

平成 30 年 6 月 18 日現在

機関番号：62601

研究種目：基盤研究(C) (一般)

研究期間：2015～2017

課題番号：15K04388

研究課題名(和文)PISA調査の検証に基づく大規模教育調査実施方法のモデル構築

研究課題名(英文)Building a standard survey model for large scale educational surveys by reviewing PISA

研究代表者

褒岩 晶(Horoiwa, Akira)

国立教育政策研究所・国際研究・協力部・総括研究官

研究者番号：00626210

交付決定額(研究期間全体)：(直接経費) 3,600,000円

研究成果の概要(和文)：本研究は、「OECD生徒の学習到達度調査」(PISA調査)の調査手法とその有用性を明らかにし、日本の様々な教育調査に活用できるような調査モデルを構築することを目的としている。特に、1)教育調査における「層化二段抽出」とデータの重み付けの必要性、2)標本抽出にともなう標準誤差の計算方法としてのBRR法の有効性、3)大規模教育調査の能力測定における項目反応モデルと母集団モデルの必要性、4)複数の能力を測定する際の多次元項目反応モデルの有用性、5)大規模教育調査に適した項目反応モデルとしての2パラメータロジスティックモデルの有効性を検証した。

研究成果の概要(英文)：This study aimed to build a standard survey model for large scale educational surveys in Japan by reviewing the survey method and its effectiveness of OECD Programme for International Student Assessment (PISA). We reviewed 1)the necessity of "stratified two level sampling" and data-weighting in educational surveys, 2)the efficacy of Balanced Repeated Replication Method for calculating standard errors of sampling, 3)the necessity of using item response model and population model for ability-measurements in large scale educational survey, 4)the usefulness of multidimensional item response model in measuring multiple abilities, and 5)the adequacy of 2 parameter logistic model as item response theory for large scale educational surveys.

研究分野：教育調査法

キーワード：PISA調査 大規模教育調査 国際学力調査 教育調査法 教育社会学 教育学

1. 研究開始当初の背景

「OECD 生徒の学習到達度調査 (PISA 調査) は、日本も参加している 15 歳児を対象とした学力に関する大規模国際比較調査であり、2000 年の第 1 回調査から 3 年ごとに実施され、その結果が日本を含む様々な国の教育政策に影響を与え、教育学、統計学といった学術分野の研究対象ともなっている。

日本における PISA 調査の学術研究は、「PISA 調査が前提とする学力観とそれを実現するための教育課程」に関するものと、「PISA 調査のデータを使った国際比較や 2 次的データ分析」とに大別される。しかし、その間をつなぐ「調査の実施」については、PISA 調査の結果報告書 (国立教育政策研究所 2013) で言及されるにとどまってきた。

日本では、大規模教育調査を実施するために必要とされる調査手法全体を焦点に据えた研究は行われてこなかった。また、海外では PISA 調査等の大規模教育調査の関係者が、調査手法に関するさまざまな業績を発表しているが、それを日本で紹介している研究も少ない。調査手法の研究が欠落し、その重要性が顧みられなかったために、日本で行われている学力調査は、国際的に標準となっている調査手法を用いてこなかった。

2. 研究の目的

本研究は、PISA 調査の調査手法を明らかにすることを目的としている。

PISA 調査では調査が行われた年の 2、3 年後に、実際の調査で行われた手法や国際報告書での分析方法の概要が、OECD (2014) から「Technical Report」として公表されている。しかしながら、その手法が採用された経緯や理論的な妥当性、実際に行われている得点の算出方法等の詳細までは記述されていない。本研究では、これらの報告書やその他の調査関係資料に基づくだけでなく、研究代表者や研究分担者が、日本における PISA 調査を実施してきた経験とそこで得られた知見や情報を用い、PISA 調査における実際の調査手法が採用された経緯と実用上の有用性を明らかにし、そしてそれらを日本の様々な学力・能力調査にも活用できるような調査モデルを構築することを目指す。

3. 研究の方法

本研究では特に、(1) PISA 調査の標本抽出法、(2) 標本抽出にともなう標本誤差の計算方法、(3) PISA 調査の得点の計算方法とデータ補完の仕組み、(4) 得点計算における多次元項目反応モデルの有効性、(5) 大規模教育調査に適した項目反応モデルを検証する。(1) では、PISA2012 年調査の日本のデータを用い、多段抽出と層化抽出の影響 (デザイン効果)、「データの重み付け」の効果を調べ、(2) の前半では、同じデータで様々な標本誤差の計算方法を比較する。(2) の後半では、日本の状況に合わせた架空の有限母集

団のデータを作成し、PISA 調査の標本誤差の計算方法が有効であることをシミュレーションで明らかにする。(3) では、能力が標準正規分布に従う無限母集団から無作為抽出したサンプルを使って、架空の解答データを作成し、項目反応モデルのみで分析した場合と、PISA 調査で用いられている「母集団モデル」を加味して分析した場合とで比較する。(4) では、多変量標準正規分布を使って、2 つの能力を測定した架空の解答データを作成し、多次元項目反応モデルの有効性を検証するシミュレーションを行う。(3) と (4) では、ラッシュモデルと呼ばれる 1 パラメータ項目反応モデル (1PL モデル) のみを扱っているが、(5) では、2 パラメータ (2PL)、3 パラメータ (3PL) の項目反応モデルに従った架空の解答データを作成し、大規模教育調査に適した項目反応モデルを考察する。

4. 研究成果

(1) PISA 調査の標本抽出方法と重み付け

PISA 調査の標本抽出は、「層化二段抽出法」で行われている。日本の PISA 調査の母集団は「日本の学校に通う 15 歳児」、すなわち高校 1 年生であるが、その母集団から生徒を選ぶ際に、まず二段抽出の第 1 抽出単位として学科 (学校単位で行うと普通科と専門学科を区別できない) を所属する生徒数に基づく確率で 200 校選び、次に第 2 抽出単位として生徒を選ばれた学科の中から 35 名抽出している。サンプルサイズは、200 学科 × 35 名の 7000 名で、PISA2012 年調査の場合、有効解答数は 191 学科 6351 名となっている。

標本調査として学力調査を行う場合、「各学校から 1 人参加してもらおう」というのは難しいため、多段抽出は避けられない。しかし、母集団の一部しか調査しないことで生じる「標本誤差」は、多段抽出を行うことで大きくなる。PISA2012 年調査の数学的リテラシーにおける日本の平均得点の標本誤差を単純無作為抽出と仮定して計算を行った場合 (以下の分散の計算は HLM7 を使用)

$$\sqrt{\frac{\text{分散}}{\text{生徒数}}} = \sqrt{\frac{8899}{6351}} = 1.18$$

であるが、二段抽出 (層化なし) を仮定した場合は (Cochran (1953) を参照)

$$\sqrt{\frac{\text{学校間分散}}{\text{学校数}} + \frac{\text{学校内分散}}{\text{生徒数}}} = \sqrt{\frac{4780}{191} + \frac{4118}{6351}} = 5.07$$

となり、デザイン効果 (標本分散を単純無作為抽出と仮定した場合のそれで割った値) は、

$$\frac{5.07^2}{1.18^2} = 4.28^2 = 18.33$$

となる。SPSS のような通常の統計ソフトでは、標本誤差は単純無作為抽出を前提に計算されるため、4 分の 1 以下の過小推定をしていることになる。ただし、PISA 調査で用いられている標本誤差の計算方法 (ここで用いる 3.59 は、標本誤差と測定誤差を加味した標準誤差の値) で、デザイン効果を計算すると、

$$\frac{3.59^2}{1.18^2} = 3.03^2 = 9.18$$

となる。このデザイン効果の違いは、「層化」の影響と考えられる。

層化は母集団をグループ（層）に分け、層別に抽出を行うことで、標本誤差を小さくする効果がある。PISA 調査では、第 1 抽出単位である学科を「公立、私立」、「普通科、専門学科」を基準に 4 層に分け、これらを「主層」と呼んで、主層別に抽出を行っている。また、主層内で、「大学・短大進学者割合」を用いて学科を 4 グループに分け、これを「副層」と呼んで、学科を系統抽出する際の並び順に反映させている。「主層」のデザイン効果は、

$$\frac{\text{級内分散}}{\text{級内分散} + \text{級間分散}} = \frac{8363}{8831} = 0.97^2 = 0.95$$

であり、数%しか標本誤差を小さくしない。「副層」の情報は公開されておらず、直接検証できないが、学科を系統抽出する際の並び順（副層ごとに学科の生徒数で並び替えている）の効果が高いと考えられる。

PISA 調査では、層ごとに生徒の抽出確率が若干異なる点や、学科や生徒の不参加による欠損を補うために、データの重み付けを行っている。ある学科 j の生徒 i の重み W_{ij} は、

$$W_{ij} = f_{2ij} \times f_{1j} \times w_{2ij} \times w_{1j}$$

w_{1j} : 学科が選ばれる確率の逆数

w_{2ij} : 学科から生徒が選ばれる確率の逆数

f_{1j} : 参加しなかった学科を補うための補正

f_{2ij} : 参加しなかった生徒を補うための補正

で求められる。PISA2012 年調査の数学的リテラシーの結果を重み付け前と後で比較すると、以下ようになる。

	重み付け前	重み付け後	差(後 - 前)
平均得点	535.9	536.4	0.5
標準偏差	94.1	93.5	-0.5

日本の場合、重み付け前後の違いが 1 点未満でほとんど違いがないが、これは生徒の選ばれる確率がほとんど同じであること、調査の学校実施率が 96%、生徒実施率が 96%とデータの回収率が高く、欠損の補正が小さいことが原因として考えられる。逆に言うと、生徒の抽出確率が異なってくる場合や、回収率が低いことが予想される場合は、重み付けが重要になってくる。

(2)PISA 調査の標本誤差の算出方法

PISA 調査のような複雑な抽出方法を用いた際に利用できる標本誤差の算出方法には、大きく分けて線形化法と複製法の 2 つがあり、後者は主に、PISA 調査で用いられている均衡反復複製法 (BRR 法)、別の国際的な学力調査である TIMSS で用いられているジャックナイフ法、コンピュータの発展によって使われることが多くなってきたブートストラップ法がある (Lohr (2010) を参照)。線形化法は、直接計算できない標本分散を近似的に求める式で算出する方法である。中央値やパーセ

ンタイル、複雑な分析には使えないという欠点がある。複製法は、リサンプリング等の方法で調査データの複製を複数作り、それと調査データとの差から標本分散を算出する方法であり、複製の作り方によって先の 3 つの方法に分かれる。BRR 法は層内の 2 つの要素のうち、片方を取り出して複製を作るが、その組合せをアダマール行列に基づいて行い、すべての組合せを試した場合と同等の複製を作り出す。この方法は、各層が 2 つの要素 (例えば学科) からなる場合に利用できるが、PISA 調査では各層の学科が 2 つになることはないため、各層内の学科を 2 つの学科からなる複数の「疑似層」に分割している。ジャックナイフ法は要素を 1 つ除いた複製を要素の数だけ作り出し、ブートストラップ法は標本抽出法に合わせたリサンプリングを繰り返して複製を作り出す。上記の 4 つの方法のうち、BRR 法以外は、どのような抽出法にも対応できるため、これらの抽出法を行ったと仮定し、PISA2012 年調査の数学的リテラシーの平均得点 (PV の値を平均したものを得点として使う、PV については (3) で説明) の標本誤差を計算すると以下ようになる (R の survey パッケージを使用)。

	単純	二段抽出	層化二段
線形化法	1.16	5.00	4.83
ジャックナイフ法	1.16	5.00	4.84
ブートストラップ法	1.14	5.05	4.85

ブートストラップ法 (リサンプリング 1000 回) は若干異なっているが、計算結果はほぼ同じである。ただしこれらは、並び順といった系統抽出に伴う効果が含まれていない。

Rust (2013) は、このような系統抽出の並び順に、2、3 の要素からなる層化と同じ効果があるとして、並び順に従った「疑似層」の有効性を主張している。要素が 2 つではない層のある場合で BRR 法を使う方法として、Kish と Frankel (1970) は層を結合したり、無作為に 2 つに分割したりする方法を、Gupta と Nigam (1987) はアダマール行列とは異なる直交行列を用いる方法を提唱している。後者は、条件に合う行列を見つけ、計算に使用するのが難しいため、国際的な学力調査では用いられていない。前者に関しては、Wu (1991) によると、層を分割して作る疑似層の数を多くするほど、標本分散の予測精度が高くなるとされている。

PISA 調査の公開データでは、抽出時の学科の並び順や副層の情報が公開されておらず、「疑似層」も要素が 2 でない場合があるため (OECD (2014: 139) を参照) 「主層」と「疑似層」の情報をもとに要素 2 つからなる本研究用の「並び順に従った疑似層」(要素が 1 つになるときは分割) を作成し、数学的リテラシーの平均得点の標本誤差をこの「疑似層」に基づく層化二段抽出として、標本誤差を算出した結果が以下の表である (R の survey パッケージを使用)。

	標本誤差
線形化法	3.63
BRR 法	3.63
BRR 法(Fay 法)	3.63
ジャックナイフ法	3.63
ブートストラップ法(1000 回)	3.52

ブートストラップ法であっても、リサンプリングを 9000 回行った場合や 10000 回行った場合の結果は 3.30 になっており、これらの方法の間に違いはない。なお、「BRR 法(Fay 法)」とあるのは、BRR 法が各層内の要素を 1 つ用いて複製を作るのに対し、層内の要素の片方の重みを増やし、片方の重みを減らすこと(ここでは PISA 調査と同様、選ばれた要素の重み付けを 1.5 倍、選ばれなかった方を 0.5 倍)、サンプル内のすべての情報を用いて複製を作る方法である。

「並び順に従った疑似層」を使った場合、どの算出方法を用いても結果に違いは見られない。しかし、それを使うことで実際の標本誤差を捉えられているのか、明らかにした研究は見当たらないため、本研究ではシミュレーションを用いてその有効性を検証した。

実際の日本の状況に近づけるため、各学科の平均得点が平均 500 点、標準偏差 100 の正規分布に従い、生徒の得点が所属学科の平均、標準偏差 100 の正規分布に従う、7000 学科×150 人=1050000 人からなる有限母集団を作成し(母平均 500、母標準偏差 139)、そこから 1)二段抽出(復元抽出)、2)層化二段抽出(非復元抽出、層は学科を平均得点順に並べて 4 分割)、3)層化二段抽出(層は学科を平均得点順に並べて 4 分割、学科の平均得点順に系統抽出)、4)層化二段抽出(層は学科を平均得点順に並べて 2 要素になるように作成、BRR 法に最適な層化)で、サンプルサイズが 200 学科×35 人=7000 人のサンプルを 10000 作成し、それぞれのサンプルで疑似層を用いた BRR 法で平均得点の標本誤差を算出し、10000 回の平均と、各標本平均と母平均の差の 2 乗平均の平方根(実際の標本誤差)を抽出法毎に比較した。その結果は以下のとおりである。

	BRR 法による平均 標本誤差	母平均に基づく 標本誤差	差
2 段抽出(復元)	7.14	7.19	-0.05
層化 2 段抽出(非復元)	2.81	2.76	0.05
層化 2 段抽出(系統)	1.08	1.17	-0.09
層化 2 段抽出(各層 2)	1.07	1.08	-0.01

系統抽出を用いた場合、BRR 法は標本誤差を過小評価しているが、その差は 0.1 点未満であり、「並び順に従った疑似層」の使用は、実用上問題ないと考えられる。いずれにしる並び順の影響を直接計算する方法がないため、「疑似層」は非常に有効な手段といえる。

(3)Plausible Value の有効性と限界

PISA 調査のような国際的な能力調査では、

項目反応理論を用いて「生徒の能力」と「問題の難易度」を同一尺度で得点化している。正答率に基づく従来の方法では、生徒全員に同じ問題を出題しなければならないが、項目反応理論を使えば、一部が共通していれば、生徒ごとに違う問題を試すことができ、経年変化を見る場合も、一部共通問題があれば同一尺度で変化を測定することができる。

項目反応理論は、資格試験や検定試験で「個人の特性」を測るためにも使われているが、大規模教育調査では、「母集団の特性」を測るため、Plausible Value (PV) が用いられている。前者では「個人の特性」を最尤推定値(MLE: 最尤値)やベイズ推定値(EAP: 事後分布からの期待値)として求めるが、PV は周辺尤度事後分布からランダムに抽出された値であり、「個人の特性」を測るためには使えない。PV は、多重代入法の考えに則って、通常は 5 から 20 個の値を各生徒に割振り(PISA2012 年調査では 5 個の PV を使用)、各 PV を使って別々に特性値を計算し、その平均を「母集団の特性」とする(各 PV の標本分散の平均値に、PV の不変分散に $1+1/(PV \text{ 数})$ をかけた値を足したのが誤差分散であり、その平方根が標準誤差となる)。

PV が何故必要なのか、MLE や EAP を使って「母集団の特性値」を推定するとどうなるのかを明らかにするために、標準正規分布からサンプルサイズ 1000 で単純無作為抽出した値を能力とし、そこから問題数 15(困難度は -3 から 3 まで等間隔)の架空の解答データを作成して、項目反応理論のモデルの 1 つであるラッシュモデルを使って MLE、EAP、PV(ここでの PV は 1 個のみ)を求め、母平均と母分散を推定するシミュレーションを 1000 回行い(分布からの抽出には R の mvtnorm パッケージを、分析には R の TAM パッケージを使用、以下同じ)、その推定値の平均を求めた。

推定値(1000 回の平均)	MLE	EAP	PV
母平均(=0)	0.00	0.00	0.00
母分散(=1)	1.59	0.67	0.99

母平均の推定に関しては、どの推定値にも問題はないが、母分散に関しては、MLE が過大推定、EAP が過小推定になっている。Mislevy ら(1992)によれば、測定される生徒の能力はすべて「欠損値」であり、その推定には「誤差」が生じるため、MLE は測定誤差が分散に上乘せられ、EAP は信頼係数に比例して分散が小さくなる。そこで能力の推定値が「真の能力」と「誤差」(正規分布に従う)からなる数理モデル(これを「母集団モデル」と呼ぶ)を項目反応モデルと統合し、能力の推定値の代わりに周辺尤度事後分布を求め、そこから推定値(PV)を抽出することで誤差の影響をなくす方法が Mislevy(1991)によって提案され、国際的な能力調査では一般的な手法となっている。

ただし、PV や母集団モデルには限界が存在する。それを示すため、平均が異なる 2 つの

下位グループの例と、能力と別の変数との相関を見る場合のシミュレーションを示す。

平均が-0.1と0.1、分散が1の2つの正規分布からそれぞれサンプルサイズ 2000 で単純無作為抽出し、併せてサンプルサイズ 4000 の標本を作り、そこから問題数 20 (困難度は-3 から 3 まで等間隔)の架空の解答データを作成して、ラッシュモデルで2つの下位グループの平均を推定するミュレーションを1000 回行い、その平均を求めた。

平均の推定 (1000 回の平均)	MLE	EAP	PV	条件付け	
				EAP	PV
平均-0.1 の群	-0.10	-0.07	-0.07	-0.10	-0.10
平均 0.1 の群	0.10	0.07	0.07	0.10	0.10

EAP と PV は、母集団を正規分布と仮定しており、下位グループの平均の差を過小評価してしまうという限界が存在する。ただし、グループの違いをダミー変数とした回帰モデルを母集団モデルとすることで(これを「条件付け」と呼ぶ)、正しい平均値を求めることが可能になる (EAP、PV の条件付けの値)。

同様の問題は、能力と別の変数との相関を見る場合にも存在する。相関 (=0.4) がある 2 変量標準正規分布からサンプルサイズ 4000 で単純無作為抽出し、そのサンプルの 2 つある変数の一方から問題数 40 (困難度は-3 から 3 まで等間隔)の架空の解答データを作成し、ラッシュモデルで MLE、EAP、PV を求め、もう一方の変数との相関を推定するミュレーションを 1000 回行い、その平均を求めた。なお、架空の解答データの正答率ともう一方の変数との相関も併せて示す。

推定値 (1000 回の平均)	正答率	MLE	EAP	PV	条件付け	
					EAP	PV
相関 (=0.4)	0.37	0.37	0.37	0.34	0.43	0.40

相関を見たい変数を使って「条件付け」を行わないと、いずれの推定値でも過小推定になるが、これは生徒の正答率であっても解答が項目反応モデルに従っていれば同じ結果となっている。相関を正確に測定できたのは、「条件付け」を行った PV だけであった。つまり、能力調査で項目反応理論を用いる場合、能力との関係を調べたい情報は、得点を計算する際にモデルに予め組み込む必要がある。

(4) 多次元項目反応モデルの有効性

PISA 調査では、読解力、数学的リテラシー、科学的リテラシーという 3 領域の能力を主に測っているが、これらは多次元項目反応モデルを使って同時に算出されている。多次元項目反応モデルには、問題が複数の能力に関連する場合と 1 つの能力にしか関連しない場合があり、PISA 調査では後者が用いられている。

このモデルの有効性を見るため、相関 (=0.8) がある 2 変量標準正規分布からサンプルサイズ 4000 で単純無作為抽出し、2 つの変数からそれぞれ問題数 40 (困難度は-3 から 3 まで等間隔)の架空の解答データを作成し、ラッシュモデルで EAP、PV を別々に求めた場合と多次元項目反応モデルで一緒に求

めた場合とで、相関を推定するミュレーションを 1000 回行い、その平均を求めた。

推定値 (1000 回の平均)	別々		多次元	
	EAP	PV	EAP	PV
相関 (=0.8)	0.68	0.58	0.88	0.81

「条件付け」に関するシミュレーションと同様、別々に算出した場合、相関は過小推定になり、多次元モデルでは EAP が過大推定、PV がほぼ正確な値が求められることがわかる。この結果は、3 次元、つまり推定する能力が 3 つの時も同様であった。

(5) 母集団モデルと項目反応モデル

PISA 調査では、問題の特徴を「難易度」という 1 つのパラメータのみで表す、ラッシュモデル (1PL モデル) が用いられているが、この「難易度」とともに問題の「識別力」(値が高いほど能力と正答・誤答の関係が明確)を考慮した 2PL モデル、さらに「当て推量」(選択式問題で能力がどんなに低くても当たってしまう確率)を加味した 3PL モデルが存在する。(3)(4)では、1PL モデルのみを用いたが、解答データが他のモデルに従うとき、どの項目反応モデルが大規模教育調査に適しているのか、つまり母集団モデルと統合したときに「母集団の特性」をよく捉えているのかを検証する。

標準正規分布からサンプルサイズ 4000 で単純無作為抽出した値を能力とし、問題数 40 で困難度は-3から3までの等間隔は同じだが、識別力を加味しない 1PL モデル、識別力が「0.2、0.4、0.6、0.8、1.0、0.2.....1.0」である 2PL モデル、さらに「当て推量」が「0.2、0.0、0.2、0.0、0.2、0.2.....0.0」(2/3 が 5 択問題を意味する)の 3PL モデルに従った架空の解答データを作成して、1PL、2PL、3PL の項目反応モデルを使って PV (ここでの PV は 1 個のみ)を求め、母分散を推定するミュレーションをそれぞれのモデルに対して 1000 回行い、その推定値の平均を求めた。

母分散 (=1) の推定 (1000 回の平均)		分析モデル		
		1PL	2PL	3PL
解答	1PL モデルで作成	1.00	1.00	0.91
	2PL モデルで作成	0.72	1.00	0.80
	3PL モデルで作成	0.46	0.99	0.67

1PL モデルで作成した架空の解答データの場合、3 種類の項目反応モデルのどれを使っても大きな違いはない。それ以外の項目反応モデルを用いて作成した架空の解答データでは 2PL モデルでしか正しい母分散の推定ができていない。この結果が数理モデルの特徴からきているのか、それとも計算に用いた統計パッケージの設定によるのかはわからないが、いずれにしろ、大規模教育調査で母集団モデルを使う場合、2PL の項目反応モデルと組み合わせるのが無難であるといえる。

< 引用文献 >

Cochran, W.G., 1953, *Sampling*

techniques, second edition, New York: John Wiley and Sons.

Gupta, V. K., and A. K. Nigam, 1987, Mixed Orthogonal Arrays for Variance Estimation with Unequal Numbers of Primary Selections per Stratum, *Biometrika*, 74(4), pp.735-42.

Kish, L, and M. R. Frankel, 1970, Balanced Repeated Replications for Standard Errors, *Journal of the American Statistical Association*, 65(331), pp.1071-94.

国立教育政策研究所編, 2013, 『生きるための知識と技能5』, 明石書店.

Lohr, S. L., 2010, *Sampling: Design and Analysis, Second Edition*, Boston: BROOKS/COLE.

Mislevy, R. J., 1991, Randomization-Based Inference about Latent Variables from Complex Samples, *Psychometrika*, 56(2), pp.177-96.

Mislevy, R. J., A. E. Beaton, B. K. and K. M. Sheehan, 1992, Estimating Population Characteristics from Sparse Matrix Samples of Item Responses, *Journal of Educational Measurement*, 29(2), pp.133-161.

OECD, 2014, *PISA 2012 Technical Report*, Paris: OECD.

Rust, K., 2013, Sampling, Weighting, and Variance Estimation in International Large-Scale Assessments, in *Handbook of International Large-Scale Assessments*, Florida: CRC press, pp.117-53.

清水裕士, 2014, 『個人と集団のマルチレベル分析』, ナカニシヤ出版.

Wu, C. F. J., 1991, Balanced Repeated Replications Based on Mixed Orthogonal Arrays, *Biometrika*, 78(1), pp.181-8.

5. 主な発表論文等

〔雑誌論文〕(計2件)

篠原真子、「情報活用能力」を「測る」とは、内外教育、査読無、6570号、2017、pp.6-9。
篠原真子、高校生「情報活用能力」の特徴と課題、内外教育、査読無、6571号、2017、pp.12-4。

〔学会発表〕(計9件)

小泉力一、篠原真子、松本博幸、「高校生の「情報活用能力」を高める要因は何か？」、全日本教育工学研究協議会全国大会、2017年11月25日、和歌山県民文化会館(和歌山県)。

裏岩晶、篠原真子、「教育調査における多次元項目反応モデルと母集団モデルの有効性について」、日本行動計量学会大会、2017年8月31日、札幌学院大学(北海道)。
裏岩晶、「学力調査に適した項目反応理論

に基づく能力推定方法を考える」、日本行動計量学会大会、2017年8月30日、静岡県立大学(静岡県)。

篠原真子、「「習熟度レベル」によって高校生の情報活用能力をどう捉えたか」、日本行動計量学会大会、2017年8月30日、静岡県立大学(静岡県)。

篠原真子、松本博幸、小泉力一、「高校生の情報活用能力を測る」、教育システム情報学会全国大会、2017年8月24日、北九州国際会議場(福岡県)。

篠原真子、松本博幸、小泉力一、「高校生の情報活用能力に足りないものは何か?」、日本情報科教育学会全国大会、2017年7月1日、大阪芸術大学(大阪府)。

裏岩晶、篠原真子、「標本抽出法を用いた教育調査における標準誤差の算出について」、日本行動計量学会大会、2016年9月1日、札幌学院大学(北海道)。

裏岩晶、「大規模教育調査とエビデンスに基づく政策」、日本行動計量学会大会、2016年8月31日、札幌学院大学(北海道)。

裏岩晶、篠原真子、「大規模教育調査における標本抽出と重みづけについて」、日本行動計量学会大会、2015年9月4日、首都大学東京(東京都)。

6. 研究組織

(1)研究代表者

裏岩 晶 (HOROIWA, Akira)

国立教育政策研究所・国際研究・協力部・総括研究官

研究者番号: 00626210

(2)研究分担者

篠原 真子 (SHINOHARA Masako)

国立教育政策研究所・研究企画開発部・総括研究官

研究者番号: 30342611

(3)研究協力者

篠原 康正 (SHINOHARA Yasumasa)

独立行政法人大学改革支援・学位授与機構
学位システム研究会専門委員

研究者番号: 90727174

松本 博幸 (MATSUMOTO Hiroyuki)

国際大学 GLOCOM・客員研究員