福岡教育大学紀要，第70号，第４分冊，19―27（2021）

教育行政が有するデータを利用したパネルデータの設計と分析

Panel Data Analysis with Local Government Survey Data

川　口　俊　明

Toshiaki KAWAGUCHI
学校教育ユニット

（令和２年９月30日受付，令和２年12月10日受理）

1．学力パネルデータの課題

2020年現在，日本でもようやく学力格差（苅谷・志水2004）を始めとする「教育の格差」が重要な社会問題として認識されるようになってきた（松岡2019，志水2020）。教育政策においてEBPM（Evidence Based Policy Making: 証拠に基づく政策立案）が重要性を増す中，近年の教育研究では一時点の格差の実態把握のみならず，「いつ／なぜ格差が生じるのか」「学力の格差は時を経て縮小するのか／拡大するのか」といった中長期にわたる変化に踏み込んだ分析が行われるようになっている（数実2017，川口・松尾・礒部・樋口2019）。

もっとも，同時に指摘しておかなければならないのは，日本の小中学校教育においては，こうした分析の基盤になる肝心のデータの蓄積が進んでいないという点である。学力研究もその例外ではなく，一般の研究者が利用可能なデータは数えるほどしか存在していない（川口2020a）。

このような状況で有効活用が期待されるのが，教育行政が蓄積してきた学力調査のデータである。全国学力・学習状況調査（以下，全国学力テスト）を筆頭に，教育行政が実施する学力調査は少なくない。ただ，日本の教育行政が実施する学力調査は，諸外国の教育研究で一般的な教育測定の知見に基づいているわけではない。全国学力テストはその典型的な例だが，IRT（Item Response Theory: 項目反応理論）を利用しているわけではないし，出題するテスト項目がテスト理論の専門家によって検討されることも稀である（川口2020a）。「学力調査の時代」（苅谷・志水2004）とは言うものの，果たしてそこで測定されている学力が信頼できる数値なのかどうかという根本的な問題は，十分に検討されてきたとは言えない。

この問題は，早くから学力研究に取り組んできた教育社会学の学力研究においても同様である。PISAやTIMSSといった国際的な学力調査を利用したものは別として，国内の学力調査を利用した研究では，調査の質が担保されているのかどうかという問題が常につきまとう。これまではデータの入手が難しいという理由で，調査の質という問題は等閑視されることが多かったとはいえ，いつまでもこの状況を放置するべきではない。何より，中長期にわたって同一個人の学力の変化を追う分析（いわゆる，縦断データの分析）では，異なる時点間で行われた学力テストが「同一の学力」を測定しているという前提が必要になる。そのため，一時点の学力データ（横断データ）の分析とは違い，学力テストの質という問題は，まさに喫緊の課題であると考えなければならない。

以上のような状況を踏まえ，本稿では，日本の教育行政が実施した学力調査の縦断データの分析を例にとり，教育行政が実施した学力調査をもとに研究を行う際に留意すべき点について論じる。分析対象とするのは，筆者が西日本のある自治体（いろは市：仮称）で行った，いろは市学力パネルデータある。同調査は2016年度の小学４年生を継続的に追跡していく調査プロジェクトの一環であり，いろは市教育委員会が独自に実施する学力調査・生活実態調査や全国学力テストのデータに，筆者らが実施する独自の生活調査・保護者調査を組み合わせ，子どもたちの学力の変化を把握できるパネルデータとしたものである（図1）。もっとも，全国学力テストはもちろん，いろは市教育委員会が実施する学力調査・生活実態調査も，もともとパ

ネルデータとすることを前提に設計されたものではない。そのため，同一個人のデータを接続する際に，手がかりとなる情報が欠落したサンプルは接続できない等の課題はあるが，データの蓄積が乏しい日本の学力研究では，貴重なデータと言える。以下では，このデータを利用し，行政が実施する学力調査を縦断研究に利用する際に考慮すべき点について検討する。なお，いろは市パネル調査の概要については，川口（2020b）を参照してほしい。

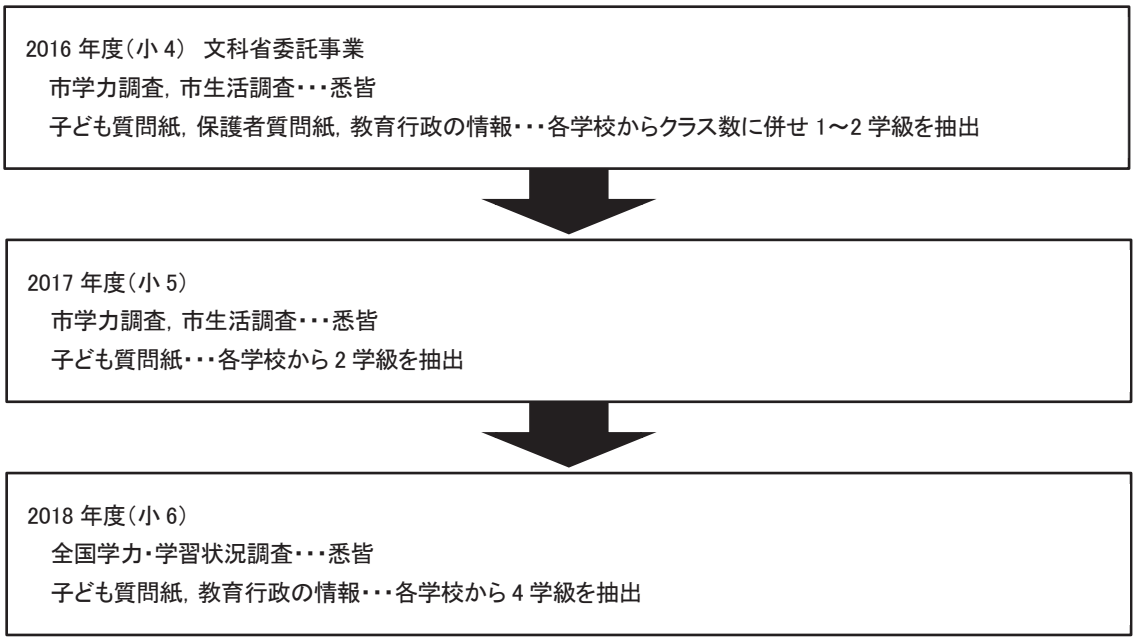

図 1．パネルデータの構造

※川口（2020b）より再掲

2. 教育行政と研究者の関心の乖離

日本の教育行政が実施する学力調査を縦断データとして研究に利用する場合，まず考えなければならないのは，教育行政と研究者のあいだで学力調査に求めるものが相当に乖離しているという点である。日本の場合，各学年で履修する事項が文部科学省の学習指導要領によって定められている。そのため，教育行政が実施する学力テストは，その学年までに履修した内容の定着を確認するという「確認テスト」の性質を帯びやすい。結果として，学力テストで測定されている「学力」の概念的定義はそれほど重要視されないし，テストの結果も全問正答の子どもが多いほど望ましいということになりがちである。他方で縦断データ分析では，学力テストで測定されている「学力」が，身長や体重のように学年が異なっても比較可能であることが大前提である。そのため，学力テストで測定されている「学力」の概念的定義は明確でなければならないし，学力テストで測定されている数値が異なる学年でも比較可能な一貫性を持っているかどうかも重要な問題となる。加えて，学力研究ではさまざまな学力層の子どもたちに対する施策や実践の影響を検討するから，テスト結果は満点や 0 点に偏ることなく，満遍なくバラついていることが好ましい。

こうしたズレがあるため，教育行政が実施する学力調査を縦断研究に利用する場合，少なくとも次のような点を考慮しなければならない。おそらくもっとも深刻な問題は，そもそも教育行政が実施している学力テストは，一体何の学力を測っている（と考えてよい）のかという点である。算数・数学，あるいは英語であれば，小学校・中学校を通して継続的に発達する数的能力・外国語の運用能力といった「学力」を想定することはそれほど不自然ではない。しかし，それ以外の教科の場合は，やっかいである。たとえば中学校の国語は，現代文・物語文の読解だけでなく，古文や漢文に関する知識を含んでいるが，そこに小学校の国語とも一貫するような「学力」は想定できるだろうか。あるいは，小学校の社会科の場合，学年に応じて，地理・日本史・世界史など異なる内容を扱うことが一般的であるが，そこに学年の進行とともに向上する「学力」を想定してかまわないのだろうか。

以上のような問題は，テスト理論において学力テストの妥当性の問題として扱われる（光永 2017）。妥当

性の問題は，統計的な枠組みだけでは解決することができず，教科教育の専門家の知見も借りなければならない。昨今では，埼玉県学力・学習状況調査のように，項目反応理論を導入することで，「学力の伸び」を測ることを謳う学力調査を行う自治体も存在する（石川ほか 2017）。ただ管見の限りでは，こうしたテストの設計に教科の専門家が加わって，そもそも「何の学力を測るのか」という点に踏み込んだ考察は行われていないようである。推測ではあるが，これは，あまりにテスト理論を重視し測定する学力の妥当性にこだわると，今度は学習指導要領に示された内容を網羅できなくなってしまい，教育行政が必要とする学習指導要領の定着を確認できなくなる恐れがあるためだと思われる。

「学習指導要領の定着」を網羅的に確認するテストと，身長のように学年の進行とともに向上していく学力を測るテストのあいだには，どうしてもある種の乖離が存在する。学習指導要領は日本の学校教育の在り方に深く結びついているから，その乖離をすぐに改善することはできないだろうが，学力の縦断研究のためには，どこかで落とし所を探す必要がある。教科教育の専門家とテスト理論の専門家の協働が求められていると言えよう。本稿では，とりあえずの対応として，「どのような学力を測るのか」という点で乖離が比較的少ないと思われる，算数・数学や英語を中心に論を進める。

もっとも教科を限定したとしても，それで問題が終わるわけではない。妥当性の問題に続いて考えなければならないのは，教育測定の知見をもとに学力調査を縦断データに「翻訳」する過程で，もともとのテストの性質が異なるためにさまざまな留保や選択を行わざるを得ないという点である。すでに指摘したように，教育行政が実施する学力調査は，もともと全員が満点をとることを理想として設計されている。そのため，あまり難易度の高いテスト項目は出題されず，平均正答率も高めになりやすい。これは，受験者の学力が満遍なくバラつくことを求める研究者の立場から見ると，偏りのあるテストになっていることを意味する。

また，大規模な学力調査で一般的に利用されている IRT の 2 パラメータ・ロジスティックモデル（2PL）は，その前提に能力の一次元性や項目の局所独立といった仮定をおいている。前者は，テストで測定した学力は一種類だけという仮定である。たとえば算数のテストは算数のテストの学力しか測っておらず，そこに理科や社会の学力は混入していないというものだ。統計的な作業としては，質的因子分析などを行い，適切なテスト項目でテストが構成されているかどうか確認しなければならない。後者は前のテスト項目の成否が後のテスト項目の成否に影響を及ぼさないという仮定である。たとえばセンター試験のような大問形式では，前のテスト項目を解くことが，後のテスト項目の前提になっている。こうしたテスト項目は IRT を利用した場合は出題できない。

もっとも前者については複数の能力を含んだモデルを利用することもできるし，後者についても部分採点モデルを利用することで対応できる（豊田編 2013）。ただ推定が煩雑になるし，どのようなモデルを選択するかによって，最終的に推定される学力に違いが出てくる。変換の過程では，どのような選択をしたか常に明示するか，あるいは複数の変換方法を試し，個々の結果を報告する必要があるだろう。

加えて，本稿が対象とするような複数時点の学力の変化を把握しようとする場合，異なる時点間の学力をどうやって比較可能にするかという問題もある。教育行政が実施する学力調査では，異なる学年には異なるテストが出題されている。テスト時点までの学習指導要領の定着を確認しているのだから，教育行政の立場から見れば，この選択がおかしいというわけではない。ただ，学力の縦断研究を行う研究者の視点から見ると，出題されたテスト項目がまったく違うのでは，異なる学年間のテストの成績を比較することができない。そのため何らかの方法を利用して，尺度をそろえる必要がある。おそらくもっとも一般的な解決策は，複数のテストを同時に受験する共通受験者を用意し，個々のテストの難易度を調整するという方法だろう（光永 2017）。ただ，この場合も，誰を共通受験者とするのか，個々のテストの難易度をどのような手法で調整するのかなど，いくつかの選択肢が存在する。そして，こうした選択は，当然のように最終的な推定結果に影響を及ぼす。そのためここでも，なぜ／どのような選択を行ったか示す必要がある。

ここまでの議論でわかるように，教育行政の実施する学力テストを縦断研究に利用できるデータに変換すること自体は不可能ではない。ただ，変換の途中で行われるさまざまな選択によって，最終的な結果が大きく変わってしまう可能性がある点には留意が必要だし，その手順はどこかに明記しておく必要がある。

ここまでの考察はテスト理論に基づいたものだが，最後にもう一つだけ変換を行うにあたって，考慮すべき点を指摘しておきたい。それは，学力の社会的構成という視点である。現在，日本社会や小中学校で流通している学力とは，あくまでも素点をもとにした正答率である。おそらく，学習指導要領の定着を正答率で測ることがあまりにも自然であるがために，項目反応理論という「科学的なテスト理論」は必要とされな

かったのであろう。こうした学力（能力）の社会的構成に関心を寄せる中村（2018）は，日本社会が，項目反応理論という科学的な得点ではなく，「素点の持つ圧倒的なリアリティのなかに「能力」を構成してきた」（p.120）ことを指摘している。

学力に関する縦断研究を行う場合は，ここまでに述べてきたようなテスト理論に基づいた変換を施すことが，学術的に望ましいことは明らかである。しかし，こうして得られた成績が，統計的には正しくても社会的にはリアリティの薄い数値でしかないとすれば，こうした変換の努力も日本社会や小中学校現場から見ると「研究者の自己満足」にしか映らないだろう。このような関心の差をどうやって埋めればよいのだろうか。私見ではあるが，EBPM の必要性が叫ばれ学力の縦断研究が求められる一方で，その要となる学力を測る方法が社会的な信認を受けていないという矛盾に，多くの人は気づいていないように見える。何か抜本的な改善策があるわけではないが，学力の縦断研究を行う上で，容易には解決できない課題があることは指摘しておきたい。

3. 既存データを利用した分析の実際

3.1. 基本的な分析

それでは，いろは市パネルデータを利用し，学力テストの成績を縦断分析が可能な形に変換する作業を行ってみよう。今回は，異なる学年のテストを比較可能にするため，2016 年度の小学 4 年生が，2016 年度・2017 年度・2018 年度に受験したテスト項目を元に作成したテスト（以下，アンカーテストと呼ぶ）を受験してもらい，その結果を利用して，各回のテストの難易度を調整（これを垂直尺度化と言う）することにした。アンカーテストの受験者は，2019 年度の小学 6 年生としている。小学 4 年生のテスト項目を小学 6 年生が受験すると，あまりに簡単すぎるという可能性はあるのだが，学校現場の負担を考えると，複数の学年に調査依頼を出すことも難しい。そのため小学 4 年生・5 年生・6 年生のすべてのテスト項目を学習済みの小学 6 年生を対象にテストを実施した。なお，受験の時期は，全国学力・学習状況調査の実施時期が 4 月末頃であることを考えると，同じ時期が望ましいのだが，学校への依頼の時期や行事との兼ね合いもあり，2 学期末に実施することになった。これはやむを得ない事情だと思うが，こうした選択が推定に影響を及ぼす可能性はある。まず，全国学力・学習状況調査は 4 月末に実施されているため，子どもたちの学力が学年の進行とともに向上していくと仮定するなら，小学 6 年生の 2 学期には小学 4 年生・小学 5 年生のテストはもちろん，全国学力・学習状況調査でさえ簡単すぎる可能性がある。この点が推定に及ぼす影響については，後ほど考察する。

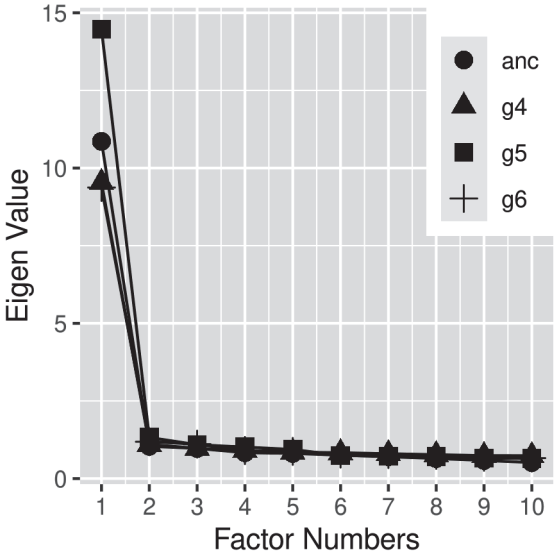

図 2. スクリープロット

アンカーテストの作成については，各年度のテスト項目について 2PL による項目分析を行い，できるだけ識別度が高い項目を中心に，難易度ができるだけバラつくように選び出した。いずれの IRT も学力調査で基本的な 2PL を設定し，受験者の能力値には標準正規分布を仮定した。なお，先に述べたように，アンカーテスト受験者の能力が高い可能性があるので，できるだけ高い難易度のテスト項目を選ぶようにしている。

続いて 2PL モデルの前提である，尺度の一次元性を確認するために質的因子分析を行う。図 1 はスクリープロットを示したもので，意外なほどに一次元性は担保されていることがわかる。それぞれの学年のテストはもちろん，アンカーテストについても一次元性が担保されているため，小学 4 年生から小学 6 年生に至るまでの，算数の学力を想定することは，少なくとも統計的な観点からは問題ないと言える。

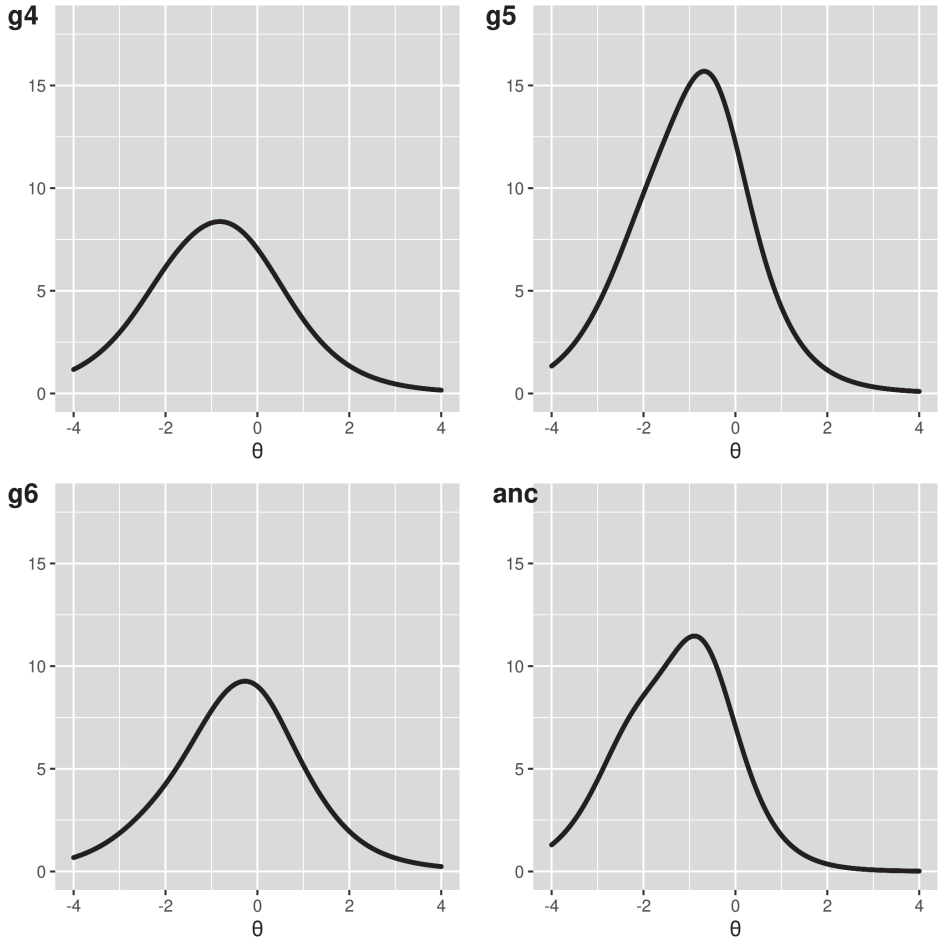

図 3. テスト情報量

図 3 は，それぞれのテストがどの程度の能力の受験者を弁別するのに適しているかを示す，テスト情報量のグラフである。小学 4 年生から小学 6 年生のテストまで，どのテストも受験者の能力値である θ が－ 1 前後を中心に山ができている。これは偏差値に換算すると 40 点前後の児童を識別することに適したテストになっていることを意味しており，テストが基礎的な内容を確認するものになっていることを示している。もともとのテスト設計が学習指導要領の定着の確認だったことを考えれば，この結果も納得できるだろう。一方で，能力の高い児童については，学力の推定に誤差が大きくなることに注意が必要である。

個々のテストの難易度が低めであることに加え，アンカーテストの受験時期が小学校 6 年生の 2 学期であることから予想されたことではあるが，アンカーテストの難易度も低めに推定されている。テスト情報関数の分布は，能力値の低い方に山ができており，小学 4 年生・小学 5 年生のテスト内容から難易度の高い問題を選んだとはいえ，小学 6 年生にとっては簡単なテストになっていることがうかがえる。当然，垂直尺度化を施した場合も，学力上位層の推定値には測定誤差が大きくなると考えられる。

3.2. 垂直尺度化とそれに関わる問題

以上のことを踏まえた上で，垂直尺度化を行う。垂直尺度化の方法には，小学 4 年生と小学 5 年生，小学 5 年生と小学 6 年生の能力をそれぞれ別に推定する独立尺度調整法と，まとめて推定を行う同時尺度調整法がある（澁谷 2019）。今回はアンカーテストの受験者が小学 6 年生しかおらず，別々に推定する必要が薄いために，同時尺度調整法を採用した。推定には，R の mirt パッケージを利用する。

IRT の能力推定法もいくつかあるが，今回は EAP と PVs（Plausible Values: 推算値法）による推定を行う。通常の IRT 分析では，個人の能力推定に関心があるため EAP や WLS 等の点推定値を利用するのが一般的である。ただ，教育社会学などの学力研究では，個人の能力に関心があるというよりは，黒人と白人の学力差といった集団の能力推定に関心を寄せている。この場合，EAP や WLS などの推定法では標準誤差に偏りが生じることが知られており，PISA や TIMSS などの大規模学力調査では PVs の使用が推奨されている（袰岩ほか 2019）。PVs は個々人の能力分布からあり得る複数の能力推定値を算出し，それを母集団の推定に利用する技法であり，今回は PISA2012 にならい，5 つの PVs を生成した。推定後の能力値は，表 1 のようになる。なお，煩雑になるため 5 つの PV のうち，1 つめの PV1 の結果のみ表示している。

表 1. 垂直尺度化後の算数学力

	EAP 推定			推算値法（PV1）		
	小 4	小 5	小 6	小 4	小 5	小 6
平均値	50.2	59.4	58.6	50.2	59.5	58.7
標準偏差	9.0	10.2	9.4	9.8	10.9	10.2

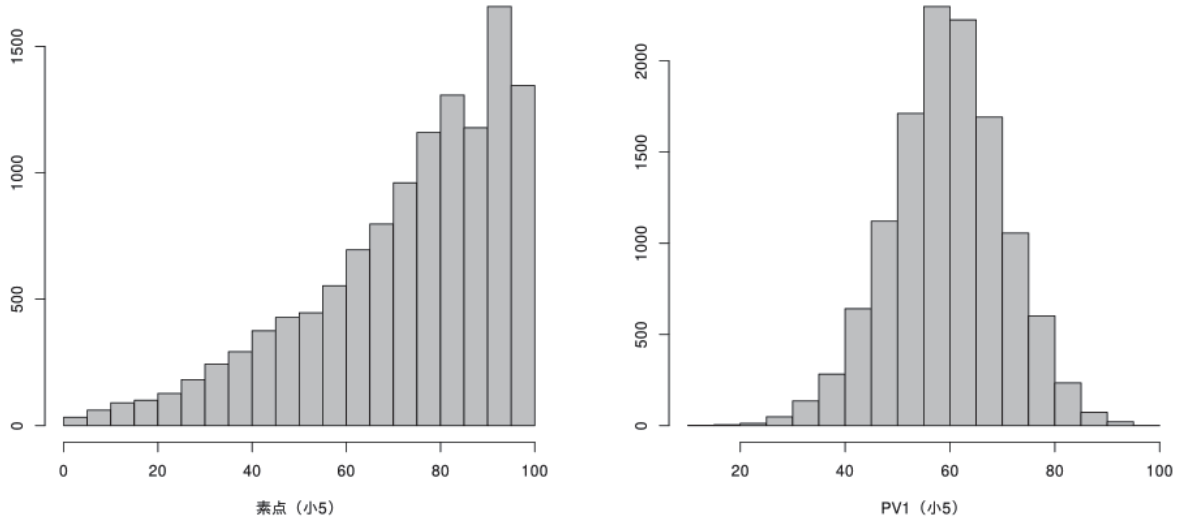

図 4. ヒストグラムの比較

表 1 を見ると，EAP 推定と PV1 で平均値に多少にズレはあるものの，2016 年度の小学 4 年生の成績は，小学 4 年生から小学 5 年生にかけて上昇し，小学 5 年生と小学 6 年生はほぼ同じ成績になっている。確かに学力は向上しているものの，それは小学 4 年生から 5 年生までであり，小学 5 年生と小学 6 年生では学力の向上が見られないという結果になった。こうした現象は垂直尺度化に関する研究では「尺度の縮小」として知られており，学年の進行とともに「成績の伸び」が緩やかになる現象が生じることがあるという（澁谷 2019）。尺度の縮小が生じた理由はいくつか考えられるが，今回の場合，小学 5 年生と小学 6 年生のテスト時期が近く，そもそも能力の向上が検出しづらかったのかもしれない。教育行政が実施する学力調査の実施

時期に対して，研究者側が介入することは難しいが，仮に能力の向上を捉えるのであれば，できるだけ同じ間隔で学力テストを実施することが好ましいと考えられる。

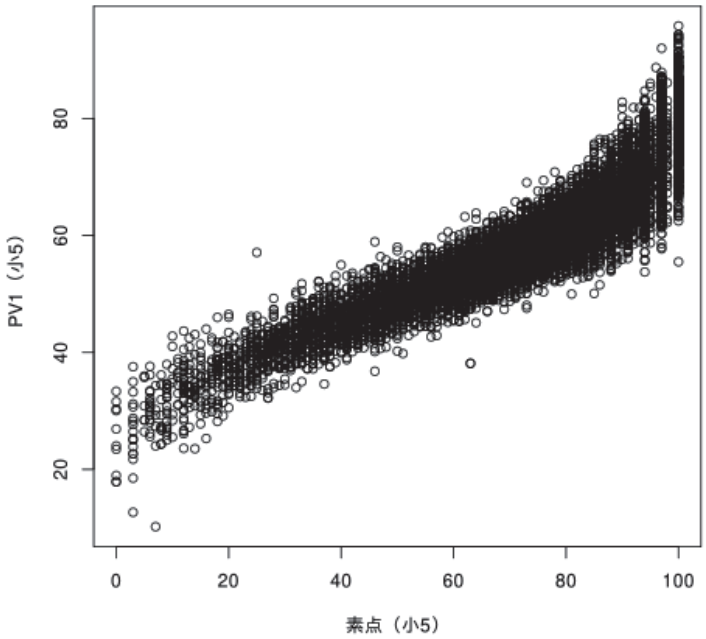

図 5. 素点と PV1 の比較

個々人の得点の推定値についても確認しておこう。素点と PV1 の成績分布は，図 4 のようになる。素点の分布は定着を確認するという目的に応じて，全体に右に偏った分布になっている。一方，PV1 は，正規分布を仮定しているため，正規分布に近い分布になっている。こうした分布の違いは，高得点層の成績の推定値にバラツキが生じるためである。実際，図 5 のように散布図を描いて，素点と PV1 の推定値を比較してみると，とくに高得点層で推定に差が生じている。教育社会学の学力研究では，さまざまな学力層における学力の変化を把握したいわけだから，PVs を利用することは理にかなっていると言える。

なお，PVs と EAP の推定値を比較してみると，EAP は標準誤差を過小評価することが知られている（袰岩ほか 2019）。実際，表 1 を見るとわかるように，EAP 推定の場合の標準偏差は PVs を利用した場合に比較し，常に小さく推定されている。このような問題を回避するためにも，教育社会学の縦断研究においては PVs の利用が好ましいと考えられる。

ただ，学力テストの点数を学校現場に返却することを考えたとき，PVs では個々人の点推定値が得られない点に注意しなければならない。PVs で得られる得点は，ランダムに生成された個々の児童の「あり得た得点」であり，これを成績として返却することはできない。実際，図 4 や図 5 を見るとわかるように，PVs を利用した場合の得点は，特に成績上位層でバラツキが大きく，この結果を個人の成績とするのは無理がある。IRT を利用して個々人に成績を返却することが目的なのであれば，PVs ではなく，EAP 推定など点推定値が得られる推定法を利用すべきだろう。用途に合わせて推定法を使い分けるというのも一つの手ではあるが，同じテストの結果が複数存在するのは，時に混乱を招きかねない。おそらくもっとも有効な解決策は，学力テストを設計する段階で，個人の能力推定に関心があるのか，母集団の推定に関心があるのか確定させておくことである。事前に調査目的が確定していれば，本稿で分析した事例のように，成績上位層の推定値にバラツキが大きいという問題も回避できる。

さて，垂直尺度化の利点は，異なる時点間で測定された学力が，同一尺度上にあることを保障できるというものであった。最後に，こうした垂直尺度化が，どのような実践上・研究上の利点をもたらすか確認しておこう。図 6 は，横軸に小学 4 年生の時点の成績，縦軸に小学 6 年生時点の成績を学校ごとにプロットし，成績の変化を示したものである。図中の線分は，小学 4 年生の成績と小学 6 年生の成績が同じだった場合の線である。図 6 を見ると，1 校を除き，ほとんどの学校が線の上側にある。つまり，ほとんどの小学校は小学 4 年生から小学 6 年生になるまでに成績を向上させることに成功しているということである。IRT を利

用しない場合も，成績の変化を捉えることは不可能ではないが，その場合の成績の変化は，あくまで相対的な位置の変化に過ぎないため，成績が向上する学校があれば，他方に成績の低下する学校が存在することになってしまう。IRT による垂直尺度化を施せば，こうした相対的な競争ではなく，もともとの自校の成績と比較して，成績を向上させることができたかどうかという課題設定も可能になる。相対的な評価では，他者を敵視する排他的競争が常態化するという批判もある（西岡ほか 2015）が，IRT を利用した垂直尺度化を行えば，こうした批判を回避することもできるだろう。

なお，いろは市教育委員会によれば，図 6 で唯一，線分より下にあった小学校は，小学 5 年生時点で学級崩壊を起こし，授業がほとんど成立しなかった学校であるとのことだった。分析結果が判明した後に得られた回顧的な情報であるため，学級崩壊が常に学力に悪影響を及ぼすかどうかはわからないが，学級崩壊と学力の関連について今後検討していく必要があると言えるだろう。

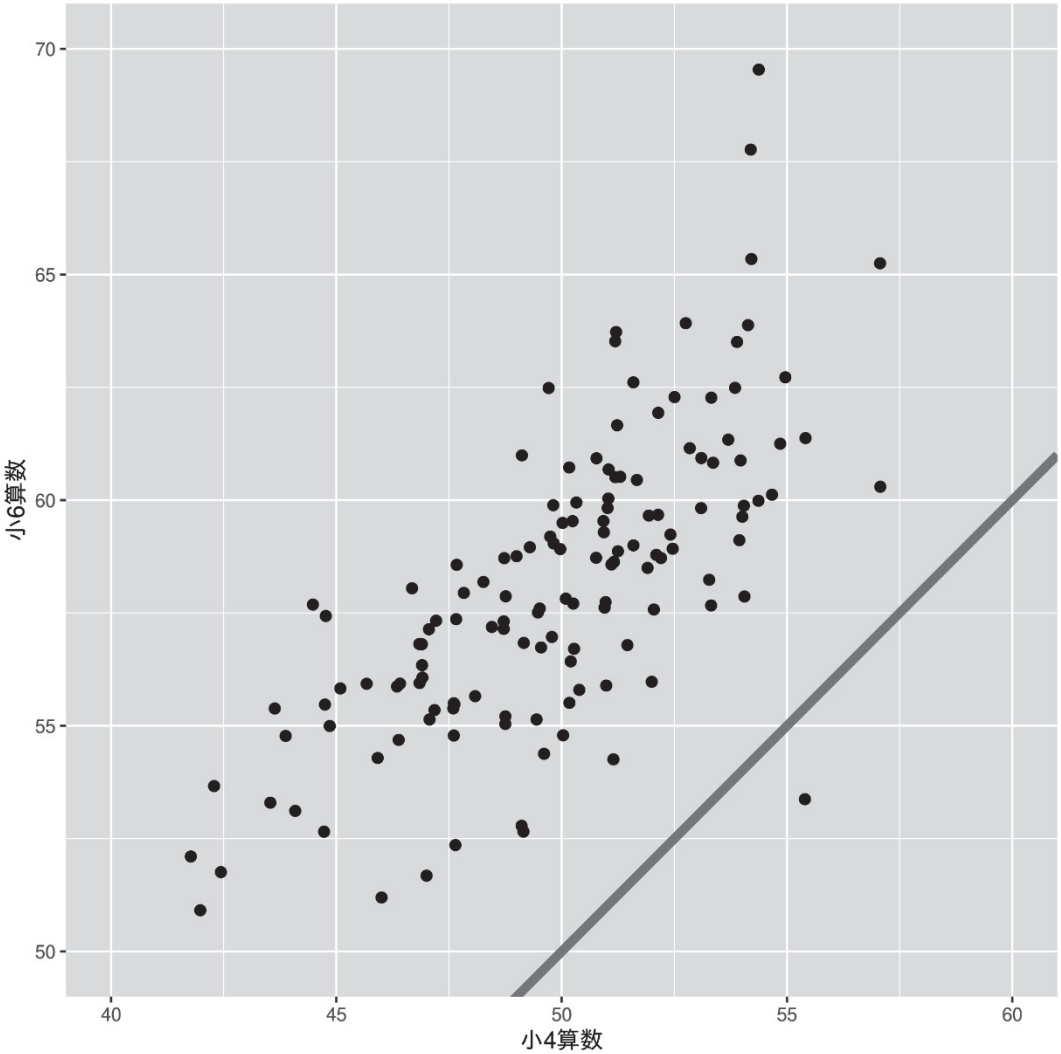

図 6. 成績の「伸び」

4. 考察

本稿では，実データを利用して IRT による翻訳を試みてきた。明らかになったことをまとめておく。まず指摘すべきは，日本の教育行政が学力調査に求めているものと，学力の縦断データの分析を目指す研究者が求めているものが，そもそも乖離しているという点である。少なくとも現時点では，前者が必要としているのは学習指導要領の定着を確認するためのテストであり，必ずしも学術的なテスト理論の前提を満たす必要はない。一方で後者は，テスト理論を前提に，学力の概念的定義の明確さをはじめ，IRT を始めとする統計的な分析に耐えるデータを必要としている。そもそもテストに求めるものが違う時点で，両者のあいだ

には深刻な溝があると考えざるを得ない。

もっとも，こうした溝があるとは言え，教育行政が実施する学力テストを利用して縦断的な教育研究を行うことは不可能ではない。本稿で示したように，教科を限定した上で，尺度の一次元性などの IRT モデルが必要とする前提を満たすように成績を変換することはできる。ただ，その場合でも，成績上位層の推定のバラツキが大きいといった問題はどうしても残る。また，本稿の分析がそうだったように，小学 5 年生と小学 6 年生の学力に伸びが見られないといった，解釈の難しい結果が生じる可能性もある。何より，こうした変換は，「素点の持つ圧倒的なリアリティ」（中村 2018）を失わせてしまう。

この事態を打開するには，おそらく「何のために学力調査を行うのか」という社会的な合意が必要なのだろう。現在，多くの自治体で実施されている学習指導要領の定着を確認するためのテストは，日本の学校教育制度や文化に根ざしたものであり，その変更は容易ではないだろう。しかし，学術的な視点から政策を検討しようというのであれば，テスト理論に則った学力テストがどうしても必要である。そのためには，教科の専門家とテスト理論の専門家の対話が日常化する必要がある。さらに PVs の技法がそうであるように，テスト理論に則った学力テストを実施するのであれば，個々人にテスト結果を返却するという現在の日本の学校教育の慣習は改めるしかない。

筆者は，拙著『全国学力テストはなぜ失敗したのか』において，全国学力テストの失敗を考察し，各学校・学級で行う「指導のためのテスト」と，文部科学省が政策立案の基礎資料とするために行う「政策のためのテスト」を区別する必要があること，及び文部科学省は「政策のためのテスト」に注力すべきであることを論じた（川口 2020c）。ここでいう「指導のためのテスト」とは，学習指導要領の定着を目指す，現在の教育行政の学力テストのことに他ならない。「指導のためのテスト」を続けるのか，それとも「政策のためのテスト」に舵を切るのか。文部科学省のみならず，地方教育行政にとっても，何のために学力テストを行うのか問い直す時期が来ているように思われる。とりあえず教育研究者にできることは，現状を整理し，「何のために学力テストを行うのか」考え直すことが必要であることを繰り返し訴えていくことであろう。

【参考文献】

袰岩晶・篠原真子・篠原康正，2019,『PISA 調査の解剖』東信堂.

石川善樹・伊藤寛武・植村理・田端紳・外山理沙子・中室牧子・分寺杏介・星野崇宏・松岡亮二・山口一大，2017,「子どもの能力を計測するための学力テストの現在と展望」RIETI Policy Discussion Paper Series 17-P-010.

苅谷剛彦・志水宏吉，2004,『学力の社会学－調査が示す学力の変化と学習の課題』岩波書店.

川口俊明，2020a,「学力調査の政治」『教育社会学研究』106 集，pp.55-76.

川口俊明，2020b,「教育行政が有するデータを利用した教育格差の実態把握」『福岡教育大学紀要』69, pp.17-25.

川口俊明，2020c,『全国学力テストはなぜ失敗したのか』岩波書店.

川口俊明・松尾剛・礒部年晃・樋口裕介，2019,「項目反応理論と潜在クラス成長分析による自治体学力調査の再分析」『日本テスト学会誌』15 巻，pp.121-134.

数実浩佑，2017,「学力格差の維持・拡大メカニズムに関する実証的研究－学力と学習態度の双方向因果に着目して」『教育社会学研究』101 集，pp.49-68.

松岡亮二，2019,『教育格差－階層・地域・学歴』筑摩書房.

光永悠彦，2017,『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.

中村高康，2018,『暴走する能力主義－教育と現代社会の病理』筑摩書店.

西岡加名恵・石井英真・田中耕治，2015,『新しい教育評価入門－人を育てる評価のために』有斐閣.

澁谷拓巳，2019,『異なる難易度のテスト項目の IRT 垂直尺度化－尺度化テストデザインによる垂直尺度構成』東北大学修士論文.

志水宏吉，2020,『学力格差を克服する』筑摩書房.

豊田秀樹編，2013,『項目反応理論【中級編】』朝倉書店.