永悠彦，2018『テストは何を測るのか―項目反応理論の考え方―』ナカニシヤ出版．

宮本聡介・宇井美代子編，2015『質問紙調査と心理測定尺度―計画から実施・解析まで―』サイエンス社．

文部科学省，2021「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」．https://www.nier.go.jp/21chousakekkahoukoku/kannren_chousa/pdf/21keinen_tech_01.pdf [2023.09.14最終閲覧]

村上道夫・永井孝志・小野恭子・岸本充生，2019『基準値のからくり―安全はこうして数字になった―』講談社．

村山航，2012「妥当性―概念の歴史的変遷と心理測定学的観点からの考察―」『教育心理学年報』第51集，118-130.

柴山直，2020「大学院 教育測定学概論05 古典的テスト理論」．https://researchmap.jp/sbym_tds/works/37056546 [2023.09.14最終閲覧]

田端健人，2023「『教育の現象学』のデータサイエンス的転回―全国学力・学習状況調査結果の分析から―」『学ぶと教える

の現象学研究』, 20号, パイデイア出版, 64-130.
田端健人編著, 2022『IRT分析ソフト EasyEstimation による全国学力・学習状況調査の検証と経年比較』パイデイア出版
田端健人・菅原敏・板垣翔大・原田信之・丸山千佳子・久保順也・本図愛実, 2022「学力／非認知能力に対する対話・探究学習効果のデータサイエンス―全国学力・学習状況調査の分析を中心に―」『宮城教育大学教職大学院紀要』, 第4号, 91-109.
東北大学, 2018「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（成果報告書）その4」, pp.115-117. https://www.mext.go.jp/component/a_menu/education/micro_detail/__icsFiles/afieldfile/2018/08/27/1408400_4_1.pdf [2023.09.14最終閲覧]

（令和6年2月6日受理）

The Quality Verification of National Assessment of Academic Ability

—The Comparison of 2016, 2018, 2021, and 2022—

TABATA Taketo

Abstract:

The purpose of this paper is to examine the quality of National Assessment of Academic Ability (NAAA) by MEXT. Chapter 1 pointed out that in light of the purpose of NAAA, it is essential to verify the accuracy of the "test" that is the measurement tool. Chapter 2 referred to test theory and explained that the quality of a test lies in its validity and reliability. In Chapter 3, Cronbach's α coefficient was calculated to verify the reliability of NAAA. In order to critically develop the previous research on the quality verification of the same survey, the years targeted for verification in this paper are 2016 and 2018, when the number of questions was high, and 2021 and 2022, when the number of questions decreased. The data is anonymous data lent by MEXT and consists of approximately 100,000 students in each grade each year. The standard values were set as Japanese language (Jpn) α =0.75 and mathematics (Math) α =0.80, following the previous research. As a result of the calculations, in the 2016 and 2018 tests, all Grades 6 and 9, all subjects Jpn and Math passed the standard, but in the 2021 and 2022 tests, they all fell below the standard. Chapter 4 compares the four years in terms of test information amount and measurement error. The IRT analysis software EasyEstimation was used for the analysis, and the output values were visualized using the statistical software R. The results showed that in the 2021 and 2022 tests, where the number of questions was reduced, the amount of test information decreased, and the measurement error increased. Chapter 5 calculated and visualized the correlation coefficients of “item number”, "difficulty", and "non-response", and verified whether the basic principles of a good test were being followed. As a result, the basic principle of “ordering questions from easy to difficult” was not observed in Grade 6 Jpn in 2021, Grade 6 Math in 2024, and Grade 9 Jpn in 2021. On the other hand, in Grade 9 Math, this principle was followed in all four years. In addition, by analyzing the correlation between item difficulty and non-response, we found a result that goes against the natural phenomenon that “the more difficult the problem, the more non-answers there are”, in Grade 6 Jpn in 2021, in Grade 6 Math in 2022, and in Grade 9 Jpn in 2016, 2018, 2022. In Chapter 6, to verify the validity of the test, we calculated the correlation between individual scores in the same subject between the main survey and the secular change analysis survey. The verification target was 2013, 2016, and 2021, when the secular change analysis survey was conducted. The maximum correlation coefficient was 0.86 and the minimum was 0.70. In 2021, the correlation coefficients for all Jpn and Math in the Grade 6 and 9 decreased. However, the evaluation of this correlation coefficient requires caution due to the lack of discussion regarding its “standard value”. Therefore, in this paper, we calculated the correlation coefficient between different subjects and estimated the correlation coefficient between different surveys on the same subject. As a result, we tentatively concluded that the correlation coefficient of 0.70 between different surveys on the same subject is a value that should be improved. Chapter 7 summarizes the above conclusions and recommends improvements.

Key Words : Validity and Reliability of Tests, Amount of Test Information, Item Error,
Secular Change Analysis Survey