参考資料3

UCHIDA

本体調査の個票データの匿名化に関する調査研究

平成30年6月27日

株式会社内田洋行　教育総合研究所

本研究の目的と内容

- 背景と目的
  - 全国学力・学習状況調査が行われるようになって10年が経過しており、**さらなる分析の充実を求める声**は大きい。
  - 現在、**本調査データを大学等の研究機関に条件つきで公開**し、研究者独自の視点に基づく調査研究を促進することで、教育施策や指導の改善に役立てるための検討が進んでいる。
  - そこで、本研究では、**学力調査データの匿名化**に関する調査研究を行う。
- 調査研究内容
  1. パブリックユースデータ（疑似データ）化に関する調査分析
  2. パブリックユースデータ（疑似データ）の作成【平成27年度分】
  3. 匿名データ化に関する調査分析
  4. 匿名データの作成【平成19年度～29年度分】
  5. 今後の可能性検討

個票データ・匿名データ・パブリックユースデータの位置づけ

- **個票データ**：学校名も含む、全ての情報が含まれるデータの中から、申出される研究に必要最小限の範囲のデータに限って、有識者会議による審査の上、貸与（個々の児童生徒の解答用紙番号は貸与しない）。国が公表していない教育委員会名・学校名が明らかになるデータを貸与する場合及び公表する場合には、貸与前は文部科学省が、公表前は申出者が当該学校の設置管理者の同意を得る。
- **匿名データ：都道府県名を含む地域情報や、一定水準以下の小規模校に関するデータを削除するなどの匿名化を行った**上で、全国の児童生徒から一定割合、無作為に抽出されたデータ。ガイドラインに基づいた利用を行う場合に、児童生徒個人、学校、設置管理者を特定することは困難。有識者会議による、より簡易な審査の上、貸与。
- **パブリックユースデータ**：調査結果の統計的性質を一部保存した上で、集計表の統計量から乱数を発生させて作成した疑似データをホームページ上に公表。特定の児童生徒個人、学校、設置管理者を示すものではない。データは教育目的等のため試行的に、個別情報の秘匿を気にすることなく自由に利用できるが、**導かれた分析結果は実証研究の結果とみなすことはできない**。

	抽出規模（想定）	地域情報（教育委員会名、学校名）	解答状況（教科）	回答状況（児童生徒質問紙）	回答状況（学校質問紙）
①個票データ 申出により貸与するデータが異なる。	貸与申出されたデータ	○	○	○	○
②匿名データ	無作為抽出（一定割合）	×	○	○	○
③パブリックユースデータ ホームページで公表	疑似データ化されたもの	×	○	○	○

本調査研究ではこちらについて検討・作成

※全国学力・学習状況調査個票データ等の貸与・公表について
http://www.mext.go.jp/a_menu/shotou/gakuryoku-chousa/sonota/__icsFiles/afieldfile/2017/06/12/1386492_002.pdf

各種データの利用目的

目　　的	申出者の範囲	貸与するデータの種類
１．学術研究の発展に資するもの	（１） 国が公募により補助する調査研究の代表者 （２） 国の委託調査研究又は共同研究の代表者 （３） 次のいずれかの機関に所属する研究者 ①国の行政機関 ②調査に参加する学校の設置管理者 ③都道府県教育委員会 ④独立行政法人 ⑤地方独立行政法人 ⑥大学及び高等専門学校 ⑦大学共同利用機関 ⑧その他科学研究費補助金取扱規程第２条第１項第４号に規定する研究機関（同条第８項の規定により研究機関とみなされるものを含む。）	個票データ 匿名データ
２．公的機関における施策の推進に適切に反映されるもの	次のいずれかの機関に所属する常勤の役員又は職員 ①国の行政機関 ②都道府県教育委員会 ③市町村教育委員会 ④独立行政法人 ⑤地方独立行政法人	個票データ 匿名データ
３．大学院生の教育目的利用等の高等教育の発展に資するもの	次のいずれかの機関の教育責任者（教員） ①大学及び高等専門学校 ②大学共同利用機関	匿名データ

※個票データ等の貸与の体系
http://www.mext.go.jp/a_menu/shotou/gakuryoku-chousa/sonota/__icsFiles/afieldfile/2017/06/12/1386492_004.pdf

パブリックユースデータ

成果物（パブリックユースデータ）

- H27匿名データをベースに、共分散行列からデータを生成したもので、文部科学省ウェブサイトからダウンロード可能
- データ件数
  - 小学校：児童2000件、学校100件　中学校：生徒2000件、学校100件
- 教科調査については国語・算数／数学・理科の全問題の調査結果を収録
- 質問紙については以下の質問を収録
  - （児童生徒質問紙）
    01朝食を毎日食べている
    02毎日同じくらいの時刻に寝ている
    03毎日同じくらいの時刻に起きている
    04ものごとを最後までやりとげてうれしかったことがある
    05難しいことでも失敗を恐れず挑戦している
  - （学校質問紙）
    16児童生徒は熱意をもって勉強している
    17児童生徒は授業中の私語が少なく落ち着いている
    18児童生徒は礼儀正しい
    19児童生徒は学級やグループでの話合いなどの活動で自分の考えを相手にしっかりと伝えることができている
    20児童生徒は学級やグループでの話合いなどの活動で相手の考えを最後まで聞くことができている

※パブリックユースデータ（疑似データ）
http://www.mext.go.jp/a_menu/shotou/gakuryoku-chousa/sonota/1404609.htm

パブリックユースデータ

匿名データ

パブリックユースデータのデータレイアウト(匿名データも同様)

児童生徒データ

項目	データ内容
実施年	※西暦にて表記、数字4桁
地域規模	1: 大都市 2: 中核市 3: その他の市 4: 町村
解答用紙番号	※個票データの解答用紙番号とは異なる番号を連番で付与する
性別	1: 男子 2: 女子 0: 不明
児童生徒質問紙種別	1: 児童生徒質問紙 I 2: 児童生徒質問紙 II 3: 児童生徒質問紙 III ※平成25年度のみ、児童生徒質問紙が分冊で行われたため、当該学校がどの児童生徒質問紙を実施したかを示すもの。以降年度は空欄固定
正答数_(教科)	
正答率_(教科)	
学力層_(教科)	1: A層 2: B層 3: C層 4: D層 ※全国の正答数ヒストグラムを4分割し、正答数が高いほうからA層～D層としたときに当該児童生徒がどの層に当てはまるかを示す

項目	データ内容
	もの。
正答数_(教科)_(領域・観点・形式)	
類型_(設問)	1: 類型1 2: 類型2 3: 類型3 4: 類型4 5: 類型5 6: 類型6 7: 類型7 8: 類型8 9: 類型9(上記以外の解答) 0: 類型0(無解答)
正誤_(設問)	1: 正答 2: 誤答 0: 無解答
正答率_(教科)_(領域・観点・形式)	
児童生徒質問紙回答	1: 選択肢1 2: 選択肢2 3: 選択肢3 4: 選択肢4 5: 選択肢5 6: 選択肢6 7: 選択肢7 8: 選択肢8 9: その他(選択肢以外の回答や複数回答) 0: 無回答
学校質問紙回答	1: 選択肢1 2: 選択肢2

パブリックユースデータ

匿名データ

項目	データ内容
	3: 選択肢 3 4: 選択肢 4 5: 選択肢 5 6: 選択肢 6 7: 選択肢 7 8: 選択肢 8 9: 選択肢 9 0: その他（選択肢以外の回答や複数回答）・無回答 ※児童生徒が所属する学校の学校質問紙回答を児童生徒 1 行に紐づいて示すもの。

学校データ

項目	データ内容
実施年	※西暦にて表記、数字 4 桁
地域規模	1: 大都市 2: 中核市 3: その他の市 4: 町村
児童生徒質問紙種別	1: 児童生徒質問紙 I 2: 児童生徒質問紙 II 3: 児童生徒質問紙 III ※平成 25 年度のみ、児童生徒質問紙が分冊で行われたため、当該学校がどの児童生徒質問紙を実施したかを示すもの。以降年度は空欄固定
正答数_（教科）	※平均正答数
正答率_（教科）	※平均正答率（整数値に四捨五入）

項目	データ内容
第 1 四分位_（教科） 第 2 四分位_（教科） 第 3 四分位_（教科） 第 4 四分位_（教科）	※教科の正答数を基準に、学校内を A 層～D 層の 4 層に分割したときに区分点となる正答数の値。
最頻値_（教科）	※学校内の正答数の最頻値を示すもの。
標準偏差_（教科）	※平均正答数の標準偏差
正答数_（教科）_（領域・観点・形式）	※領域・観点・形式別平均正答数
正答率_（教科）_（領域・観点・形式）	※領域・観点・形式別平均正答率（整数値に四捨五入）
無解答数_（教科）	※平均無解答数
無解答率_（教科）	※平均無解答率
無解答数_（教科）_記述式	※記述式のみの無解答数を示すもの。
無解答率_（教科）_記述式	※記述式のみの無解答率を示すもの。
割合_（教科）_A 層 割合_（教科）_B 層 割合_（教科）_C 層 割合_（教科）_D 層	※全国の正答数ヒストグラムをもとに算出した A 層～D 層（学校の A 層～D 層とは異なる）に、当該学校の児童生徒が当てはまる割合を示すもの。
児童生徒質問紙回答割合_（質問番号）	※児童生徒質問紙の当該学校の肯定的回答割合を示すもの。
児童生徒質問紙回答割合_当日実施のみ_（質問番号）_（選択肢番号）	※児童生徒質問紙の当該学校の質問ごと・選択肢ごとの回答割合を示すもの。
学校質問紙回答_（質問番号）	
結果チャート_全国_学校質問紙_（領域） 結果チャート_全国_児童生徒質問紙_（領域）	※全国基準の「結果チャート」の得点を示すもの。
結果チャート_県_学校質問紙_（領域） 結果チャート_県_児童生徒質問紙_（領域）	※都道府県基準の「結果チャート」の特典を示すもの。

匿名データ

成果物（匿名データ）

各年度のデータ件数

		H19	H20	H21	H22	H24	H25	H26	H27	H28	H29
小学校	児童	109,899	109,104	108,151	25,134	24,320	105,426	102,848	101,341	97,352	100,450
	学校	2,060	2,040	2,017	517	489	1,918	1,888	1,871	1,814	1,912
中学校	生徒	96,674	97,354	97,656	39,535	40,171	96,847	95,896	95,859	93,730	97,493
	学校	928	938	909	406	403	892	895	888	869	921

※H22, H24は抽出調査（抽出率3割程度）のため、他年度よりも件数が少ない

匿名化の対象とするレコード

- 公立学校
  - 国立学校・私立学校は除く
- 当日実施の児童生徒
  - 調査実施日に調査を受けた児童生徒のみ
- H28までの調査結果について「匿名データ公開可」としている市町村教育委員会に属する児童生徒
  - H29からは匿名データの研究利用について実施要領に記載

匿名データ

サンプリング率
（児童生徒データ・学校データ）

- 【結論】文献および有識者委員会の議論を踏まえ、
  サンプリング率は　10%　とした
  - 平均正答率の信頼区間を一定区間に収めるという意味では、より小さなサンプリング率でも十分な値が得られる
  - しかし、本調査で扱われる変数は平均正答率のほかにも設問別の解答類型や児童生徒質問紙回答までさまざま
  - 多様な研究ニーズを保証するという意味で、匿名性が保たれる範囲で可能な限りデータ件数を多く確保することも重要

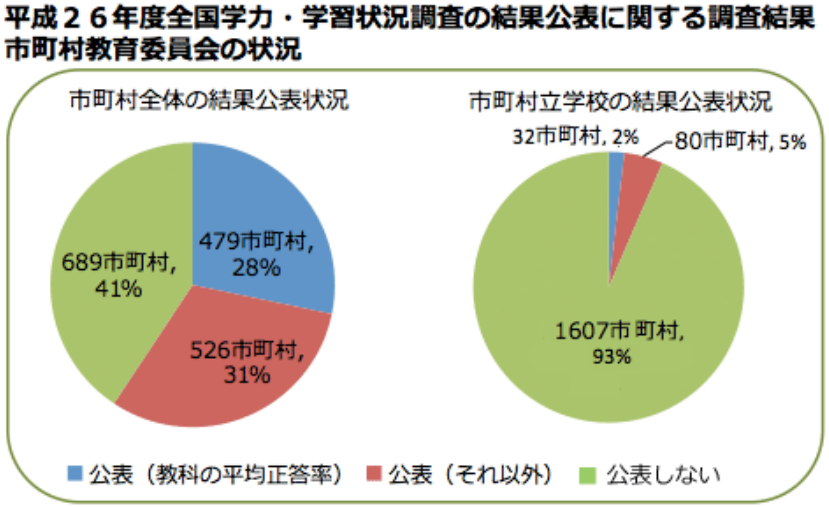
※参考：
永田靖（2003）統計ライブラリー　サンプルサイズの決め方．朝倉書店
山口拓洋（2010）臨床家のための臨床研究デザイン塾テキスト中級編2　サンプルサイズの設計．健康医療評価研究機構

匿名データ

平均正答率の丸め処理（学校データ）

- 「教科の平均正答率」を公表している学校・教育委員会が一部存在するため、ある教科の平均正答率ごとに学校数が何校あるか（特定されることがないか）を検討した

- 【結論】データの可用性を維持しつつ、学校の特定化を防ぐため、学校データの平均正答率を**整数値に丸める**こととした
  - たとえば、平均正答率が70%の学校はほかにもあるため、特定の学校であると言い切れなくなる
  - 平均正答数については、毎年教科の設問数が変わり解釈が異なることもあり、平均正答率と比べて公表されていないことが多いため、丸め処理は行わない

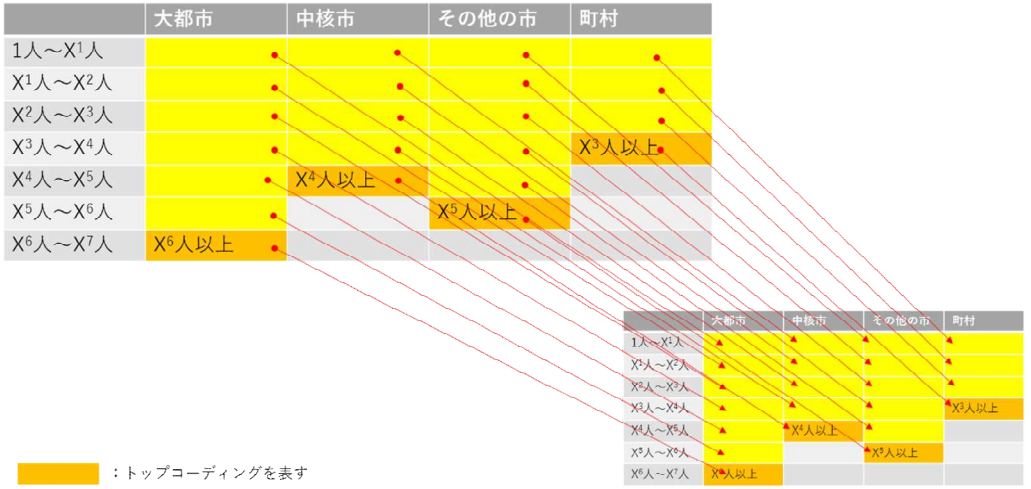


匿名データ

学校規模・地域規模による サンプリング（学校データ）

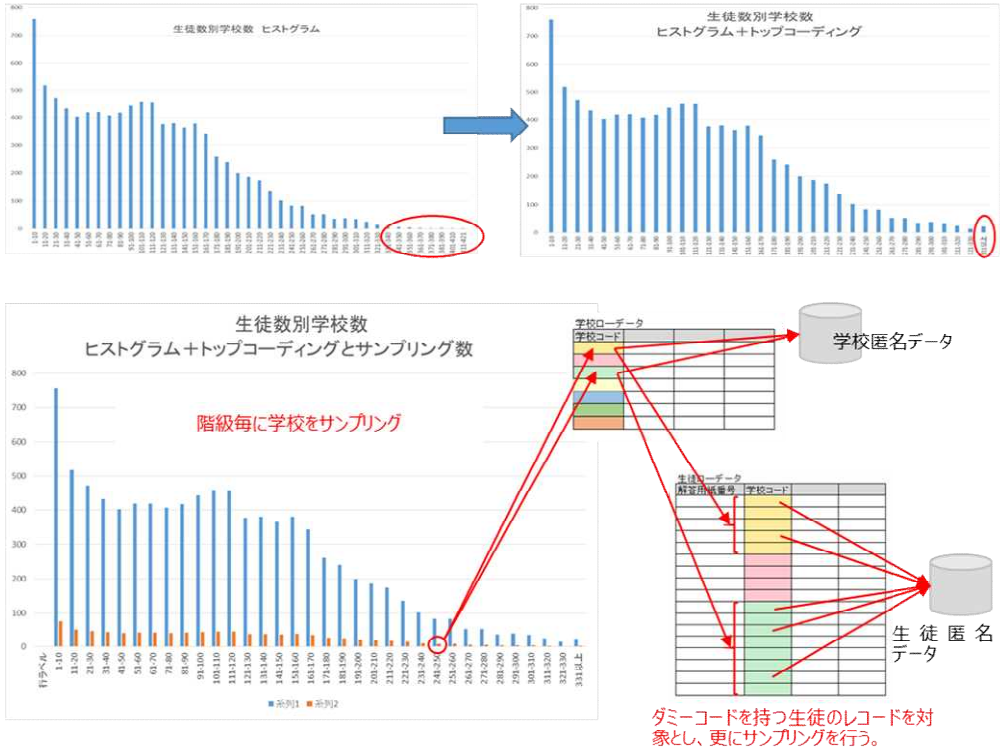
- 中学校では大規模校・小規模校が削れることで生徒数が2割弱減少するため、学校規模による単純なレコード削除は行うべきでない

- **【結論】学校規模（児童生徒数）および地域規模で学校を層化**し、各階級に一定数の学校が存在することを確認した上で一定割合をサンプリングした
  - 地域規模ごとにクロス集計表のセル内の学校数があらかじめ決めた基準に満たなくなったところでトップコーディング（たとえば「201人以上」など）を行った
  - 上記基準については全年度共通とした



今後の課題①

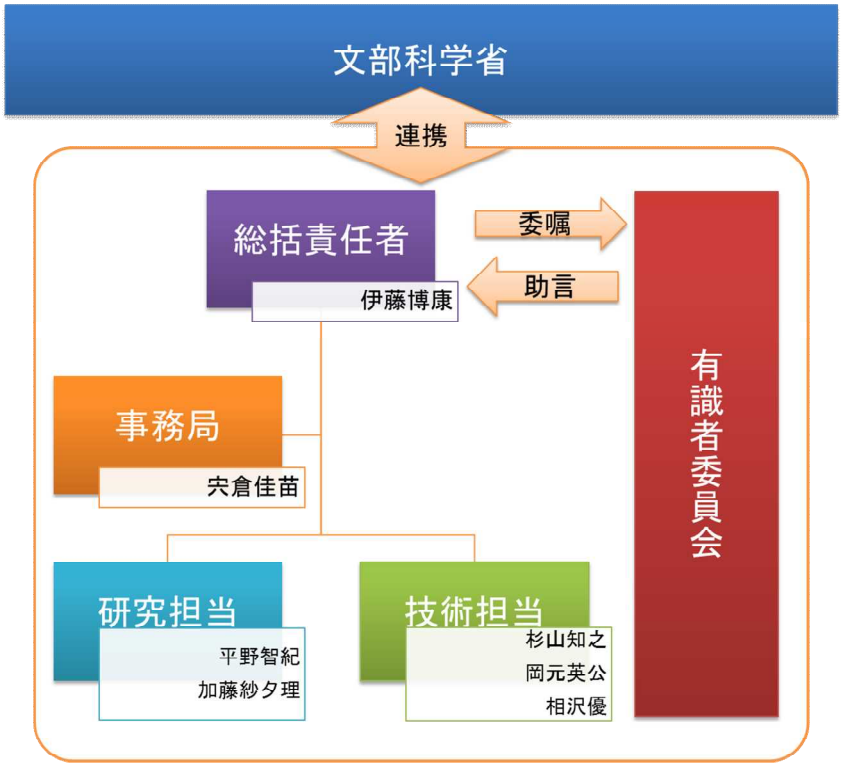
- 学校と児童生徒の紐付けは以下の理由から見送り
  - 匿名化の度合いが下がること
  - 児童生徒データで求めた学校平均が学校データとずれること
  - 学校内の児童生徒を二段抽出する場合、研究テーマによっては分析しづらいこと
  - 分析に資するデータ件数を確保したほうがよいこと
- **学校と児童生徒を紐づけた匿名データも検討可能**
  - 学校規模（児童生徒数）および地域規模に合わせて層化し、階級ごとに一定割合の学校を抽出する
  - 抽出した学校内から児童生徒を抽出（二段抽出）する
  - ダミーの学校コードを振り、同一学校の児童生徒を判別できるようにする



今後の課題②

- 児童生徒データ・学校データの正規化とデータベース化
  - RDBMSの考え方に基づき、マスターテーブルとデータテーブルを分割
  - 文字コード等の統一
  - クラウド上での分析環境の提供（cf: PISA調査）
- バリエーションのあるパブリックユースデータの作成
  - たとえば「読書」に関する質問項目に特化したパブリックユースデータ等、より分析に資する形でのデータ生成・公開

実施体制



有識者委員会（五十音順・敬称略）

氏名	所属・役職
伊藤　伸介	中央大学　経済学部 教授
後藤　康志	新潟大学　教育・学生支援機構 准教授
福田　幸男	横浜薬科大学　薬学部 教授
南　和宏	統計数理研究所　モデリング研究系 准教授
村山　功　※主査	静岡大学大学院　教育学研究科 教授