

大規模学力調査における推算値法

Plausible Values in Large Scale Educational Assessment

川 口 俊 明

Toshiaki KAWAGUCHI

学校教育ユニット

(令和6年9月30日受付, 令和6年12月23日受理)

1. PVs (Plausible Values : 推算値) とは何か

2020 年前後から, 日本の小中学校を対象とした学力調査において, IRT (Item Response Theory : 項目反応理論) が採用されることが増えてきた。おそらく先駆的な例は, 埼玉県が実施する埼玉県学力・学習状況調査¹⁾であろう。IRT を採用すれば, 内容が異なるテストであっても受検者の学力が比較可能になり, 児童生徒の「学力の変化」や「学力の伸び」を把握できるようになる。学習指導や教育政策の効果を見たい教育研究者にとっても, IRT のこうした特性は魅力的である。国際的な研究の潮流を鑑みても, 同一個人の変化を追跡するパネルデータが流行しており, IRT を採用した学力調査が教育研究の前提になっていると言っても過言ではない (川口 2020)。

ただ, 単純に IRT を採用しただけでは, 教育研究に利用できる学力調査にはならない。とくに本稿で注意を促したいことは, IRT の能力推定法によって母集団の推定に偏りが生じるという点である。詳しくは2節で示すが, IRT の能力推定法としてよく利用される MLE (最尤推定値) や EAP (期待事後推定値) は, 基本的に受検者個人の能力を推定するための手法であり, 母集団の推定には不向きである。PISA や TIMSS, あるいは PIAAC といった大規模学力調査では, MLE や EAP に代わる能力推定法として PVs (Plausible Values : 推算値) と呼ばれる方法が採用されている (巖谷・篠原・篠原 2019; 廣田 2023)。これらの調査で利用されている PVs の概要については日本語の解説も存在するが, 教育測定に関する知識を有していることが前提になっており, 必ずしも学力調査を分析する研究者の関心に沿っているわけではない。そこで本稿は, PVs の概要と利用時の留意点について, 教育測定を専門としない者を念頭に解説を行う。さらに, 全国学力・学習状況調査に代表される日本の学力調査の特性も踏まえた上で, PVs の有効性と具体的な利用方法について論じる。

最初に, PVs の概要を確認しておこう。すでに述べたように, PVs は IRT の能力推定法の一つである。一般に IRT を用いた学力調査で受検者の能力を推定する際は, 最初に出題されたテスト項目の困難度 (≡ 個々の設問の難しさ) や識別力 (≡ 個々の設問が異なる能力を有する受検者を弁別できる程度) といった項目パラメータを推定する。その上で推定された項目パラメータと, 個々の項目への回答状況をもとに受検者の能力を推定することになる。能力推定法は, 大きく分類すると最尤推定法 (代表は MLE) と, 事前分布を置くベイズ推定法 (代表は EAP) に分類できる (加藤・山田・川端 2014)。

このうち MLE は, 母集団の推定という観点から見ると問題がある。なぜなら, 学力調査においてテストに出題できる設問の数には限りがあり, 推定された個々人の能力に測定誤差が存在するからである。単純に推定された受検者の能力を平均すると測定誤差が増幅され, 全体としてみると大きな偏りが生じる危険がある。とくに日本の小中学校を対象とした学力調査は「確認テスト」の意味合いが強く, テストの難易度が低くなりがちである (川口 2024)。この場合, テストが受検者の能力に比して簡単すぎるために, 満点やそれに近い成績を取る受検者が多数生じ, より大きな測定誤差が生じる可能性が高い。一方で EAP も, 事前分布 (通常は標準正規分布) に受検者の能力推定値が引き寄せられるため, 下位集団の平均値の差や能力の分

散を過小推定する危険がある。

こうした問題を改善するための発想が PVs である。数学的には PVs は、ベイズ推定を利用した、受検者の能力母数 (θ) の事後分布からの無作為標本である (文部科学省 総合教育政策局調査企画課学力調査室 2023)。今、受検者の項目反応パターンを x 、能力母数を θ 、能力を推定する尤度関数を $f(x|\theta)$ としよう ($f(x|\theta)$ には、たとえば 2 母数・ロジスティックモデルなどが想定される)。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平均 μ 、分散 σ^2 の正規分布 $g(\theta) \sim N(\mu, \sigma^2)$ を仮定する。このとき事後分布 $h(\theta|x)$ は次の式になり、ここからの無作為標本が PVs となる。

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

ただし、この形の PVs は学力調査の能力推定値として十分ではない。たとえば男女のように下位集団で受検者の能力分布が異なる事態を想定すると、男女を合わせた能力の分布は単峰ではなく双峰型に近いはずである。このような状況を考慮し、事前分布を $g(\theta) \sim N(u + \alpha x, \sigma^2)$ と修正する。ここで男子 ($x=0$)、女子 ($x=1$)、 α は男女の能力値の差であり、両者の分散は同一とする。この形の PVs を、とくに条件付けた PVs と言い、 x は条件付け変数 (Conditioning Variables) と呼ばれる (Wu 2005)。より一般に、複数の条件付け変数が想定できるなら、事前分布は $g(\theta) \sim N(u + \alpha x + \beta y + \gamma z + \dots, \sigma^2)$ という形に拡張できる。大規模学力調査における PVs は、こうした項目反応モデルと回帰分析 (「潜在回帰モデル」と呼ばれる) を組み合わせたものになっている。なお、条件付け変数には連続変数が想定されているが、ダミー変数を利用すれば、男女のようなカテゴリカル変数を投入することも可能である。

条件付けた PVs は、次のような利点を持つ (Wu 2005)。まず MLE や EAP で見られる分散の過大／過小推定を避けることができ、母集団の分散を適切に推定できる。しかもこの性質は、テストの項目数が少ない場合やテストの難易度が受検者の能力に比して低すぎる (あるいは高すぎる) 場合であっても成立する。日本の学力調査のようにテストの難易度がそれほど高くない調査を母集団の推定に使う場合、条件付けた PVs は有効な選択肢である。

本稿では PVs を利用する利点に加え、学力調査を分析する者が気になると思われる点についても解説を加える。具体的には、下位集団の平均値や相関係数の推定値といった一般的な話題だけでなく、条件付け変数の欠測、PVs を利用したマルチレベルモデル、PVs を独立変数として利用するといったテーマを扱う。最後にフリーの統計ソフトである R を利用して、学力調査から PVs を生成する方法についても紹介する。

2. PVs を利用しないとどのような偏りが生じるか

PVs を利用しないとどのような偏りが生じるのだろうか。この点を、学力と SES・ジェンダーの関連を想定したシミュレーションで確認しよう。よく知られているように、児童生徒の学力と SES (Socio-Economic Status: 社会経済的地位) には強い関連がある (松岡 2019)。さらに学力には男女差があり、SES と学力の関連も男女で異なると仮定しよう。具体的には、男子の学力の平均値は -0.2、女子は 0.2、学力と SES の相関は男子が 0.5、女子が 0.4 とする。SES の平均、学力と SES の分散は男女ともに 1 であり、サンプルサイズも男女ともに 2000 としよう。数理的には、男女それぞれについて下記のような平均 (μ) と共分散行列 (Σ) を持つ多変量正規分布を想定し、そこからそれぞれ 2000 の標本を取り出すことにする。

$$\mu_{\text{男子}} = \begin{pmatrix} -0.2 \\ 0 \end{pmatrix}, \Sigma_{\text{男子}} = \begin{pmatrix} 1.0 & 0.5 \\ 0.5 & 1.0 \end{pmatrix} \quad \mu_{\text{女子}} = \begin{pmatrix} 0.2 \\ 0 \end{pmatrix}, \Sigma_{\text{女子}} = \begin{pmatrix} 1.0 & 0.4 \\ 0.4 & 1.0 \end{pmatrix}$$

受検者の回答は、受検したテストから IRT の 2 母数・ロジスティックモデルに従って生成され则认为る。2 母数・ロジスティックモデルは全国学力・学習状況調査の経年変化分析調査で採用されているから、この仮定はそこまで無理のあるものではない。

受検するテストは、次のような特徴を持ったテストを想定する。まず、項目の困難度は -3 から +1 のあいだで等間隔に分布し、項目の識別力はすべて 0.9 とする。これは、全国学力・学習状況調査の悉皆調査を念頭に置いている (川口 2024)。すでに指摘したように、一般に日本の学力調査のテストの難易度は低くな

りがちである。そこで項目の困難度の最大値をやや低めに取っている。一方で識別力は比較的高いので、0.9 としても大きな問題はないと思われる。テストの設問数は、保護者の学歴や年収を調べる「保護者に対する調査」が行われた 2021 年度の全国学力・学習状況調査の悉皆調査が小学校・中学校ともに 15 題程度であることから、15 題に設定する。なお、過去の「保護者に対する調査」と同時に行われた 2013 年度・2017 年度の全国学力・学習状況調査は、国語・算数（数学）のそれぞれについて A 問題・B 問題が存在し、両者を合わせると 1 教科あたりの設問数が 30 題程度であった。そこで設問数の増減が推定に与える影響を検討するため、設問数を 30 題に設定した場合についてもシミュレーションを行う。

以上のような条件のもとで作成した仮想の解答データを、IRT で分析し受検者の能力を推定する。推定には R の TAM²⁾ を利用した。項目パラメータの推定法は周辺最尤推定法、能力推定法は MLE、EAP、条件付けない PVs (PV1 と呼称)、条件付けた PVs (PV2 と呼称) の 4 種類である。PVs は本来複数の値を抽出するが、今回は 1 つだけを抽出している。さらに参考のために、受検者の正答率を標準化した素点についても記載する。テストが 15 題と 30 題の 2 種類存在するため、個々の推定値は「素点_1」「素点_2」といった具合に添え字をつけて区別する。1 が 15 題、2 が 30 題のテストの推定値である。

シミュレーションは 1000 回を行い、以下の表に示す値は、その平均値である。なお、真値は多変量正規分布から抽出された標本の値である。カッコ内の数値は、真値と推定値の差から計算した RMSE (Root Mean Square Error: 二乗平均平方根誤差) であり、値が 0 に近いほど推定の誤差が小さいことを意味している。

2.1. 下位集団の平均値・標準偏差

最初に、下位集団の平均値と標準偏差について検討しよう。表 1 がシミュレーションの結果である。表 1 を見ると、素点・MLE・EAP・条件付けない PVs のいずれについても、真値と比べ 0.05 から 0.02 程度の偏りが生じていることがわかる。偏りは設問数が 15 題の方が大きく、とくに MLE_1 の RMSE は女子で 0.06 と大きな値を示している。これに対し、条件付けた PVs (PV2) は設問数によらずほぼ真値を復元しており、RMSE も小さい。

標準偏差については、素点・MLE・EAP のいずれも真値からの偏りが生じている。平均と同じく、その差は設問数が 15 題の時の方が大きい。PVs は条件付けの有無や設問数に関わらず真値の近い値を復元しており、RMSE も推定法の中でもっとも小さくなっている。

表 1. 男女の平均値・標準偏差の推定値

	真値	素点_1	素点_2	MLE_1	MLE_2	EAP_1	EAP_2	PV1_1	PV1_2	PV2_1	PV2_2
平均値 (男子)	-0.20	-0.17 (0.03)	-0.18 (0.02)	-0.16 (0.05)	-0.17 (0.04)	-0.16 (0.05)	-0.17 (0.03)	-0.16 (0.05)	-0.17 (0.03)	-0.20 (0.02)	-0.19 (0.02)
平均値 (女子)	0.20	0.17 (0.03)	0.18 (0.02)	0.26 (0.06)	0.24 (0.05)	0.16 (0.05)	0.17 (0.03)	0.16 (0.05)	0.17 (0.03)	0.19 (0.02)	0.19 (0.02)
標準偏差 (男子)	1.00	1.02 (0.02)	1.02 (0.02)	1.16 (0.16)	1.09 (0.09)	0.89 (0.11)	0.93 (0.07)	0.99 (0.02)	0.99 (0.02)	0.99 (0.02)	0.98 (0.02)
標準偏差 (女子)	1.00	0.95 (0.05)	0.94 (0.06)	1.17 (0.17)	1.11 (0.12)	0.86 (0.14)	0.91 (0.09)	0.98 (0.03)	0.98 (0.03)	0.98 (0.03)	0.98 (0.03)

○内は RMSE

2.2. 相関係数の偏り

続いて学力と SES の関連について検討する。表 2 は、学力と SES の相関係数の推定値を示したものである。これを見ると、素点・MLE・EAP・条件付けない PVs のいずれも真値から過小推定となっていることがわかる。とくに条件付けない PVs で差が大きい。また、設問数が 15 題の場合と 30 題の場合では、30 題の偏りが小さいこともわかる。条件付けた PVs は、設問数によらず真値を復元しており、RMSE も推定法の中でもっとも小さくなっている。

表 2. SES との相関係数の推定値

	真値	素点_1	素点_2	MLE_1	MLE_2	EAP_1	EAP_2	PV1_1	PV1_2	PV2_1	PV2_2
相関係数	0.44	0.39	0.41	0.39	0.41	0.39	0.42	0.35	0.39	0.44	0.44
		(0.05)	(0.03)	(0.05)	(0.03)	(0.05)	(0.03)	(0.09)	(0.05)	(0.01)	(0.01)

()内は RMSE

表 1・表 2 の結果から言えることは、母集団の推定には条件付けた PVs が有効であるということである。とくにテストの設問数に推定値が左右されないという点は重要である。他の推定法は、設問数の多寡によって推定値が変わるという欠点を抱えている。条件付けた PVs 以外の能力推定法を使っている場合は、実際には母集団の学力実態が変わっていなくても、出題される設問数が減少（あるいは増加）すると、推定値が変化する。全国学力・学習状況調査の設問数が 30 題前後から 15 題前後へ減少してきたことを鑑みれば、設問数に推定値が左右されるという特性は、学力の変化を捉えるという観点から問題がある。設問数に推定値が左右されないという 1 点だけを取り上げても、大規模な学力調査に条件付けた PVs を取り入れる重要性は明らかであろう。

3. PVs に関わる、その他の話題

3 節では、PVs を実際に利用する際に分析者が気になるだろう論点をいくつか取り上げる。具体的には、PVs の利用方法、条件付け変数に含まれる欠測、PVs を使ったマルチレベルモデル、PVs を独立変数として扱う場合の 4 点である。

3.1. PVs の利用方法

先ほど 2 節のシミュレーションで確認したように、PVs は事後分布から 1 つしか取り出さなかったとしても、母集団の特性を適切に推定することが可能である。もっとも、より精度を高めるため、通常は 5 ないし 10 程度を取り出した方が望ましい。ここで重要になるのが、複数の PVs の推定結果をどのように統合すればよいかという点である。原則として、PVs による推定結果 (θ) は、Rubin のルールと呼ばれる次の式で統合する (OECD 2009)。

$$\theta = \frac{1}{M} \sum_{i=1}^M \theta_i$$

ここで θ は統合された推定値、 M は PVs の数 (通常 5 ないし 10)、 θ_i は i 番目の PV を使った推定値である。たとえば 5 つの PV を使い、学力の平均値を計算したところ、その値が 1 番目の PV から順に、49, 50, 51, 50, 49 になったとしよう。このとき、最終的な推定値は、49.8 (計算方法は、 $(49+50+51+50+49) \div 5$) である。この式は、平均値だけでなく、回帰分析やマルチレベルモデルの係数、相関係数、 R^2 値の統合にも適用可能である。

推定値の分散は、以下の式で計算できる。

$$V = U + \left(1 + \frac{1}{M}\right) B_m, \quad B_m = \frac{1}{M-1} \sum_{i=1}^M (\theta_i - \theta)^2$$

ここで V は統合された推定値の分散、 U は個々の PV の推定値の分散の平均、 M は PVs の数、 θ_i は i 番目の PV を使った推定値、 θ は統合された推定値である。先ほどの例で、個々の PV の分散が順に、1.2, 1.3, 1.2, 1.1, 1.4 だったとしよう。このとき、 U は 1.24 (計算方法は、 $(1.2+1.3+1.2+1.1+1.4) \div 5$)、 B_m は 0.7 (計算方法は、 $((49-49.8)^2 + (50-49.8)^2 + (51-49.8)^2 + (50-49.8)^2 + (49-49.8)^2) \div 4$) なので、最終的な推定値の分散は、 $1.24 + (1+1/5) \times 0.7$ で、2.08 となる。

θ や V の計算式を見て、多重代入法を想起した人もいるかもしれない。それも当然で、PVs は多重代入法の考えを教育測定に応用したものである (巖・篠原・篠原 2019)。なお、多重代入法に詳しい人の中には、

Rubin のルールは相関係数や R^2 値のように分布が正規ではないものには適用できない（高橋・渡辺 2017）ことを知っている人もいるだろう。これは PVs も同じであり、PVs の推定結果を Rubin のルールで統合する際は、分布が正規かどうか確認する必要がある（Scharl & Zink 2022）。ただ、大規模学力調査の場合はサンプルサイズが大きいこともあって、相関係数や R^2 値も分布が正規であると仮定して問題はない。実際に PISA では、相関係数や R^2 値の推定結果を Rubin のルールで統合している（OECD 2009）。

PVs を利用する際に、重要なポイントはおそらく次の 2 点である。一点目は、PVs は受検者個人の能力推定値としては不適切という点である（巖・篠原・篠原 2019）。そもそも現実的な問題として、受検者の得点が複数あるようではテストとして問題がある。入試を想像するとわかりやすいと思うが、ある受検者の得点が 40 点かもしれないが 60 点かもしれない、はたまた 50 点かもしれないとなったら、合否判定に使うことができない。

二点目は、PVs の推定値を統合する際に、個々の PV を個人のレベルで平均してはならないというものである（文部科学省 総合教育政策局調査企画課学力調査室 2023）。1 節で述べたように、PVs は受検者の能力母数 (θ) の事後分布からの無作為標本である。PVs を個人レベルで平均すると、その値は EAP 推定に近い値になる。すなわち、PVs を個人レベルで平均すると、EAP を利用した場合と同じような偏りが生じてしまうのである。PVs を個人レベルで平均するよりも、1 つだけ PV を使って分析を行った方が、適切な推定値を得ることができる。

3.2. 欠測

ここまで言及しなかったが、実は PVs の推定では条件付け変数に欠測がないことが前提とされている（Weirich et al. 2014）。しかしこの仮定は、現実にはほとんどありえない。そのため PISA や TIMSS のような大規模学力調査では、条件付け変数の欠測を示すダミー変数を作成し、条件付け変数に加えることが一般に行われている。具体的な手順は、巖（巖・篠原・篠原 2019）などを参照してほしい。その意味では、PISA や TIMSS で提供されている PVs は、条件付け変数の欠測は考慮済みであり、ユーザー側がそこまで欠測を意識する必要はないとも言える。

ただダミー変数を利用した方法は、条件付け変数の欠測の割合が大きい場合には、推定に偏りが生じる可能性がある。条件付け変数の欠測にどう対処するかという点について、現時点でダミー変数を使う以外の一般的な合意は存在しないようだが、いくつか対処法に関する研究が進んでいる。中でも注目されるのが、条件付け変数の欠測を多重代入法で処理し、作成した複数のデータセットのそれぞれから PVs を生成するというものである（Weirich et al. 2014）。この手法を使うと、条件付け変数の欠測を多重代入法で補完しつつ PVs を生成することが可能になる。もっとも求められる計算量はかなり多く、たとえば多重代入法で 5 つのデータセットを生成し、それぞれのデータセットについて 5 つの PVs を生成したとすると、 5×5 で 25 の PVs を統合することになる。大規模学力調査はサンプルサイズが大きいことを考えると、あまり多数のデータを生成することは難しいだろう。この手法は、ドイツの学力調査に実装されており、R を使った自動化も行われている（Scharl & Zink 2022）。

3.3. マルチレベルモデル

教育研究では、学校ごとに児童生徒の学力の平均や分散が大きく異なる（≡データに階層構造がある）ことを考慮し、マルチレベルモデルと呼ばれる分析手法が利用されることがある。PVs をマルチレベルモデルで利用するにはどうすればよいのだろうか。

まず確認しておくことは、PVs の推定の際に利用されている回帰モデルはマルチレベルモデルではないから、何の工夫もなしに PVs をマルチレベルモデルの従属変数として利用することはできないという点である。条件付けない PVs を従属変数としてマルチレベルモデルを実行した場合、推定値には偏りが生じる。端的には、個々の学校の平均点に差があることが考慮されていないため、学校間分散が過小推定されてしまう（Monseur & Adams 2009）。

この問題を回避するため、PISA や TIMSS といった大規模学力調査では、それぞれ次のような対応が取られている（Zheng 2023）。まず PISA では、条件付け変数には学校 ID がダミー変数として含まれている。個々の学校ごとに平均点が異なるということを、学校 ID で表現しているわけである。他方 TIMSS では、各学校／学級の児童生徒の学力推定値（EAP）の平均が、条件付け変数として含まれている（EAP の算出

方法については後述する)。いずれの方法でも学校間の平均値の差を考慮することが可能になるが、前者は学校の数だけ条件付け変数にダミー変数を投入する必要があるため、学校の数次第では回帰モデルの独立変数が膨大になるという欠点がある。

なお、これらの手法は、学校ごとの平均点が異なるという情報を潜在回帰モデルに組み込んだだけである。そのため Zheng (2023) によればマルチレベルモデルのランダム切片モデルには対応できるが、係数が学校ごとに異なるモデル (=ランダム切片・係数モデル) について適切に推定できる保証はない。PVs の推定に利用する回帰モデルに階層構造を組み込むことができれば、ランダム切片・係数モデルにも対応した PVs を生成可能かもしれない。ただ現時点では、階層構造を組み込んで PVs を生成できるソフトウェアは存在していないようである。

3.4. PVs を独立変数として扱う

大規模学力調査における PVs は、基本的に従属変数として利用することが想定されている。ただ、研究テーマによっては学力を独立変数として利用したいというケースもあるだろう。たとえば相澤・池田 (2022) の研究では、「いじめ反対意識」を従属変数、数学的リテラシーの PV1 を独立変数として重回帰分析が行われている。こうした PVs を独立変数として扱う場合の留意点については、管見のかぎり検討は行われていないようである。そこでシミュレーションを使い、PVs を独立変数とした場合に適切な推定値が得られるかどうか検討してみよう。

以下では、2 値データが従属変数、PVs が独立変数の場合の二項ロジスティック回帰分析を例にシミュレーションを行う。真の回帰係数は 0.2 とし、テスト（困難度は -3 から 3 で等間隔に分布、識別力はすべて 0.9、設問数は 15 題）を受検した場合に得られる解答データに、IRT の 2 母数・ロジスティックモデルを適用して得られた 5 つの PVs の最初の 1 つ (PV1) を独立変数として、1000 回のシミュレーションを繰り返した。PVs は、通常の大規模学力調査を想定し、従属変数を条件付け変数として生成したものである。参考のために、PV1 以外に素点、MLE、EAP、および 5 つの PVs を個体レベルで平均したものを独立変数とした場合の推定値も計算した。分析結果は、表 3 の通りである。

表 3. 回帰係数の推定値

	真値	素点	MLE	EAP	PV1	PVs を平均
回帰係数	0.20	0.17	0.15	0.20	0.20	0.25
		(0.03)	(0.05)	(0.02)	(0.03)	(0.06)

() 内は RMSE

表 3 を見ると、条件付けた PVs の一つを独立変数として投入すること (PV1) が適切な推定に繋がると考えられる。従属変数として PVs を扱う場合と同じく、個々の PV を個人レベルで平均することは偏りに繋がるので避けた方がよい。興味深いことに、EAP 推定で算出した能力値は、独立変数として使うのであれば適切に母集団の値を復元することが可能なようである。これに対して、素点や MLE 推定は過小推定に繋がるため、避けた方がよい。RMSE を見ても、EAP か PVs を 1 つだけ利用する方法が適切と言えるだろう。

4. R を利用した推算値の算出

図 1. 仮想の解答データの生成

```
# パラメータの定義
size <- 4000 # サンプルサイズ
item_n <- 15 # 項目数

b <- seq(-3, 1, length.out = item_n) # 項目困難度
a <- 0.9 # 項目識別度

# theta×SES
sig1 <- matrix(c(1.0, 0.4, 0.4, 1.0), nrow = 2) # 女子
sig2 <- matrix(c(1.0, 0.5, 0.5, 1.0), nrow = 2) # 男子

sim1 <- mvtnorm::rmvnorm(n = size / 2, mean = c(0.2, 0), sigma = sig1)
sim2 <- mvtnorm::rmvnorm(n = size / 2, mean = c(-.2, 0), sigma = sig2)

sim <- rbind(sim1, sim2)

theta <- sim[, 1]
ses <- sim[, 2]
sex <- c(rep(1, size / 2), rep(-1, size / 2)) # 性別

# 反応パターンを生成
resp <- mirt::simdata(
  a = rep(1.7 * a, item_n),
  d = -b * 1.7 * a,
  Theta = theta,
  itemtype = "2PL"
)
```

PISA や TIMSS といった大規模学力調査では、公開されたデータセット内に PVs が含まれていることが一般的である。これに対して日本の学力調査では、条件付けた PVs がデータセットに含まれていることは稀なので、研究者が自身で PVs を計算せざるをえないといった場面もあるだろう。4 節では、フリーの統計ソフトである R を使い、PVs を算出する方法を紹介する。具体的には、IRT 分析用の package としてよく知られている MIRT³⁾ と TAM を利用する。なお、R の使い方は解説しないので、加藤・山田・川端 (2014) といった入門書を参照してほしい。

最初に、MIRT を利用して仮想の解答データを作成する (図 1)。パラメータの設定は、2 節で利用した男女と SES の関連に関するシミュレーションを流用する。反応パターンは、“resp” 内に格納される。

4.1. MIRT を利用した場合

図 2 は、MIRT を利用した PVs の推定法を示したものである。条件付けを行っていない場合 (mod1, pvs1) と行った場合 (mod2, pvs2) のそれぞれを示す。抽出する PV の数は fscores 関数の引数 (plausible.draws) を指定することで調整できる。図 2 の場合、5 つの PV を抽出している。条件付けに利用する変数は、引数 (covdata) にデータフレームを指定し、さらに引数 (formula) で回帰式を指定することで設定できる。ここでは、単回帰分析を想定しているが、“formula = ~SEX * SES” のように指定すれば、性別と SES の交互作用を含んだ回帰モデルを適用することも可能である。

図 2. MIRT による PVs の推定

```
# 条件付け無
mod1 <- mirt::mirt(resp, model = 1, itemtype = "2PL")
pvs1 <- mirt::fscores(mod1, plausible.draws = 5)

# 性別で条件付け
mod2 <- mirt::mirt(resp,
  model = 1,
  itemtype = "2PL",
  covdata = data.frame(SES = ses, SEX = sex),
  formula = ~SEX
)
pvs2 <- mirt::fscores(mod2, plausible.draws = 5)
```

4.2. TAM を利用した場合

図 3 は、TAM を利用した推定法を示したものである。条件付けを行っていない場合（mod3, pvs3）と行った場合（mod4, pvs4）のそれぞれを示す。TAM を使って条件付けを行う場合、MIRT と同様に引数（dataY, formulaY）でデータフレームと回帰式を指定してもよいが、引数（Y）に条件付け変数の行列を指定するだけでもかまわない（コメントアウトされている側の mod4）。実際の大規模学力調査の分析では条件付け変数が百を超え、回帰式を指定するのが面倒なことも珍しくないので、この仕様は便利である。

図 3. TAM による PVs の推定

```
# 条件付け無
mod3 <- TAM::tam.mml.2pl(resp)
tampv1 <- TAM::tam.pv(mod3, nplausible = 5)
pvs3 <- TAM::tampv2datalist(tampv1)

# 性別で条件付け
mod4 <- TAM::tam.mml.2pl(resp,
  formulaY = ~ SEX, dataY = data.frame(SEX = sex, SES = ses)
)
# 引数 Y に条件付け変数の行列を指定してもよい
# mod4 <- TAM::tam.mml.2pl(resp, Y = sex)

tampv2 <- TAM::tam.pv(mod4, nplausible = 5)
pvs4 <- TAM::tampv2datalist(tampv2)
```

TAM の利点は、いったん条件付けを行わずに項目パラメータを推定し、その推定値を利用して条件付けを行うという 2 段階の推定が可能な点である（図 4）。大規模学力調査では複数の能力間の相関を考慮し、多次元項目反応理論が利用されることがある。このような複雑なモデルを適用すると同時に条件付けを行うと、推定に時間がかかるか、あるいは収束しない場合がある。この場合、最初に条件付けをせずにパラメータを推定し（mod3, likeli）、続いて条件付けを行う方法（mod5）が有効である。なお、TIMSS の条件付けに利用されている EAP 推定値は、第 1 段階の推定（mod3）の情報をもとに推定した EAP であり、これを学校・学級単位で平均したものが第 2 段階の推定（mod5）で条件付け変数として利用されている。

図 4. TAM による PVs の推定 (2 段階推定)

```
# 2 段階で条件付け
likeli <- CDM::IRT.likelihood(mod3)
mod5 <- TAM::tam.latreg(likeli,
  formulaY = ~ SES + SEX,
  dataY = data.frame(SEX = sex, SES = ses)
)
# mod4 と同じく、引数 Y に条件付け変数の行列を指定してもよい
# mod5 <- TAM::tam.latreg(likeli, Y = cbind(sex, sex))

tampv3 <- TAM::tam.pv(mod5, nplausible = 5, normal.approx = TRUE)
pvs5 <- TAM::tampv2datalist(tampv3)
```

5. まとめ

本稿では、大規模学力調査における PVs の概要とその利点についてまとめてきた。最後に、大規模学力調査の分析者が知っておくべきポイントを 6 点にまとめておこう。

第一に、条件付けた PVs を利用することで下位集団の平均や分散、あるいは学力と他の変数の相関係数を適切に推定することができる。それ以外の手法 (MLE や EAP、あるいは条件付けを行っていない PVs) では推定値に偏りが生じる。この偏りはテストの設問数の多寡にも左右される。条件付けた PVs は、テストの設問数の多寡によらず母集団を適切に推定するため、大規模な学力調査では条件付けた PVs を利用した方がよい。

第二に、PVs を個体レベルで平均してはならない。個体レベルで平均すると、その値は EAP に近くなり、EAP を離礁した場合と似たような偏りが生じることになる。

第三に、条件付け変数の欠測は、通常はダミー変数で処理されているため、そこまで気にする必要はない。ただ条件付け変数に多数の欠測がある場合、多重代入法で補完することも可能である。この場合、補完した個々のデータセットごとに PVs を生成し、その結果を統合する手法が存在する。

第四に、マルチレベルモデルを利用する場合、学校 ID か学校ごとの学力推定値で条件付けを行うことで、ランダム切片モデルに対応することができる。ただし、ランダム切片・係数モデルについては不明瞭な点がある。

第五に、PVs を独立変数として扱いたい場合は、従属変数の場合と同じく、PV ごとに推定を行うことが望ましいと考えられる。個体レベルで PVs を平均した場合、その推定値には偏りが生じる。

第六に、学力調査のデータセットに条件付けた PVs が含まれていない場合、R の MIRT や TAM を利用して自身で PVs を生成することが可能である。とくに TAM は、PISA や TIMSS でも利用されている 2 段階の推定法が利用できるため、大規模学力調査の分析に最適であると考えられる。

冒頭でも述べたように、日本の学力調査にも IRT が採用されることが増えている。今後は、こうした調査データを使った学力研究が進むであろう。その際、より適切な推定を行うために本稿の知見が役立てば幸いである。

【注】

- 1) <https://www.pref.saitama.lg.jp/f2214/gakutyoku/20150605.html>
- 2) TAM (<https://cran.r-project.org/web/packages/TAM/index.html>)
- 3) MIRT (<https://cran.r-project.org/web/packages/mirt/index.html>)

【参考文献】

相澤真一・池田大輝, 2022, 「別学と共学の違いから見る男女のいじめに対する意識の計量分析—PISA2018 データを用いた日韓英豪四ヶ国比較教育学研究」『教育学研究』89, 670-682.

- 廣田英樹, 2023, 「PIAAC の Plausible Values の理解のために —Plausible Values を用いる理由とその算出方法」『国立教育政策研究所紀要』152, 71-87.
- 褓岩晶・篠原真子・篠原康正, 2019, 『PISA 調査の解剖—能力評価・調査のモデル』東信堂。
- 川口俊明, 2020, 『全国学力テストはなぜ失敗したのか—学力調査を科学する』岩波書店。
- 川口俊明, 2024, 「全国学力・学習状況調査（保護者に対する調査・経年変化分析調査）における多次元項目反応モデルと推算値法の有効性の検証」『日本テスト学会誌』20, 73-89.
- 加藤健太郎・山田剛史・川端一光, 2014, 『R による項目反応理論』オーム社。
- 松岡亮二, 2019, 『教育格差—階層・地域・学歴』ちくま新書。
- 文部科学省 総合教育政策局調査企画課学力調査室, 2023, 『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（改訂版）』文部科学省。
- Monseur, C., & Adams, R., 2009, Plausible Values: How to Deal with Their Limitations, *Journal of applied measurement*, 10(3).
- OECD, 2009, *PISA Data Analysis Manual: SPSS Second Edition*, OECD Publishing.
- Scharl, A., & Zink, E., 2022, NEPSscaling: Plausible Value Estimation for Competence Tests Administered in the German National Educational Panel Study, *Large-scale Assessments in Education*, 10(1), 28.
- 高橋将宜・渡辺美智子, 2017, 『欠測データ処理—R による単一代入法と多重代入法』共立出版。
- Weirich, S., Haag, N., Hecht, M., Böhme, K., Siegle, T., & Lüdtke, O., 2014, Nested Multiple Imputation in Large-scale Assessments, *Large-scale Assessments in Education*, 2, 1-18.
- Wu, M., 2005, “The role of plausible values in large-scale surveys,” *Studies in Educational Evaluation*, 31(2-3), 114-128.
- Zheng, X., 2023, On Generating Plausible Values for Multilevel Modelling with Large-scale-assessment Data, *British Journal of Mathematical and Statistical Psychology*, 77(1), 212-236.