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 ―

The Reason for and Method of Producing “Plausible Values” in PIAAC

廣田　英樹*
HIROTA Hideki

Abstract

Many international proficiency assessments like PIAAC, PISA and so on use a number of “Plausible Values”（PVs）instead of a single score of individual proficiency. Although a single score can be estimated by using Item Response Theory（IRT）, why do they use PVs and how are those PVs produced? In Japan, we can hardly find comprehensive and easy-to-understand commentaries on PVs. I suspect it is one of the reasons why there are not many secondary analyses which use the date of “Public Use File” of PIAAC in Japan. Therefore, I tried to make brief commentaries. In this paper, I start my explanation with IRT for beginners, then I proceed to the uncertainty of the estimation by IRT. Next, I discuss the integration of the IRT model and a population model that uses the information obtained from the background questionnaire. Finally, I explain the method of multiple imputation and the procedure of producing 10 PVs. I also add a brief guide on how we use those 10 PVs. The new PIAAC assessment was held in 2022 and 2023 and its data will be available soon. I hope these brief commentaries will be of some help to researchers who are interested in the new PIAAC.

* 生涯学習政策研究部　総括研究官

１．はじめに

PIAAC（国際成人力調査）は、PISA などと同様に OECD の主導で行われる国際的な調査であり、2011 年に第１回の調査が実施されて以来 11 年ぶりに、2022 年から 2023 年にかけて第２回の調査が実施された。OECD によって調査結果が報告書として取りまとめられるのは 2024 年になる予定だが、各国の調査を通じて収集されたミクロデータ（個票データ）も、OECD による精査・加工を経て、第１回と同様に “Public Use File” として公開されると思われるので、日本の大学や研究機関においても幅広く研究のために活用されることが望まれる。

一方で、第１回の PIAAC については、公開されたデータを活用した研究事例が必ずしも多くないように感じられる。その理由の一つに、PIAAC のデータの取扱いが難しいと思われていることがあるのかもしれない。中でも大きな問題は、PIAAC の基本的な目的が、各国の成人の認知スキルを測定することにある中で、OECD によって公表されているデータには、読解力、数的思考力、IT を活用した問題解決能力の３分野の回答者のスコアはなく、それに代わるものとして、10 個の “Plausible Values”（以後 “PVs” と呼ぶ。）が収載されていることではないだろうか[1]。図１は、OECD が公開している日本のデータファイルから作成した。一番上にある列番号は筆者が付したものであるが、図中の一つの行が一人の回答者に対応しており、年齢、性別から始まって、教育、仕事、家族など、背景調査の設問から得られた回答や、３分野のアセスメントでの解答に関するデータなどが記載された後に、1159 列目から 10 個の PVLIT（読解力の PVs）が記載されている。

図１　PIAAC の公開ミクロデータのファイルに収載された PV

1	2	3	4	5		1159	1160	1161	1162	1163	1164	1165	1166	1167	1168
CNTRYID	CNTRYID_	SEQID	AGE_R	GENDER_R		PVLIT1	PVLIT2	PVLIT3	PVLIT4	PVLIT5	PVLIT6	PVLIT7	PVLIT8	PVLIT9	PVLIT10
392	392	1	45	2	～	290.8536	275.3755	301.8967	300.1982	295.5246	317.5466	296.7449	304.7697	304.8909	299.4686
392	392	2	48	1	～	277.1685	278.3115	284.6797	291.2799	274.0447	297.0942	260.007	289.7718	297.4921	279.8756
392	392	3	47	1	～	302.4788	304.202	303.3461	303.8161	311.3669	312.94	287.3272	278.205	306.6005	292.7362
392	392	4	58	1	～	304.5744	282.5685	305.774	311.4683	300.1175	302.5375	299.5183	299.1208	338.1857	288.8517
392	392	5	29	2	～	335.5443	367.095	308.1292	348.1189	310.7282	337.2817	315.2988	312.6049	336.1294	329.9094
392	392	6	35	2	～	339.3943	371.0809	361.9547	367.3884	352.9698	393.6736	346.5537	333.9634	377.9273	317.3439
392	392	7	60	1	～	303.7377	314.3853	298.8021	306.748	288.9831	279.601	292.4033	298.5717	317.4293	274.7586
392	392	8	48	2	～	338.4039	304.2485	332.7174	301.0753	335.4304	358.4662	304.8877	329.7203	296.6114	308.0711
392	392	9	56	1	～	221.493	238.3714	219.3401	198.1977	231.1964	217.587	235.1607	226.4605	223.974	233.3929
392	392	10	24	1	～	322.1816	303.8807	301.8995	328.8217	321.2318	310.6124	322.0015	307.1173	320.0848	313.1191

この PVs とは何だろうか。10 個の PVs はなぜ、そしてどうやって算出されるのか。またそれらをどのように分析に利用したらよいのか。これらの疑問に答えるのが本稿の目的だが、そのためには、項目反応理論をはじめとした、関係する理論や手法に一定程度触れることが必要になる。一方で、PIAAC は成人の認知スキルの測定を目的としていることから、例えば教育社会学や高等教育研究、あるいは労働経済学などの分野でも関心の対象となると考えられるが、こうした分野では、項目反応理論等について未だなじみが薄い研究者も多いと思われる。このため本稿では、PVs をデータ分析に活用する上で必要な理解の促進ということに的を絞って、できるだけ直感的な理解を重視しながら、最小限の範囲で理論的なことについても説明することとする。

1　PIAAC のミクロデータを用いて分析を行う場合は、ジャックナイフ法を用いて標準誤差を計算する必要があり、このことも一つのハードルとなっている可能性がある。これについては脚注 34 を参照。

２．項目反応理論に基づく推計に伴う不確かさ

日本の教育現場で通常行われているテストでは、予め問題ごとに配点が決められており、それに基づいて採点されるのが一般的である。しかし PIAAC や PISA のような国際調査では、問題ごとの配点は存在せず、「項目反応理論」(Item Response Theory）に基づき、回答者の解答結果に応じてその「特性値」[2]（本稿では特性値を θ で表す。）を推計するという方法が取られている。特性値の推計に際しては、各項目[3]の「困難度」β や「識別力」α[4]（これらは項目パラメーター[5]と呼ばれる）が推計されるが、困難度が高い項目に正答することが直接的に高い特性値の推計につながるわけではない。

項目反応理論の基本的な考え方を大まかに説明すれば、特性値が高い回答者は高い確率で多くの項目に正答すると考えられるので、その関係を逆にたどって、解答の正誤のパターンすなわち「項目反応」から､回答者の特性値を推計しようとするものと言える。関連して､一定の条件下で､異なる項目群が出題された回答者集団間で推定した特性値を、相互に比較可能にする手法も用いられる。

このような推計は信頼できる数学的手法で裏付けられている。しかしまた、特性値は項目反応を通して「推計」することしかできない潜在的な構成概念であり、身長や体重のように実測することはできない。そして推計には一定の誤差あるいは不確かさがつきものであり、それを適切に考慮に入れないと、例えば平均値の信頼区間を正しく計算できないし、回帰係数の有意性も正しく評価できないことになる。最初に一つの結論を言えば、PIAAC の公表データに、各回答者について推計したただ 1 個の特性値ではなく、異なる 10 個の PVs が収載されているのはこのためである。10 個の PVs は、そのばらつきを通して推計値の「不確かさ」に関する情報を補っていると考えられる[6]。

３．PVs の算出に至る具体的なプロセス

しかしながら、PVs が特性値の推計の不確かさを表していると述べただけではもちろん十分な説明とは言えない。なぜ推計に不確かさが伴うのか、それをどうやって補っているのか、結局 PVs はどうやって算出されるのかを知ることが重要である。このため、PVs の算出に関わる主要な事項について、以下の順でこれから説明を行う。数式も用いているが、回帰分析に関する基本的な知識を有する方であれば概ね理解できると思われる内容となるように適宜解説を加えてい

2 「英語では “trait”、もしくは直接観察できない潜在的な概念として “latent trait”（潜在特性）と呼ばれる。PIAAC が測定している trait は「認知スキル」(cognitive skills）の「習熟度」(proficiency）である。しかし豊田（2012, p2）は、「学力試験で測られる特性は一般的に能力と呼ぶことができるが、性格検査や臨床検査で測っている対象は能力ではない。また能力が高いとか、能力が低いという表現は耳障りである。」として「特性値」という語を用いており、これを踏まえて本稿でも「特性値」という言葉を用いる。

3 「項目」とは回答者に出題される設問のことであり、“Item Response Theory” の “Item” である。

4 「困難度」を本稿では β で表記し､「識別力」を本稿では α で表記する。困難度は文字通りの意味で解してもよいが、モデルにおける具体的な機能は次頁の３（１）①２パラを参照。識別力の機能については同４パラを参照。

5 パラメーターを母数と訳する用例もあるが、本稿では単にパラメーターと表記する。パラメーター（parameter）という語は、元来ギリシャ語の $\pi\alpha\rho\alpha$（そばで）と $\mu\varepsilon\tau\rho o\nu$（尺度）という語からつくられた造語であるとされる。

6 OECD（2016, p2）も、「PVs の方法論は、多重代入された習熟度（PVs）を用いることで、個人のレベルにおける誤差（若しくは不確かさ）を、こうした不確かさをゼロと推定するよりも正しく説明する。」としている。

る。

（１）項目反応理論に基づく特性値の推計方法

（２）幅広い層を対象に幅広く特性値を測定することに伴う制約

（３）項目反応に基づく推計方法と回帰式による推計方法との統合

（４）多重代入法の基本的な考え方と 10 個の PVs の算出手順

なお、本稿では特に OECD（2016）と Maehler et al.（2020）を多く参照している。OECD（2016）は Chapter ごとに頁が付されているが、参照したのは Chapter17 が殆どなので、Chapter17 での参照については頁数だけを記載し、他の Chapter を参照したときだけ Chapter 番号を記載した。また Maehler et al. も Chapter 3 しか参照しておらずその著者は Khorramdel et al.[7] なので、本文では後者の名前で記載した。数式中のアルファベットや添え字の用い方は、基本的に OECD（2016）の用法に準拠したが、文献によって用法が異なっている場合は、引用元の用法をそのまま用いて、注記で相違を示した。

（1）項目反応理論に基づく特性値の推計方法

① 項目反応理論のモデル

PIAAC は項目反応理論のモデルとして、２パラメーターロジスティックモデル（2-parameter logistic model。以後 “2PL” と呼ぶ。）と一般化部分採点モデル（generalized partial credit model）を用いている（OECD, 2016）。前者は解答が正答か誤答の何れかである項目を対象としており、後者は解答の評価が多段階で行われる項目を対象としている。読解力と数的思考力の項目が前者に当たり、ICT を活用した問題解決能力の項目には前者と後者の双方があるが、本稿では専ら 2PL を用いて説明を行う。

2PL のモデルは以下の１－１式[8] に基づいている。$p_i(\theta_j)$は回答者 j が項目 i に正答する確率である。回答者の特性値 θ と項目の困難度 β は同じ尺度を共有し、「引き算」が可能であるとされている。１－１式を変形した１－２式の右辺に注目すると、$\theta_j - \beta_i$ が大きくなる、即ち回答者 j の特性値が項目 i の困難度よりも大きくなると、$p_i(\theta_j)$が 1 に近づいていき、正答確率が高まるようにモデルが仕組まれていることが分かる。

$$p_i(\theta_j) = \frac{exp\left(D\alpha_i(\theta_j-\beta_i)\right)}{1+exp\left(D\alpha_i(\theta_j-\beta_i)\right)} = \frac{1}{1+exp\left(-D\alpha_i(\theta_j-\beta_i)\right)} \qquad 1-1\text{式}$$

$\exp(x) = exp^x$ であり、$exp(-x) = \frac{1}{exp^x}$ なので、１－１式の右辺は以下に変形される。

$$p_i(\theta_j) = \frac{1}{1+exp\left(-D\alpha_i(\theta_j-\beta_i)\right)} = \frac{1}{1+\frac{1}{exp\left(D\alpha_i(\theta_j-\beta_i)\right)}} \qquad 1-2\text{式}$$

また、θ_j と β_i が同じ値のときに正答確率は 0.5（五分五分）になり、θ_j が β_i よりも小さく

7 Khorramdel et al.（2020）の著者は一人を除いて OECD（2016）の Chapter 17 の著者と同一であり、前者の内容は後者のそれを補足修正したものになっている。

8 exp は exponential の略であり自然対数の底（e = 2.718…）を意味する。また中央の項と右端の項とが等しくなるのは $\frac{a}{1+a} = \frac{1}{1+\frac{1}{a}}$ だからである。D は定数であり、$D = 1.7$ である

なると正答確率はそれよりも低下していく。なお、θ が無限大となるとき正答確率は 1 に限りなく近づき、逆に θ が負の無限大となるとき正答確率が 0 に限りなく近づくが、これは数式上でのことであり、現実的な話ではない。

１－１式あるいは１－２式をグラフ化したのが図２に描かれた曲線である。横軸は特性値 θ の大きさであり、縦軸は正答確率である。ここでは $\beta = 0$ に固定しているので、$\theta = 0$ のときに正答確率が0.5になっている。曲線が３本描かれているのは識別力 α の値の違いによる。α は“slope parameter”とも呼ばれ、α の値が大きくなるにつれて、$\theta = \beta$ のところで特性値 θ のわずかな変化が正答確率に大きな変化を与えるようになるが、今後の説明では簡単のために $\alpha = 1$ に固定する。

② 同時確率分布による正誤パターンの表現

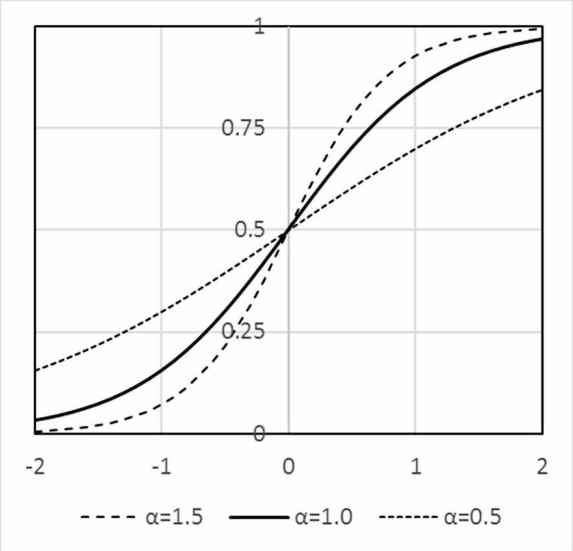
ここまでの説明で、回答者 j の特性値 θ_j，項目 i の困難度 β_i, 識別力 α_i から正答確率 $p_i(\theta_j)$ が推計されることを見た。しかし実際には、特性値 θ_j, 困難度 β_i, 識別力 α_i の何れ(いず)も既知ではない。

このことについて２で、「関係を逆にたどって、解答の正誤のパターンすなわち『項目反応』から、回答者の特性値を推計」すると述べたが、そのために、まず解答の正誤パターンが数学的にどのように表現されるのかを見てみよう。

回答者 j の項目 i への反応を x_{ij} とし、正答の場合は $x_{ij} = 1$ 誤答の場合は $x_{ij} = 0$ と記すことにする。引き続き $p_i(\theta_j)$ は回答者 j が項目 i に正答する確率を表すが、誤答する確率を $q_i(\theta_j)$ で表すことにする。$q_i(\theta_j) = 1 - p_i(\theta_j)$ である。

図２ 特性値の大きさと正答確率との関係

（横軸：特性値，縦軸：正答確率，困難度 β =0 の場合）



次に複数の項目反応を数式で表す方法を見る。項目反応理論では、「局所独立の仮定」[9] を前提として、回答者の解答の正誤のパターンを同時確率分布として表す手法が採られる。具体的には、回答者 j の n 個の項目への反応：$x_j = [x_{1j}, x_{2j}, \cdots x_{nj}]$ となる確率を、$x_{1j}, x_{2j}, \cdots x_{nj}$ という事象が同時に生起する確率として定義し、これを個々の反応の確率分布の積として以下の２式で表すのである[10]。式の左辺の項の$\prod$の記号は、$\prod$の右側の各要素をすべて掛け合わせることを意味しており、これを展開すると右辺の項となる。

$$\prod_{i=1}^{n} p_i(\theta_j)^{x_{ij}} q_i(\theta_j)^{1-x_{ij}} = p_1(\theta_j)^{x_{1j}} q_1(\theta_j)^{1-x_{1j}} \times p_2(\theta_j)^{x_{2j}} q_2(\theta_j)^{1-x_{2j}} \cdots \times p_n(\theta_j)^{x_{nj}} q_n(\theta_j)^{1-x_{nj}} \quad \text{２式}$$

9 局所独立の仮定について豊田（2012, p45）は「θ_i（筆者注：OECD の用法では添え字 j が回答者を示し、添え字 i が項目を示しているが、豊田の用法では添え字 i が回答者を示している。）が所与である場合には、項目反応は互いに独立である仮定」としている。局所独立の成立を妨げる要因として登藤（2012, p.82）は３つのタイプがあるとしている。このうち２つは項目間の内容的なつながりに関係するものだが、３つ目は、正答確率が特性値 θ 以外の想定外の特性値 θ' の影響を受ける場合であるとしている。このことに関連して塗師（1989, pp.137-139）は、「その尺度が測定している特性は確かに一つであるかという尺度の一次元性」を取り上げて、「一次元性とは、Lumsden（1961）によれば、すべての項目が同じ一つの特性をしかもそれだけを測定していること」であると述べている。

10 正答の時には $x_{ij} = 1$、誤答の時には $x_{ij} = 0$ であり、$\alpha^1 = \alpha$，$\alpha^0 = 1$ なので、$p_i(\theta_j)^{x_{ij}} q_i(\theta_j)^{1-x_{ij}}$ の各項は、正答の場合は $p_i(\theta_j)$、誤答の場合は $q_i(\theta_j)$ で表される。

③ 最尤推定による特性値の推計

最後に、解答の正誤のパターンから、その「原因」となる特性値をどのように推計するのかを見てみよう。回答者 j の特性値がある値 θ_j を取るときに、項目 i に正答する確率を条件付確率と呼び、$p_i(\theta_j)=P(x_{ij}=1 \mid \theta_j)$ と表す。括弧の中の縦棒の右側は先に生起した事象を意味し、左側はその後に生起した事象を意味する。この関係を逆転して、回答者 j が項目 i に正答したという結果から、回答者 j の特性値が θ_j であったと考えることの妥当性（もっともらしさ）を尤度（likelihood）と呼び、これを $L(\theta_j \mid x_{ij}=1)$ 表す。

x_{ij} は既に起こったことなので固定されるが、θ_j の値を変化させることで尤度も変化するので、$L(\theta_j \mid x_{ij}=1)$ は尤度関数と呼ばれる。関数の式は条件付確率と同じである。尤度を最大にする θ_j の値を求めることを最尤推定（maximum likelihood estimation）と呼び、これによって得られた値（最尤推定値）が回答者 j の特性値 θ_j の推計値となる。

θ_j を動かすことで尤度が変化することを１－２式の右側の項を用いて見てみよう。解答が正答のときは尤度関数が $p_i(\theta_j)=1/(1+1/exp(D\alpha_i(\theta_j-\beta_i)))$ となるので、θ_j を β_i より大きくしていくと尤度が高まる（$p_i(\theta_j)$ が大きくなる）が、誤答のときは尤度関数が $q_i(\theta_j)=1-1/(1+1/\exp(D\alpha_i(\theta_j-\beta_i)))$ になるので、逆に θ_j を小さくしていくと尤度が高まる（$q_i(\theta_j)$ が大きくなる）ことが分かる。同様なことはより複雑な２式[11] でも観察できる。

実際に、困難度 $\beta=0$ の５項目が出題された場合の特性値を最尤推定した結果が図３である。正答率が 0.5 となる 2.5 項目が正答のときに $\theta_j=0$ と推計されるが、現実には 2.5 項目を正答する

図３　困難度 $\beta=0$ の５項目が出題された場合の正答数別の特性値の最尤推定値

（横軸が正答数、縦軸が特性値）

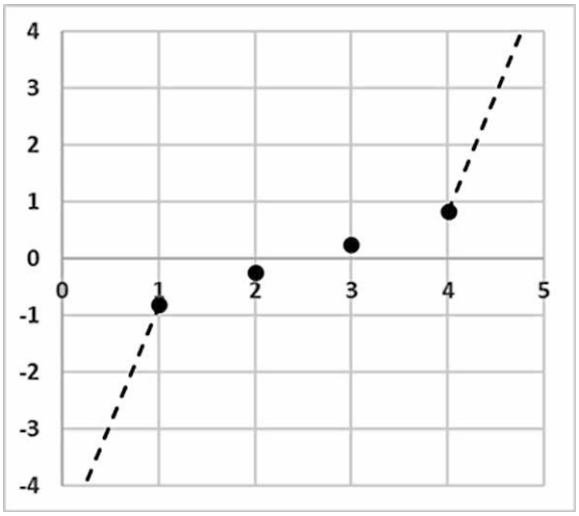
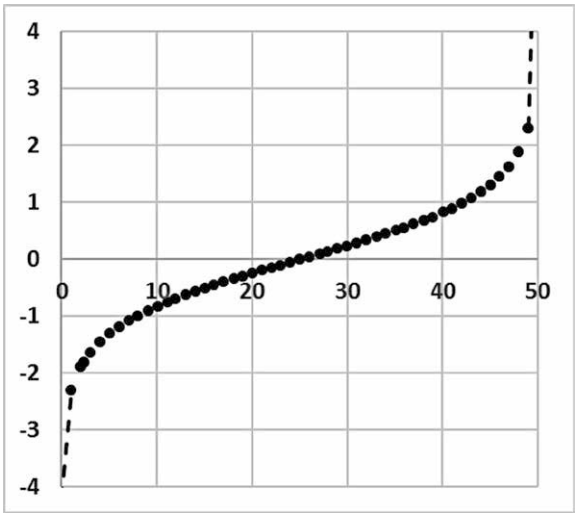


図４　困難度 $\beta=0$ の 50 項目が出題された場合の正答数別の特性値の最尤推定値

（横軸が正答数、縦軸が特性値）



11　ただし数多くの確率分布の積の連なりからなる尤度関数の扱いは容易ではない。このため実際の最尤推定は、まず尤度関数の対数値を取った対数尤度関数を作成し、これを θ で偏微分した対数導関数を求めて、さらにその対数導関数の解が 0 となる（＝尤度が最大となる）θ を求めることで行われる。またこの解を求めるには、複数回の計算を反復して漸近的に解に近づいてゆく「数値解法」を用いる必要があるとされる（豊田, 2012, pp.53-54）。PIAAC でも数値解法の一種である「EM アルゴリズム」が用いられているが（OECD, 2016, p.6）、EM アルゴリズムについては(4)② i で具体的に説明する。なお、最尤推定値の標本分布は、項目数 n が大きくなるに従って限りなく正規分布に近づくとされる（豊田, 2012, pp.88-89）。

ということはない。また、全項目正答と全項目誤答の場合は、最尤推定値が正の無限大と負の無限大になり現実的な特性値が求まらない。このため１項目正答，２項目正答，３項目正答，４項目正答の４パターンを推計すると、それぞれ約－0.83、約－0.24、約 0.24、約 0.83 となる。

一方、困難度 β＝0 のまま項目数を 50 に増やした場合が図４である。５項目しかない場合は特性値がプラスマイナス 0.83 を外れる回答者の特性値は推計できないし、プラスマイナス 0.83 に含まれる回答者の特性値の違いも４段階でしか区別できないが、50 項目の場合は、プラスマイナス 2.31 の間の回答者の特性値の違いを 49 段階で区別できる。このことから、項目数の多さが測定の精度を規定することが分かる。

ただし図４を見ると、特性値がプラス１程度以上若しくはマイナス１程度以下では推定の精度が次第に大まかになり、プラス２以上もしくはマイナス２以下では殆ど特性値の違いを識別できなくなっている。しかし項目の困難度を β＝2（図５）若しくは β＝－2（図６）にすると、それを中心とした特性値を持つ回答者を高い精度で測定することができるようになる。いずれの場合も項目の困難度から離れた回答者の特性値の測定精度は低下するので、対象となる回答者に項目の困難度を適切に合わせることが重要であることが分かる。

図５　困難度 β=2 の 50 項目が出題された場合の正答数別の特性値の最尤推定値

（横軸が正答数、縦軸が特性値）

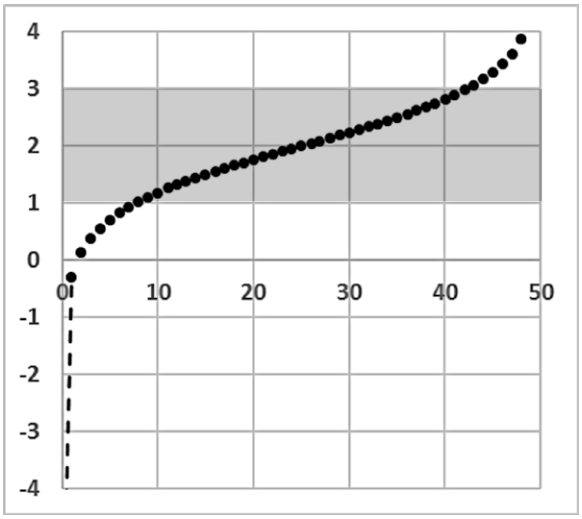
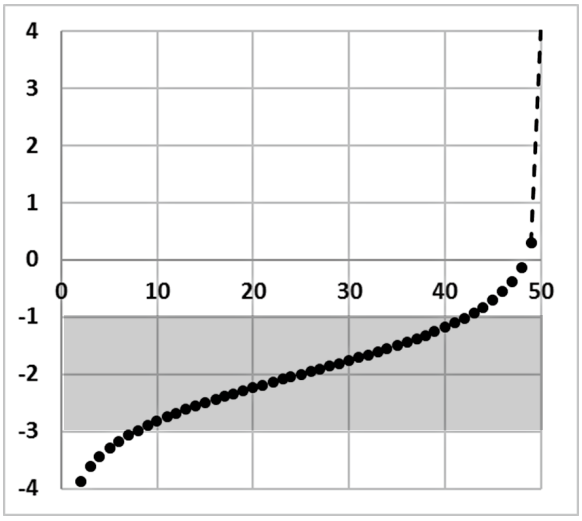


図６　困難度 β= －2 の 50 項目が出題された場合の正答数別の特性値の最尤推定値

（横軸が正答数、縦軸が特性値）



④　用語について－周辺分布と事後分布

今まで、結果としての解答パターンから原因である特性値を推計する方法を紹介してきたが、その際、困難度 β と識別力 α も未知であることは捨象してきた。これらの項目パラメーターも推計が必要だが、項目パラメーターはすべての回答者に共通する値であり、すべての回答者の解答結果を条件として θ，β，α を最尤推定しなければならない。この関係を表したのが以下の３式であり、式の左端の項にある大文字の X は N 人の回答者全員の、n 個の項目すべてに対する反応の束を意味している。

$$L(\theta,\beta,\alpha|X)=\prod_{j=1}^{N}L(\theta_j,\beta,\alpha|x_j)=\prod_{j=1}^{N}\prod_{i=1}^{n}L(\theta_j,\beta_i,\alpha_i|x_{ij}) \qquad 3\text{式}$$

しかし実際には、３式からθ，β，αの３つを同時に最尤推定するのには困難が伴うので、それを避けるためにいったんθを変数から除去してβとαとを最尤推定する「周辺最尤推定法」が採られるとされる[12]（豊田, 2012）。周辺という言葉が用いられるのは、複数の確率変数の組を確率要素とする同時確率分布から、一部の確率変数を消去した確率分布を周辺確率分布（marginal probability distribution）、若しくは単に周辺分布（marginal distribution）と言うからである[13]。

また、ベイズの定理に基づくベイズ統計学では、ある事象が生起したという条件の下で、別の事象が生起する確率を事後確率と呼ぶが、これは条件付確率と基本的に同じである。ベイズの定理は以下の４式で定義される[14]。

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \times P(B|A)}{P(B)} \qquad 4\text{式}$$

式の左辺の項の$P(A \mid B)$は、B が生起したと言う条件の下で A が生起する確率を意味し、式の右辺の項の分子にある$P(B \mid A)$は、A が生起したと言う条件の下で B が生起する確率を意味する。A と B を様々な値を取る確率変数と捉えれば、P は確率分布を示すことになり、$P(A \mid B)$は A の「事後分布」（posteriori distribution）と呼ばれ、$P(B \mid A)$は B の事後分布と呼ばれる。$P(\theta_j \mid x_{ij})$も、回答者 j の項目 i への反応x_{ij}が観察された場合の特性値θ_jの事後分布である。

OECD（2016, p.1）も、「PIAAC は、多重代入であり、IRT による認知項目（cognitive items）の尺度の測定と、母集団モデルでの背景調査の情報を用いた潜在回帰モデルとを統合した事後分布から抽出された plausible values を用いている。」（下線は筆者による）として「事後分布」の語を用いている[15]。このような複雑な言い方をしている理由は、この後順を追って説明をしていく中で明らかになるだろう。

（２） 幅広い層を対象に幅広く特性値を測定することに伴う制約

ここまで項目反応理論の基本的な考え方について説明してきたが、PIAAC のような大規模国際調査において、理想的な形でアセスメントを実施するのには大きな困難が伴うとされる。図７は読解力と数的思考力の項目群の構成である。ステージ１では３つの難易度別に分類された９項目から成る項目群（Testlet）の何れか一つが出題され、ステージ２では４つの難易度に分類された 11 項目から成る項目群の何れか一つが出題される。そして学歴やコアステージ（スクリーニングテスト）の解答などから特性値が高いと予想される回答者には高い比率で難易度の高い項目群が出題され、特性値が低いと予想される回答者には高い比率で難易度の低い項目群が出題される仕組みになっている（OECD, 2016, Chapter1, pp.10-16）。

---

12 θを変数から消去するためには関数をθで積分する必要があり、そのためにθの分布が標準正規分布に従うことが仮定される（豊田, 2012, pp.67-71）。

13 ここではθを変数から消去することでβとαの周辺分布を得るが、それによって推計したβとαの定数を用いて得られるθの確率分布も周辺分布である。

14 ベイズの定理については色々なウェブサイトでも解説されているので、適宜参照いただきたい。

15 「周辺」と言う語は、例えば国立教育政策研究所（2013, p.70）が用いている。曰く、「PIAAC で用いられている個々の成人の習熟度は、問題群への反応から推定される習熟度についての周辺事後分布からランダムに 10 個の推算値（plausible values）を取り出したものである。この分布の推定には、背景調査の一部の回答も用いられている。」（下線は筆者による）。言い方は異なっているが、基本的には OECD と同じことを述べている。

図７　読解力と数的思考力の項目群の構成

STAGE 1 (18 unique tasks – 9 tasks per testlet. Each respondent takes 1 testlet)							
	Block A1	Block B1	Block C1	Block D1			
Testlet 1-1	4 tasks	5 tasks					
Testlet 1-2		"	4 tasks				
Testlet 1-3			"	5 tasks			
STAGE 2 (31 unique tasks – 11 tasks per testlet. Each respondent takes 1 testlet)							
	Block A2	Block B2	Block C2	Block D2	Block E2	Block F2	Block G2
Testlet 2-1	6 tasks	5 tasks					
Testlet 2-2		"	3 tasks	3 tasks			
Testlet 2-3				"	3 tasks	5 tasks	
Testlet 2-4						"	6 tasks

図の出典：OECD（2016, Chapter1, p.13）

こうした出題方法を採ることは、回答者の負担を軽減しながら、極めて広い範囲に散らばっている一国の「成人」の特性値をできるだけ適切に推計するために必要なことであるが、読解力、数的思考力ともに一人の回答者に出題される項目数は 20 しかない[16]。村木（2009, p.40）は、個々の「受験者」に 60 項目かそれ以上の十分な項目数を与えることができれば、特性値の推定誤差を無視できるほど小さいものにすることができるとしているが、それとはかなり離れた状況にあると言える。

回答者に出題される項目群の内容が異なることも、推計の不確かさを増す要因となる。項目反応理論には、異なる回答者集団に課された異なる項目群の中に一定数の同一の項目が存在するか、若しくは、異なる項目群を課された異なる回答者集団の間に一定数の同一の回答者が存在すれば、推計された特性値を相互に比較可能にする手法があるが[17]、そこにも一定の不確実性が伴ってくる[18]。

（３）　項目反応に基づく推計方法と回帰式による推計方法との統合

前節で項目反応に基づく特性値の推計方法には相当な不確かさが伴うと考えられることを述べたが、PIAAC では、背景調査の回答から得られた情報を基に回答者の特性値を回帰式によって推計する方法も採用しており、これと項目反応に基づく推計方法とを統合することで、特性値の推計がより妥当なものになるようにしている。二つの異なる方法をどのように統合する[19]のかは後で説明することとして、まず、回帰式による推計方法を見てみよう。

①　回帰式による推計方法

回帰式による推計方法では次の５式のように、回答者の特性値はy Γを平均としΣを分散とする正規分布に従うものと想定されている。式中のΓとΣとが「潜在回帰パラメーター」（latent

16　Khorramdel et al.（2020, p29）は、個々の回答者は「わずかな数の答えを返すのみ」（provides only a small number of answers）で、「その分野の一部分を解答するのみ」（will only answer a subset of the domains）としている。

17　このための具体的な方法については豊田（2012, pp.114-126）を参照いただきたい。

18　このことについては脚注 27 を参照

19　このことに関して OECD（2016, p.5）は、「PIAAC で用いられた母集団モデル（population model）は、IRT モデルと潜在回帰モデル（latent regression model）との統合（combination）である。」と言う言い方をしている。

regression parameters）であるとされ、Γが回帰係数の行列、Σが残差の分散共分散行列（residual variance-covariance matrix）である。「潜在回帰」という言葉が用いられているのは、回帰式の推計に用いられる特性値θが、直接的には観察されない「潜在変数」（latent variable）[20]だからである。

$$\theta \sim N(y\Gamma, \Sigma) \qquad 5式$$

具体的に用いられる背景調査の設問から得た情報の例としてOECD（2016, p5）は、「性別、出身国、教育、職業、就業形態、読解力の使用状況（reading practices）、etc.」と記しているが、どこまでの範囲が用いられたのか詳しいことは分からない[21]。ただし設問の回答をそのまま変数として用いずに、主成分分析によって情報を集約し、主成分を説明変数としてγの推計を行ったとされ、また、説明変数と特性値との関係には国による違いがあることを考慮して、主成分分析とγの推計は国別に行ったとされる（Khorramdel et al., 2020, p.37）。

OECD（2016, p.5）は、Θを特性値の列ベクトル：$\Theta = (\theta_1, \cdots \theta_s)^t$ [22]、Γを回帰係数の行列：$\Gamma = (\gamma_{sl}, \quad s = 1, \cdots, S; \quad l = 0, \cdots, L)$[23]、$Y$を背景調査で得られた情報の列ベクトル：$Y = (1, y_1, \cdots, y_L)^t$として、$Y$の$\Gamma$による$\Theta$への潜在回帰モデル（latent regression model）を以下の６式で表し（式中のε_Sはスキルsの推計に関する誤差項を示す。）、またΣをその次の７式で表している[24]。

$$\theta_s = \gamma_{s0} + \gamma_{s1} y_{j1} + \cdots + \gamma_{sL} y_L + \varepsilon_s \qquad 6式$$

$$\Sigma = \Theta\Theta^t - \Gamma(YY^t)\Gamma^t \qquad 7式$$

② 項目反応に基づく推計方法と回帰式による推計方法との統合

次に２つの推計方法の統合について説明する。前節で見た６式は一見通常の重回帰式と変わらないが、しかし普通に考えると、被説明変数であるθを観測できないのだから、回帰式自体を作成することができないのではないだろうか。実はこの問題を解決するのが２つの推計方法の統合である。

このことに関してOECD（2016）を見ると、「各回答者jのPVsは、この条件付分布から求められる」として、次の８式の条件付確率分布が掲げられている。これは、項目反応x_j、背景調査の情報y_j、及び潜在回帰パラメーターのΓとΣのすべてに条件付けられた特性値θ_jの条件付確率を意味している。その次に掲げられている式が、８式をベイズの定理を用いて変形したのが９

20 潜在変数という言葉はMislevy（1991）で用いられているが、OECD（2016）では使われていない。

21 この手法について考察した初期の論文であるMislevy（1985）では、米国で行われた調査を用いて人種と性別の影響について分析を行っている。

22 添え字のsはPIAACで測定したスキルの数（＝3）を表し、tはベクトルや行列の転置を表している。

23 添え字のl（エル）は背景調査で得られた情報の数、sはスキルの数を表している。

24 ７式は確率変数の分散の公式である$V(\theta) = E(\theta - \mu)^2 = E(\theta^2) - \mu^2$を表していると思われる。ここでは$\mu = y\Gamma = \gamma_{s0} + \gamma_{s1} y_{j1} + \cdots + \gamma_{sL} y_L$である。

式[25] だが、これがどういうことを意味しているのか特に説明がない。そこで OECD も引用している Mislevy（1985）を参照してみよう。

$$P(\theta_j \mid x_j, y_j, \Gamma, \Sigma) \qquad 8\text{式}$$

$$P(\theta_j \mid x_j, y_j, \Gamma, \Sigma) \propto P(x_j \mid \theta_j, y_j, \Gamma, \Sigma)\, P(\theta_j \mid y_j, \Gamma, \Sigma) = P(x_j \mid \theta_j)\ \mathrm{P}(\theta_j \mid y_j, \Gamma, \Sigma) \quad 9\text{式}$$

Mislevy（1985, pp.993-994）はまず、「完全なデータがある場合の解法」（the complete-data solution）として、変数 x からパラメーターである Γ と Σ とを導出する方法を示し、次にそれと対置される「不完全なデータしかない場合の解法」（the incomplete-data solution）として、変数 x が観測できない場合に、$P(y \mid x)$ が既知である変数 y を用いて Γ と Σ とを導出する方法を示している。そこで示されているのが以下の 10 式であり、式中の $P(y \mid x)$ は項目反応理論に基づく回答者の特性値 x と回答者の項目反応 y との関係を表している。用いられている記号が異なるので分かりにくいが、９式の右側の項と 10 式の右側の項は基本的に同じ構造であり、10 式は、$P(y \mid x)$ $f_k(x \mid \Gamma, \Sigma)$ から x を積分で変数から消去することで $g_k(y \mid \Gamma, \Sigma)$ が得られること、これにより y を用いて Γ と Σ とが推測できることを示している[26]。

$$g_k(y \mid \Gamma, \Sigma) = \int_x P(y \mid x)\ f_k(x \mid \Gamma, \Sigma)\, dx \qquad 10\text{式}$$

Mislevy（1985, p.994）はまた、10 式の関数の対数値を取ってそれを Γ で偏微分することでその最尤（さいゆう）推定値を求めるという方法を示している。これは脚注 11 で述べた最尤（さいゆう）推定値を求める方法と同じであり、すなわち 10 式の条件付確率を最大化する Γ を推計することを意味している。このようにして得られた Γ を用いて、項目反応に基づく推計方法と回帰式による推計方法とを統合した９式の右辺の項で求めるのは、２つの条件付確率の積である $P(x_j \mid \theta_j)$ $P(\theta_j \mid y_j, \Gamma, \Sigma)$ を最大化する θ_j であるということになる。そしてこのようにして求められた θ_j は、項目反応理論に基づいて $P(x_j \mid \theta_j)$ の尤度（ゆうど）だけを最大化する θ_j よりも好ましいと考えられている[27]。

---

25 ベイズの定理による８式の９式への変形方法を単純化して示すと以下となる。∝は比例を意味する記号である。

① $P(A \mid B \cap C) = \dfrac{P(A \cap B \cap C)}{P(B \cap C)}$ であり、$\mathrm{P}(B \mid A \cap \mathrm{C}) = \dfrac{P(A \cap B \cap C)}{P(A \cap C)}$ なので、

$$P(A \mid B \cap C) = P(B \mid A \cap C) \times \frac{P(A \cap C)}{P(B \cap C)}$$

② また、$P(\mathrm{A} \cap \mathrm{C}) = P(A \mid C) \times P(C)$ であり、$P(B \cap C) = P(B \mid C) \times P(C)$ なので、

$$P(A \mid B \cap C) = P(B \mid A \cap C) \times \frac{P(A \mid C) \times P(\mathrm{C})}{P(B \mid C) \times P(\mathrm{C})} = P(B \mid A \cap C) \times \frac{P(A \mid C)}{P(B \mid C)} \ \propto\ P(B \mid A \cap C) \times P(A \mid C)$$

となる。
なお、９式の中央の項にある $P(x_j \mid \theta_j, y_j, \Gamma, \Sigma)$ が右端の項では $P(x_j \mid \theta_j)$ となっているのは、x_j が θ_j にのみ依存し、他の変数に対しては独立であるとの仮定に基づくものと考えられる（村木, 2009, p.41）。

26 添え字の k は回答者が属する集団の属性を示しており、g_k 及び f_k は集団の属性によって特性値の分布が異なることを想定したモデルであることを意味している。

27 このことに関して Khorramdel et al.（2020, p.41）は以下の通り述べている。「PIAAC の IRT の尺度は、異なる項目群を出題された回答者群の比較可能性の問題を、項目と特性値の双方に同じ尺度を適用することで解決している。しかし測定に伴う不確かさのために、IRT から得られた個人の特性値の点推定は深刻な歪みをもたらす可能性がある。このため PIAAC は、回帰モデルから得られた PVs を提供することで、集団レベルの効果を回帰によって適切に制御し、測定誤差を減少させながら、集団レベルでの比較の歪みを取り除いている。」

（4） 多重代入法の基本的な考え方と 10 個の PVs の算出方法

最後に本節では、なぜ、そしてどうやって 10 個の PVs が算出されているのかということについて説明する。そのためにまず多重代入法（multiple imputation）の基本的な考え方を説明して、次に具体的な PVs の算出方法について説明する。

① 多重代入法の基本的な考え方

多重代入法は、データの欠測に対応するための代入法の一つである。例えば、PIAAC の背景調査では、回答者の性別、年齢、学歴、収入等について聞いているが、収入について回答が得られずに欠測となっている回答者がいるかもしれない。そのような場合に、欠測のない回答者のデータを用いて、以下の式のように収入を被説明変数として、性別、年齢、学歴を説明変数とする回帰式を作成すれば、欠測している収入について一定程度信頼できる推計値を得られるかもしれない。

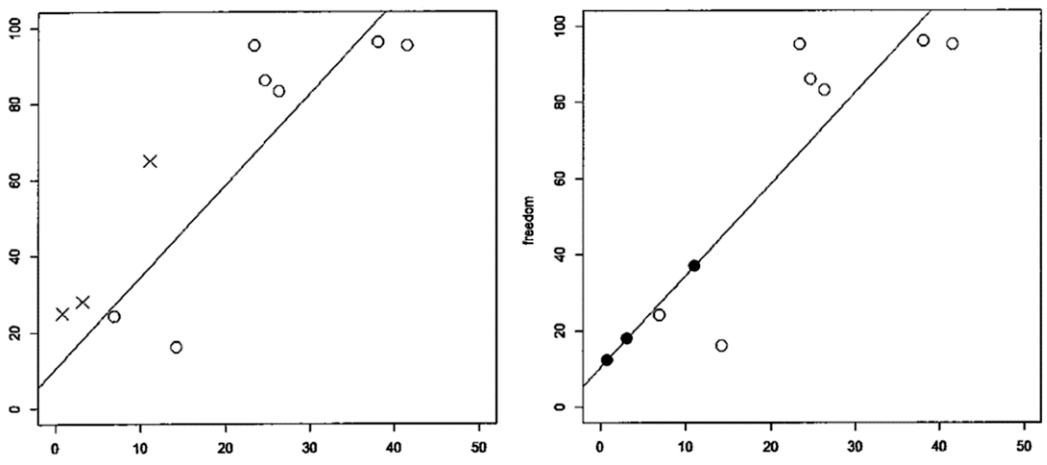
$$収入_j = \gamma_0 + \gamma_1 性別_j + \gamma_2 年齢_j + \gamma_3 学歴_j \qquad 11\text{ 式}$$

このようにして推計した値で欠測を補うのが「代入法」（imputation）だが、回帰式に基づく推計は、５式のように平均値と分散という形で得られるものであり、そこには一定の「ばらつき＝不確かさ」が伴っている。「単一代入法」（single imputation）は、１個の推計値、例えば回帰式で得られた平均値だけで欠測を補う方法だが、推計のばらつきが考慮されていないことから、これによる代入を行った場合は変数全体の分散の大きさが過小評価されることになる。図８は、本来の分布と単一代入法で欠測を補った場合の分布とを比較した図である。ここでは欠測した変数を推計する説明変数が一つしかない単回帰であるために、右側の図ではすべての代入値が回帰直線の真上に並ぶという極端な結果となっているが、単一代入法の欠点を良く表している。

これに対して「多重代入法」は、推計値の確率分布から無作為に複数個の推計値を抽出して用いる方法であり、複数個の推計値のばらつきによってデータを用いた分析での誤差の評価が妥当なものになることが企図されている。PIAAC で各回答者にそれぞれ 10 個の PVs が算出されているのもこのためである。高橋等（2017, p.45）は、異なる複数の推計値が代入されたデータセットについて、代入されたデータを１つに統合してから統計分析を行ってはならず、それぞれを別々

図８ 本来の分布（左の図）と、単一代入法で代入した分布（右の図）

（〇が観測値、×が本来の欠測値、●が単一代入法による欠測値の代入値）



図の出典：高橋等（2017, p.36）

に使用して統計分析を行ってから複数の結果を１つに統合して最終結果とするべきであるとしているが、推計の不確かさを表すために敢えて複数の推計値が算出されていることに鑑みれば、この指摘の意味もよく理解できる。

以上が PVs が 10 個あることの理由の説明となるが[28]、しかし PIAAC の場合は、ここで述べたように一部の回答者についてだけ特性値が欠測しているわけでない。この違いに関する考え方を Mislevy（1991, pp.179-178）が述べているので以下に要約して紹介する。

ⅰ　データの欠測に関しては “Missing at random”（MAR）という概念があり、それはデータの欠測が当該データの値に依存しない[29] ことを意味している。MAR の場合は、観測されたデータ（Y_{obs}）の分布を用いて、観測されないデータ（y_{mis}）の分布の条件付確率 $p(y_{mis} \mid y_{obs}, z)$ を仮定してパラメーターを推測することができる（筆者注：z は調査のデザインに関するパラメーター）。

ⅱ　直接観測できない潜在変数についても、すべての回答者の回答が欠測していると考えれば MAR に該当すると見なすことができる。潜在変数である特性値 θ の分布は、属性を示す Y と調査デザインを示す Z、及び母集団における分布を示す α を用いる母集団モデル $P(\theta \mid Y, Z, \alpha)$ と、項目反応を示す X と項目パラメーターを示す β を用いる潜在変数モデル $P(X \mid \theta, \beta)$ の２つから推測することができる[30]。

上記の ii に出てくる α は θ の母集団分布 $P(\theta \mid \alpha)$ を示すパラメーターであるとされ、PIAAC で用いられる Γ と Σ とに相当するものである。Mislevy（1991, p.179-180）は、α に焦点が置かれる場合は、個人を単位とする点推定が大きく歪(ゆが)みうることを指摘しているが、こうした指摘を考慮すると、「母集団モデル」は、個人の特性値の測定精度を補完的な情報を用いて高めるというより、集団における特性値の分布を正しく測定するという独自の目的のために採用されていると考えられる。

②　10 個の PVs の算出手順

ⅰ　EM アルゴリズムについて

10 個の PVs の具体的な算出手順を説明するに当たり、まず、それに用いられる EM アルゴリズムから説明したい。EM アルゴリズムについては赤穂（1996, p.44）が分かりやすく解説しているので、以下にその概略を記す（数式の添え字は筆者が変更して使用している）。

観測されるデータ y と、観測されないデータ x とがあり、x が分かっていれば関心のあるパラメーター ξ の最尤(さいゆう)推定が容易になる場合が前提とされる。まず ξ について、その初期値を適当な点 $\xi = \xi^{(0)}$ にとる。次に y と $\xi^{(0)}$ とが与えられたときの x の分布をベイズの公式に基づいて推計し、対数尤度(ゆうど) $\log f(x \mid \xi)dx$ の、データ y とパラメーター $\xi^{(0)}$ に関する条件付平均（期待値）を求める（E ステップ：Expectation step）。数式で表すと以下の 11 式を計算することとなる。

28　10 個という数自体に特に強い根拠があるわけではないと思われる。高橋等（2017, p.53）は、代入する推計値の数は多いほどよく、100 程度にするのがよいという考えも述べている。

29　たとえば所得を尋ねる設問について、所得が高い人ほど回答を拒否する率が高くなるようなら MAR ではない。

30　脚注 19 に記したように OECD は、IRT モデルと潜在回帰モデルとを統合したものが母集団モデルだという言い方をしており、微妙に用語の使い方が異なっている。

$$Q(\xi) = E\left[logf(x|\xi)|y,\xi^{(0)}\right] = \int f\left(x|y,\xi^{(0)}\right)logf(x|\xi)dx \qquad 11\text{ 式}$$

次に、推計した x の分布を用いた上記の Q（ξ）を最大化する ξ を新たに推計して $\xi^{(0+1)}$ とし（M ステッ Maximization step）、そこから更に E ステップと M ステップとを繰り返していくことで、最終的に ξ について収束した解を得ることができるとされる。

PIAAC での EM アルゴリズムを用いた操作に関して OECD は具体的な数式を示していないが、上記の赤穂の数式の変数を PIAAC のモデルで用いる変数で置き換えると以下の 12 式のようになる。

$$Q(\Gamma,\Sigma) = E\left[logf(\theta|\Gamma,\Sigma)|x_j,y_j,\Gamma^{(p)},\Sigma^{(p)}\right] = \int f\left(\theta|x_j,y_j,\Gamma^{(p)},\Sigma^{(p)}\right)logf(\theta|\Gamma,\Sigma)d\theta \qquad 12\text{ 式}$$

この式に基づいて行うと考えられる操作を簡単に説明する。最初に仮の $\Gamma^{(0)}$, $\Sigma^{(0)}$ を置き、その下で潜在変数 θ（赤穂の式の x に相当）の仮の値を計算し、その仮の θ を用いて対数尤度（ゆうど）の期待値を求める（E ステップ）。次にその仮の θ の下で対数尤度（ゆうど）の期待値が最大になる $\Gamma^{(0+1)}$, $\Sigma^{(0+1)}$ を求める（M ステップ）。この２つのステップを繰り返すことで、計算が収束して最終的な Γ と Σ の確率分布が求まる。

ii　PVs の具体的な算出手順

最後に、OECD（2016, p.6）に記された PVs の具体的な算出手順の概要を以下に紹介する。

① EM アルゴリズムを用いて Γ と Σ とを推計する[31]。② Γ と Σ の推計が完了したら、正規近似する確率分布 $P(\Gamma,\Sigma\,|\,x_j,y_j)$ から一つの値の Γ を抽出し、それによって Σ の値を $\widehat{\Sigma}$ に固定する。③ 得られた Γ と Σ とに条件付けられた θ の事後分布の平均値 m_j^p と分散 Σ_j^p を計算する[32]。④ 平均 m_j^p、分散 Σ_j^p の多変量正規分布から一つの値の θ を抽出する[33]。⑤ 以上の手順を 10 回繰り返して回答者ごとに θ の 10 個の代入値を算出する。

４．PVs を用いた分析を行う場合の計算方法

PVs を用いた分析を行う場合の具体的な計算方法は、OECD（2009, Chapter 8）に詳しく解説されており、ここでは概略のみを掲げる。平均値、比率、回帰係数の何（いず）れも同じ操作で処理できる。

PVs によって取り込まれた特性値の推計の不確かさは、複数個の PVs を用いた推計値間の分散という形で表れてくる。高橋等（2017, p.44）は、代入モデルの精度が高い場合は推計値間の分散が小さくなり、精度が低い場合は分散が大きくなるとしているが、そのことは直感的にもよく理解できる。

31　この際、赤穂は前述の通り E ステップで潜在変数の分布を推計し、さらにそれを用いた対数尤度の期待値を求めているのに対して、PIAAC では潜在変数とその分散の期待値を求めることを以て E ステップとしており（Mislevy（1985, p.994））、その点に違いがある。

32　右肩の添え字 p は事後分布（posterior distribution）の p を示していると思われる。

33　OECD（2016）には具体的な記載がないが、PISA のテクニカルレポートである OECD（2014, pp.146-147）には PVs の具体的な抽出方法が記されており、これを見ると、推計される θ の事後分布からの様々な値の PVs の抽出確率は、当該値の確率密度に比例している。（つまり、θ の平均に近い値ほど抽出確率が高まる。）

1．最初に 10 個の PVs を 1 個ずつ（ここでの 1 個の意味は、図 1 にある縦の列 1 列という意味である。）使って、関心のある推計値（ここでは平均値 μ を取り上げた。）を求めてから、最後にその平均をとって最終的な推計値とする。

$$\mu = \frac{\mu_1 + \mu_2 + \mu_3 + \mu_4 + \mu_5 + \mu_6 + \mu_7 + \mu_8 + \mu_9 + \mu_{10}}{10} \qquad 13 \text{ 式}$$

2．PVs を用いて得られた推計値の分散については 2 種類の分散を合計する。1 つは 10 個の推計値それぞれの標本分散の平均であり、もう 1 つは PVs を用いて行った 10 個の推計値間の分散である。ただし後者は 10 個の推計値の標本分散を求めてから、それに 1+1/10 を乗ずることとされている。

$$\sigma_E^2(final\ Error\ vriance) = \sigma_S^2(final\ Sampling\ variance) + \sigma_I^2(Imputation\ variance) \qquad 14 - 1 \text{ 式}$$

$$\sigma_S^2 = \frac{\sigma_{\mu_1}^2 + \sigma_{\mu_2}^2 + \sigma_{\mu_3}^2 + \sigma_{\mu_4}^2 + \sigma_{\mu_5}^2 + \sigma_{\mu_6}^2 + \sigma_{\mu_7}^2 + \sigma_{\mu_8}^2 + \sigma_{\mu_9}^2 + \sigma_{\mu_{10}}^2}{10} \qquad 14 - 2 \text{ 式}$$

$$\sigma_I^2 = \frac{1}{10-1}\sum_{i=1}^{10}(\mu_i - \mu)^2 \times \left(1 + \frac{1}{10}\right) \qquad 14 - 3 \text{ 式}$$

OECD（2009）には上記の計算を行うための PISA 用の SPSS のマクロも掲載されており、また、SPSS と SAS については、国際教育到達度評価学会の International Database Analyzer（IDB Analyzer）を用いることで、PIAAC の調査デザインを考慮した推定を行うことも可能である（卯月, 2022, p.67）。筆者はフリーの統計ソフト R を使っており、同様な計算を R のパッケージ survey にある withPV 関数を用いて行っている[34]。

5．終わりに

国立教育政策研究所（2013, p.70）は、「この推算値（筆者注：PVs を指す）は、各国の母集団やサブ・グループを推定するために算出したものであり、成人個人の習熟度を明確に特定することはできない。」としている。また Khorramdel et al.（2020, p.38）も、「回帰モデルに基礎を置いた PVs は集団の特性を推定するためにデザインされている。PVs は個人のテストのスコアに代替するものではなく、決して個人レベルの推測のために用いてはならない。」としている。まずは筆者自身が、本稿を作成することを通じてこれらの指摘の意味をよく理解できたと感じている。

最後に 2 つのことを追加的に取り上げてみたい。冒頭の図 1 に掲げた 10 人の回答者の各 10 個の PVs の分散を計算し、± 1σ と ± 2σ の区間を計算したものが図 9 である。それぞれの PVs の分散にはかなりの違いがあり、6 番目の回答者では ± 1 シグマ区間の幅が 45.4、± 2 シグマ区間の幅が 90.9 あるが、10 番目の回答者ではそれぞれ 18.1 と 36.2 しかない。このような違いが生ま

34 PVs を分析に用いるかどうかに関わりなく、PIAAC のデータを用いて推計を行う場合は、ジャックナイフ法と呼ばれる方法を用いて標準誤差を計算する必要があるが、これも R のパッケージ survey にある svrepdesign 関数を使うことで計算が可能である（PVs を用いた推計では、svrepdesign 関数と withPV 関数を組み合わせて計算する。）。ジャックナイフ法を用いた計算を行うためには、PIAAC のミクロデータファイルの 1200 列目にある SPFWT0 と、1201 ～ 1280 列にある 80 個の replicate weight を用いる必要があるが（列数は日本のファイルのもの）、その R への取り込み方は Lemley（2010, p.25）に書かれている。

図９　10 個の PVs の分散から計算した±１シグマ区間と±２シグマ区間

SEQID	PVLIT1	～	PVLIT10	10個のPVsの平均値	10個のPVsの分散	10個のPVsの標準偏差	－１σ	＋１σ	－２σ	＋２σ
1	290.8536	～	299.4686	298.726929	118.101661	10.8674588	287.8595	309.5944	276.992	320.4618
2	277.1685	～	279.8756	282.972492	133.391067	11.549505	271.423	294.522	259.8735	306.0715
3	302.4788	～	292.7362	300.301866	119.694518	10.940499	289.3614	311.2424	278.4209	322.1829
4	304.5744	～	288.8517	303.271672	219.166511	14.8042734	288.4674	318.0759	273.6631	332.8802
5	335.5443	～	329.9094	330.083982	354.894721	18.8386497	311.2453	348.9226	292.4067	367.7613
6	339.3943	～	317.3439	356.225006	516.251301	22.7211642	333.5038	378.9462	310.7827	401.6673
7	303.7377	～	274.7586	297.542002	193.955974	13.9268077	283.6152	311.4688	269.6884	325.3956
8	338.4039	～	308.0711	320.963222	425.261172	20.6218615	300.3414	341.5851	279.7195	362.2069
9	221.493	～	233.3929	224.51736	134.374949	11.5920209	212.9253	236.1094	201.3333	247.7014
10	322.1816	～	313.1191	315.095046	81.8249814	9.04571619	306.0493	324.1408	297.0036	333.1865

れてくるのは、もちろん抽出の偶然性の影響もあるだろうが、主には、回答者の特性値と出題された項目の困難度の合致の度合いに違いがあったからだと考えられる。よく合致していれば分散は小さくなり、そうでなければ分散は大きくなると考えられる。

もう一つ、４で記した方法を用いて日本の回答者の読解力の平均値を、実際に 10 個の PV を使って、サンプルサイズを 10、100、1000 と変えて計算してみた結果が表１である。③が 10 個の PVs を用いた推計値間の分散だが、②の通常の意味での標本分散に較(くら)べてかなり小さく、またサンプルサイズが大きくなるほど分散が小さくなっていることが分かる。脚注 28 で代入する推計値の数についても議論があることを記したが、これを見る限り、PVs の数を更に大幅に増やして誤差の計算を精密化することの意味は大きくないと思われる。

表１　サンプルサイズの違いによる、PVs を用いた読解力の平均値の分散の変化

	① 10 個の PVs の平均値の平均値	② 10 個の PVs の標本分散の平均値	③ 10 個の PVs の平均値の標本分散× 11/10	④ 最終的な分散（②＋③）	平均値の 95% 信頼区間
サンプルサイズ：10	303.0	1374.9	42.10	1417.0	279.6 ～ 326.3
サンプルサイズ：100	299.0	1785.4	1.48	1786.8	290.7 ～ 307.3
サンプルサイズ：1000	298.2	1614.4	0.44	1614.9	295.7 ～ 300.7

付　記

本稿の作成に当たり、未熟な原稿を辛抱強く読んでいただき、貴重な御意見を頂いた二人の審査員の方に対して、深く御礼を申し上げます。また、いつも多大な励ましを頂いている畏友 K 氏にも深く御礼を申し上げます。

本研究は JSPS 科研費 JP20K02623 の助成を受けたものです。

【参考文献】

Khorramdel, L., von Davier, M., Gonzalez, E. & Yamamoto, K.（2020）. Plausible Values: Principles of Item Response Theory and Multiple Imputations.

（※ D. B. Maehler, B. Rammstedt 編（2020）. *Large-Scale Cognitive Assessment – Analyzing PIAAC Data*. Springer に Chapter 3 として収録）

Lemley, T.（2010）. *Complex Surveys*. Wiley.

Mislevy, R. J.（1985）. Estimation of latent group effects. *Journal of the American Statistical Association*, *80*（392）, 993-997.

Mislevy, R. J.（1991）. Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*（2）, 177-196.

OECD（2009）. *PISA Data Analysis Manual SPSS SECOND EDITION.* OECD.

OECD（2014）. *PISA 2012 Technical Report*. OECD.

OECD（2016）. *Technical Report of the Survey of Adult Skills*（*PIAAC*）（*2nd edition*）. OECD.
（※ https://www.oecd.org/skills/piaac/PIAAC_Technical_Report_2nd_Edition_Full_Report.pdf（最終閲覧日 2023 年 2 月 15 日））

Thomas, N.（1993）. Asymptotic Corrections for Multivariate Posterior Moments with Factored Likelihood Functions. *Journal of Computational and Graphical Statistics*. Vol. 2, No. 3（Sep., 1993）, 309-322.

赤穂昭太郎（1996）.「EM アルゴリズムの幾何学」『情報処理』Vol.37, No.1, 43-51.

卯月由佳（2022）.「国際比較データ」『日本労働研究雑誌』2022 年 4 月号（No.741）, 65-69.

国立教育政策研究所編（2013）.『成人スキルの国際比較 OECD 国際成人力調査（PIAAC）報告書』明石書店.

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.

豊田秀樹（2012）.『項目反応理論［入門編］【第２版】』朝倉書店.

登藤直弥（2012）.「項目反応間の局所依存性が項目母数の推定に与える影響 項目母数の比較可能性を確保した上での検討」『行動計量学』第 39 巻第 2 号, 81-91.

塗師斌（1989）.「二値データに基づく尺度の一次元性の評価の方法」『横浜国立大学教育紀要』第 29 巻, 137-148.

村木英治（2009）.「社会調査としての学力テスト」『社会と調査』第 2 号, 35-42.

（受理日：令和 5 年 2 月 24 日）