# サーベイ実験の再現可能性と外的妥当性 : オンラインフィールド実験による追検証

| | |
|---|---|
| その他のタイトル | Replicability and External Validity of a Survey Experiment for Charitable Giving : A Replication Study through an Online-based Field Experiment in Japan |
| 著者 | 善教 将大, 坂本 治也 |
| 雑誌名 | ノモス = Nomos |
| 巻 | 46 |
| ページ | 1-15 |
| 発行年 | 2020-06-30 |
| URL | http://hdl.handle.net/10112/00020425 |

〔論　説〕

# サーベイ実験の再現可能性と外的妥当性

## ── オンラインフィールド実験による追検証 ──

善教将大・坂本治也

## 1　問題設定

### （1）再現・追試研究の重要性

政策展開に「根拠」を用いるべきという気運が、近年の日本では、国や地方を問わず広まりつつある。印象や直感に頼るのではなく、根拠に基づき政策形成すべき（evidence-based policy making; EBPM）という意見に対してはいくらかの批判が存在するものの[1)]、この取り組みそれ自体は決して否定されるべきものではない。EBPMはより効率的ないし効果的な行政運営を可能にする手法であると同時に、施策を見直す機会を与えるものでもあるからである。

根拠としての信頼性がもっとも高いとされるのはランダム化比較試験（randomized controlled trial; RCT）、あるいは複数のRCTの結果を統合したメタ分析の結果である[2)]。もちろん、観察調査・研究（observational studies）により得られた知見や、インタビューなど質的調査から得られた知見も証拠としての価値は高い。しかし、政策の効果という点では、実験の結果がもっとも信頼性と妥当性が高いと考えられる。実験によって推定された平均処置効果（average treatment effect; ATE）は、内生性や交絡要因の影響を除去した上での効果量、すなわち因果効果を明らかにするものだからである。特にRCTのように、現実の社会を題材に処置の平均的な効果を推定する実験の結果は、妥当性の高い根拠だとみなされる傾向にある（伊藤 2017）。

しかし、たとえRCTであったとしても、1つの分析結果から普遍的な結論を導き出すことには慎重にならなければならないこともある。他の多くの手法と同様に、実験にも多くの問題や限界はある。仮にRCTを実施した結果、処置の因果効果が有意だったとしても、それが微細だった場合は、偶然有意な結果となった可能性が疑われる（第一種過誤）。逆に統計的に有意ではないという結果が示されたとしても、別の実証研究を積み重ねる中で、その解釈が覆されることもあろう。いずれにせよ発見の独自性や新規性、あるいは重要性が高い場合などは、先行研究と同一あるいは類似の実験を繰り返しながら、実験結果の信頼性を確認していかなければならない。

実験結果の再現性（replicability）を確認する研究は、科学研究の信頼性を担保する点で極めて

1）EBPMは、政策形成者の利益などに合致する形で根拠が形成されてしまう側面もある。それゆえにPBEM（policy-based evidence making）と揶揄する声もある。

2）ここでは主として行動経済学研究などに見られる、現実の社会で実験を行うフィールド実験（field experiment）研究をRCTと呼んでいる。

重要であるが、日本の社会科学領域でその重要性が認識されているとはいえない。心理学では、過去に公刊された論文の多くが再現不可能であることが発表されたことにより（Open Science Collaboration 2015）、追試や再現可能性への問題意識が急速に広まっている[3)]。これとは対照的に心理学を除く社会科学の領域では、再現研究に対する取り組みはほとんど広がっていない。これほどまでに再現性の危機が議論されているにもかかわらず、再現研究が広がらない理由として「多くの研究者が研究のオリジナリティを求められる中で、他人の研究のコピーまでする時間がないため」などという、再現研究に対する偏見を開陳する研究者が、日本には依然として存在する[4)]。

以上の問題意識に基づき、本稿では日本人の寄付行動の規定要因を明らかにした実験研究の追試に取り組む。実験結果の再現性もさることながら（King 1995）、HARKing など誤った研究慣行への問題意識の高まりに鑑みれば（Laitin and Reich 2017; 友永・三浦・針生 2016）、既存の知見の信頼性を確認する研究は、今後重要性を増すことが予測される[5)]。

研究の信頼性を確認する研究は、再分析研究（reanalysis studies）と再現研究（replication studies）の２つに大別される。前者は、主として先行研究と同じデータないし対象を用いつつもさらに洗練された方法で再度分析し、知見の頑健性を検証するものである。これに対して再現研究は、基本的には実験的手法を用いている先行研究の知見の信頼性や妥当性を同様の手法で検証することを目的とする[6)]。本稿は後者の再現研究の１つとして位置づけられる。

追試・再現の対象として本稿が取り上げるのは、善教・坂本（2017）の実験結果である。この実験の概念的追試（conceptual replication）を行うことが、本稿の具体的な課題となる。概念的追試とは、先行研究と全く同じではないが同様の知見を得ることができると期待されるような実験を新たに行い、先行研究の知見の信頼性等の検証を行うものである。これに対して、先行研究と（ほぼ）同一の実験を行うものを直接的追試（direct replication）という。善教・坂本（2017）の実験設計は、後述するように寄付行動の規定要因を十分に外的妥当性（external validity）の高い形で明らかにするものではない。したがって本稿では実験設計をそのまま踏襲するのではなく、

---

3）たとえば『心理学評論』59巻１号（2016年刊行）では「心理学の再現可能性」に関する特集が組まれ、多くの心理学者がこの問題に関する論稿を発表している。日本パーソナリティ心理学会が刊行する『パーソナリティ研究』でも、近年、事前登録制の再現研究の投稿受付を開始した。筆者らの専門とする政治学でも、*Journal of Experimental Political Science* や *Japanese Journal of Political Science* では追試や再現研究の専門セクションを設けている。

4）本稿をある学会誌に投稿した時の査読者のコメントをそのまま引用したものであるため鍵括弧をつけた。「他人の研究のコピー」という表現は、まさに再現研究に対する偏見を濃縮したものだといえるのではないだろうか。

5）HARKing とは hypothesizing after the results are known の略称で、分析結果を見てから仮説を設定する行為を指す。データに対して様々な分析を繰り返し行いながら統計的に有意な結果を見つけ出し、後付け的に「仮説」が検証されたという論文を執筆することは、データのねつ造や改ざんとは異なるけれども不正な研究慣行とされる。このような論文は実際には信頼性の低い結果を示すものになってしまうが、実際にどのような過程で当該論文が執筆されたかを第三者が判断することはほぼ不可能といってよい。再現研究の重要性は、このような研究慣行を抑止する観点からも指摘できる。

6）再分析研究と再現研究の区別は *Japanese Journal of Political Science* の Reanalysis と Replication の説明を参考にした。

外的妥当性を高める方向へと改善した新たな実験によって再現性を検証する。

一般に著者自身、あるいは著者を含むグループによる直接的追試は、出版バイアス（publication bias）などに対する配慮もあって推奨されない傾向にある[7)]。しかし、概念的追試に関しては先行研究をそのまま踏襲するのではなく、実験設計を改善するなど新規性が認められるものであることから、著者自身による追試を妨げないとするのが一般的である。ただし近年、著者以外の研究者集団による直接的追試には問題があり、それが再現の正否にも大いに関わることから、事前登録（pre-registration）を活用した自己追試の方法についての検討も進められている（Shah et al. 2019）。日本の再現研究に対する資源や理解の乏しさを勘案すれば、第三者による追検証を待つのではなく[8)]、自らが自身の実験結果を問い直す自己追試・再現をいかに行うかという視点も重要であろう。

### （2）何を再現・追試すべきか

前項で述べたように、本稿が追試する対象として取り上げるのは、サーベイ実験（survey experiment）の１つである無作為化要因配置実験（randomized factorial survey experiment; RFSE）によって寄付行動の規定要因を明らかにした善教・坂本（2017）である。一般に、アンケート調査で回答者を無作為に複数の群に配分し、それぞれに異なる質問文や選択肢を提示することで質問文の違いなどの効果を推定する手法をサーベイ実験という。善教・坂本（2017）のRFSEは、調査対象者ではなく調査画面ないしヴィネットに含まれる情報が無作為に変化するサーベイ実験となっており、その情報の変化が仮想的な寄付行動（寄付金額）に与える因果効果を推定している。

---

7）著者ないし著者を含むグループが直接的追試を行なった場合、不都合な結果（たとえばnull results）が出るとそれを公表せず、成功例だけが報告されがちとなる。ゆえに著者自身による自己追試は一般的に推奨されないとされている。もっとも、これはすべての自己追試が不正行為だと指摘するものではない。出版バイアスを生じさせうる可能性があるから自己追試は推奨されるべきではないというだけであり、その可能性がないと判断されるのであれば、Shah et al.（2019）のように、自己再現研究は正当化されうる余地はあるというのが、筆者らの理解である。それゆえに、仮に自己再現を行う研究者や研究集団に対して、あたかも不正行為を行なっているかのような指摘が根拠なくなされた場合、それは研究者としての信頼性や名誉を不当に毀損するものだと判断される可能性がある。

8）日本には「第三者の追検証を経ていない場合（概念的追試含む）自己追試や再現研究をすべきではない」といった見解を（暗に）主張する研究者や学術団体がある。たしかに直接的追試に関しては自己追試だと肯定的な結果が報告される場合が多く、そのため慎重に結果を吟味する必要がある。脚注３であげた『パーソナリティ研究』でも直接的追試に関しては、著者を含むグループなどによる投稿を認めていない。しかし概念的追試に関してはそのような制約は課されていない。また、再現性に対する関心が近年高まりつつある教育学の指針では、そもそも自己追試そのものが禁止されていない。例えばアメリカのInstitute of Education SciencesとNational Science Foundationが共同で2018年に発行した再現研究のガイドライン（*Companion Guidelines on Replication & Reproducibility in Education Research*）を見ると、十分に客観性を保証するための方策を含めるという条件付きで自己追試を認めている。再分析であれ再現であれ、第三者による追試の存在が自己追試を行う必要条件だという主張はいかなる論拠や証拠により正当化されるのだろうか。著者らはこの点を理解することができない。

本稿が善教・坂本（2017）を取り上げる理由は、大きくは2点ある。第1は、この実験研究はこれまでの寄付行動の規定要因に関する先行研究とは異なる知見を多く提示する一方、一度きりの実験に基づくものであるため、結果の信頼性が不明瞭だからである。たまたま異なる結果となったのか否かは、先行研究と善教・坂本（2017）の結果が食い違っている以上、重要な検討課題となる。第2は善教・坂本（2017）の実験は外的妥当性に乏しいからである。これはサーベイ実験全般に通ずる課題でもあるが、一般にサーベイ実験は仮想的な状況を設定した上で処置の効果を推定する。ゆえに現実社会に知見を適用可能なのかどうかが問題となる。

サーベイ実験に付随する妥当性の問題について詳しく説明する。実験研究の妥当性という時、通常、そこには異なる2つの意味が含まれている。1つは無作為配分に関する妥当性である。これは通常内的妥当性（internal validity）と呼ばれる。被験者ないし実験参加者が統制群と処置群に無作為に配分されているかどうかは、ATEを推定するにあたり重要となる。回帰非連続デザインや差分の差法といった疑似実験（quasi-experiment）は、人為的に無作為配分されていないので内的妥当性を疑問視される場合がある。実験室実験やサーベイ実験の場合は無作為配分が人為的に行われるため、この点について批判されることは稀である[9)]。したがってもう1つの外的妥当性が、本稿にとっては重要となる。

外的妥当性とは、一般的には「結果、処置、セッティング、対象者の特性を超えて因果関係が維持されるかどうか」に関わる問題をいう（Shadish et al. 2002: 21）。つまり、同じ実験を異なる国や地域の人を対象に実施しても同一の結果になるかという、分析結果ないし知見の一般化可能性（generalizability）に関わる概念である。さらにサーベイ実験に関する議論の文脈では、現実社会への知見の適用可能性やサーベイ実験の処置と現実社会の処置の一致性も外的妥当性の問題として議論される（Barabas and Jerit 2010; Hainmueller et al. 2015）。RCTや疑似実験と比べると、外部環境を統制する実験室実験や仮想状況を設定するサーベイ実験の結果は外的妥当性に問題がある。

寄付行動の先行研究は、実際にお金をどの程度寄付する、あるいはしてきたかを分析し明らかにするものが多く（Alpizar et al. 2008; Della Vigna et al. 2012; Ishida and Okuyama 2015; Jackson 2016; James and Sharpe 2007）、これらと比較すると善教・坂本（2017）のサーベイ実験の結果には現実社会への適用可能性について難があると言わざるを得ない。この実験は、あくまで仮想的なヴィネットを用いて実験参加者の寄付「意向」の規定要因を分析しているにすぎず、実際の寄付行動が分析対象となっていない。信頼性にくわえて（外的）妥当性という点でも善教・坂本（2017）はいくらかの限界を抱える。

ゆえに本稿が行う善教・坂本（2017）の結果の追検証は、実験結果の信頼性（reliability）の検証のみならず、サーベイ実験の外的妥当性を検証することにも繋がる。上述したように、サーベイ実験の結果に対しては、それらの結果は現実社会に適用可能なのかという批判が常につきまと

9）十分な数の実験参加者を確保できていない場合、たとえ無作為配分していたとしても処置群と統制群間の均質性が満たされない場合がある。サンプルサイズが小さい場合は、ブロッキングなどによって重要な共変量を統制しておくべきであろう。

う。本稿はサーベイ実験結果の追検証を行うものであるが、その作業を通じて、サーベイ実験の結果には外的妥当性があるのかという疑問にもこたえるものである。

## 2　実験設計

### （1）オンラインフィールド実験

筆者らは善教・坂本（2017）の実験結果の再現可能性と外的妥当性を検証するために、実際にある団体に寄付をしてもらうフィールド実験を、オンライン上で実施した[10)]。オンラインフィールド実験とは「政府、組織、あるいは諸個人の行動や意識の傾向を研究することを目的に、インターネット上に存在するシステムやプラットフォームを活用する」フィールド実験とされる（Muise and Pan 2019: 218）。無作為に実験参加者を統制群と処置群に配分した上で、処置群に投票動員に関するメールを送信し、それが投票率に与える影響を分析する研究などがそれに該当する。サーベイ実験とオンラインフィールド実験は、ともにオンライン上で処置を与える点は共通する。しかし前者が仮想的な状況を設定した上で「行動意欲」などの意識の変動を分析対象とする一方、後者は実際に何らかの処置を講じたり、実際の行動に対する影響を推定したりすることを目的とする点で異なる。

本稿の実験は、後述するようにディセプションを行うことで、実験参加者は「実際に存在する団体ビラを確認した上で、100ポイントのうち、いくら寄付するか」を決定するものとなっている。その意味で実験参加者の主観レベルではオンラインフィールド実験となるが、他方で実際に何かしらの団体に寄付するわけではなく、外形的にはオンラインフィールド実験よりもオンラインラボ実験に近似する（Parigi et al. 2017）。ただし本稿では実験参加者が実際に寄付することを想定しながら意思決定したものと考えている。この点を重視し、本稿の実験もオンラインフィールド実験の１つとして位置づけている。

具体的に実験の手順を説明する。まず、実験実施前に以下の事項について実験参加者に伝えた。１）調査協力の謝礼とは別に100ポイント（100円相当）の追加謝礼を支払う。２）この100ポイントは「実際に熊本で活動する団体」に対する寄付のために追加的に付与するものである。３）当該団体の活動チラシを見ながら、実際に寄付をするか否かを決める。４）寄付する場合、当該団体の活動資金として実際に調査会社を経由して寄付が行われ、その分、実験参加者が受け取るポイントは減る。たとえば50ポイントを寄付した場合、50ポイントしか追加謝礼を受け取れなくなる。５）寄付は強制されるものではなく、寄付したくない場合は寄付しなくてもよい。筆者らは上述した５点を実験実施前に説明し、これらの条件に対して同意した人にのみ実験に協力してもらった。

次に、筆者らは実験参加者に対して「実際に熊本で活動する団体」のチラシを表示し、当該団

10）詳細はすべて本稿のOnline Appendix（https://zkun.sakura.ne.jp/document/Zenkyo_Sakamoto_2020.pdf）で説明しているのでそちらを参照のこと。

体に対して謝礼ポイントのうちいくら寄付するかを尋ねた。このチラシは善教・坂本（2017）の実験設計に基づき作成した架空のチラシであるが、実験参加者には「実際に熊本で活動する団体」のチラシだと説明しているため、実験参加者は架空のチラシだとこの時点では認識していない。なおこのチラシは、後述するように掲載情報のうち主体特性、管理運営費、返礼の３要因について、その水準表記が無作為に変化するチラシとなっている。

本稿の実験では、架空のチラシという点と実際に寄付したポイントが当該団体に与えられるという２点において、実際とは異なる偽情報を実験参加者に示している。必要な情報を実験参加者に隠したり、偽情報を提示したりすることを一般にディセプションという。ディセプションは心理学などの実験研究で用いられる手法の１つであり、研究目的上必要であることや、後述するデブリーフィングやコンセントフォームへの同意などを条件に許容される場合がある。本稿のディセプションは、何かしらの対象を誹謗・中傷したり、実在する団体名を出して寄付を募ったりするものではない。ディセプションにより特定の人や団体が不利益を被る可能性は小さく、先行実験の外的妥当性を高めるという本稿の研究目的からも必要不可欠な措置だと考える[11)]。

しかしディセプションを行う以上、実験参加者が何らかの不利益を被らないように最大限配慮する必要がある。そのため本稿では実験実施後に、偽情報を与えた上での調査であったことを説明するデブリーフィングを行った。デブリーフィングでは、偽情報を与えたことについて謝罪した上で、当該団体は存在しないことや、実際には調査会社を経由する寄付は行われないことなどを説明した。また、追加した謝礼ポイントが実際には受け取れないとなると参加者が不満を募らせ、学術研究に不信感を抱いてしまう可能性があることから、回答した寄付ポイント額の大小にかかわらず、実験参加者全員に100ポイントの追加謝礼を支払った。さらに実験に疑念や不満を抱いた人は、いつでも調査から離脱可能であることを事前に同意書の中で説明しており、実験参加者はそのことに同意の上で参加している。

### （２）推定結果の補正と質の向上について

筆者らは、善教・坂本（2017）の結果の再現性等を検証するためのオンラインフィールド実験を、（株）楽天リサーチ（現在は楽天インサイト）のモニタから実験参加者を募る形で実施した[12)]。実験の実施期間は2017年２月16日から20日までの４日間である。全国18歳から79歳までの男女を対象に実験を実施し2521人から有効回答を得た。調査ないし実験はオンライン上で行う意識調査実施システムである Qualtrics を用いて実施した。次項で説明する水準表記の無作為化は、この Qualtrics のシステムと外部の php プログラムを連動させる形で行った。

---

11）もっとも、デブリーフィングなどを行うかどうかにかかわらず、ディセプションは社会に対して深刻な害を与えると批判する研究者もいるので、実験実施には慎重になる必要がある。倫理委員会から承認を得たと説明すればよいと考える人がいるかもしれないが、そのような批判をする研究者はディセプションを承認した倫理審査委員会の見識も非難するので意味はない。

12）このオンラインフィールド実験は、関西学院大学「関西学院大学人を対象とする行動学系研究倫理委員会」の承認を受けて実施したものである（承認番号2016-51）。

推定結果の妥当性を高めるために、本稿では実験結果を Hainmueller (2012) で考案されたエントロピーバランシングスコアを用いて補正する。筆者らは実験参加者を集める際に、性別（男女）、年齢（18–29歳、30代、40代、50代、60代以上）、地域（北海道・東北、関東、中部、近畿、中四国、九州）の各カテゴリの比率が2015年度国勢調査のそれと一致するように調整した。しかし、それでもなお性別、年齢、居住地域などについて、実際の比率とは異なる結果となった。特に居住地については、実験参加者の分布が都市部（政令市在住者）に大きく偏っていた。以上の理由により本稿では、エントロピーバランシングスコアを作成し、これを用いて推定結果を補正することにした。具体的には、性別、年齢（上記 5 カテゴリ）、居住する市郡規模（政令市・東京23区、人口30万以上の市、人口30万未満の市、町村）を、国勢調査の値と一致させるための重み付け変数を作成し、これを用いて結果を補正することにした[13)]。

実験の実施に際しては、実験参加者の回答努力の最小化（satisfice）をいかに抑制するかも重要である。本稿では、以下の方法でこの問題に対処する。まず実験に関する設問の数問前に instructional manipulation check（IMC）による satisfice 傾向の検証を行い（Oppenheimer, et al. 2009; 三浦・小林 2015）、ここで satisficer と識別された人に対して、読み飛ばし行為を行わないように警告文を表示し注意喚起を促した。さらに実験に関する設問の直前に direct question（DQ）形式で再度 satisficer を識別し（Maniaci and Rogge 2014）、IMC と DQ の両者で satisficer として識別された実験参加者490人については、実験画面を注視しないことが高い確率で予測されるため、実験対象者から除外した[14)]。ただし DQ では識別されたが IMC では satisficer と識別されなかった実験参加者に対しては、IMC の場合と同様に警告文を表示するに留めた[15)]。これらの措置を講じることによって本稿では推定結果の妥当性の向上に努めた。

データの代表性（representativeness）について改めて説明しておく。本稿の実験参加者は、調査会社のモニタに登録した人に限定されており、代表性がどの程度満たされているかは不明瞭である。理想論をいうなら母集団に対する平均処置効果を推定する場合、選挙人名簿などを用いて無作為に抽出された人を対象に実験を実施する必要がある。しかし、母集団から無作為に抽出されたオンライン調査用のプールが現時点で存在しない日本では、調査費用や非回答バイアス等の問題から、無作為抽出された実験参加者を対象に実験を実施することなど不可能である。したがって標本の偏りを前提に代表性をいかに向上させるかという点から、この問題について考えなければならない。本稿では推定結果の補正によりこの問題に対処している。

---

13）実験結果についてウェイトを利用し結果を補正するかについては、いくつか見解の相違がある。例えば因果効果の異質性が存在しないことを仮定できる場合、ウェイトを使って結果を補正することには意味がない。さらに結果を補正することで、かえって誤差が拡大する場合もある。しかしながら本研究の実験結果は、後述するように年齢によって因果効果の値が異なっている。このような場合は推定結果を補正した方が妥当性の高い結果を得ることができると判断し、本稿では推定結果を補正することにした。

14）調査ないし実験へのアクセス総人数は4104人であった。うち530人が同意画面の時点で非同意を選択し、さらに563人は調査途中で回答を終了した。総アクセス数からこれらの人数と（強い）satisficer として識別された人数を除外したのが有効回答者数である。

15）IMC、DQ、および警告文の詳細は Online Appendix にて説明しているので、詳しくはそちらを参照のこと。

### （3）実験刺激と推定法

本稿では、善教・坂本（2017）と同様のヴィネットを用いて、寄付行動の規定要因を分析する。つまり内容が無作為に変化するチラシを提示し、実験参加者はそのチラシを見ながら、追加的に付与された100ポイントのうち、実際にいくら寄付するのかを尋ねるという実験である[16]。ただし研究資金の制約上[17]、善教・坂本（2017）のように仮想のチラシを繰り返し提示することは不可能だと判断した。内容こそ異なるが、類似のチラシを繰り返し提示すると「実際のチラシ」という情報の信憑性が薄れるため、その意味でも一回限りの方が、本稿の目的からしても望ましいといえる。

善教・坂本（2017）では繰り返し数が5回であるのに対して、本稿は1回のみである。本稿が直接的追試を行うものではなく概念的追試を行うものであるので、繰り返し回数を揃えなければならない理由はない。また善教・坂本（2017）が5回繰り返した理由は、単純に属性および水準数が多く1回だけだと推定値の標準誤差が大きくなってしまうからである。繰り返すたびに推定値に変化が生じる場合は問題となるが、善教・坂本（2017）ではキャリーオーバー効果（carryover effect）はなかったとされている。繰り返し数の相違は再現性の検証に何ら影響を与えない。

1回限りの実験であることを踏まえて、本稿では、無作為に変化する要因を善教・坂本（2017）でいう主体特性、管理運営費、返礼の3要因とした。また水準数についても、いずれも2とした。要因数や水準数が多すぎると推定結果の信頼性が低下するためである。

上述した3つの要因を取り上げた理由を説明する。第1に主体特性については、NPO法人格を持つ方が寄付行動に負の影響を与えるという、通説とは異なる結果が善教・坂本（2017）にて示されたからである。そのためここでの水準設定は「ボランティア団体」と「NPO法人」とした。第2に管理運営費は善教・坂本（2017）でもっとも寄付行動に強い影響を与えていた要因だからである。管理運営費の水準は最小値と最大値に対応する「0％」と「30％」とした。第3に返礼は、これまで寄付者の名前を記した返礼だと寄付が増えるとの知見が示されてきたにもかかわらず（Mason 2016）、善教・坂本（2017）では有意な影響はないという結果が示されたからである。返礼の水準は「手書きのメッセージカード」と「寄付者の名前を記した活動報告書」としている。

これら3要因の因果効果の推定法は善教・坂本（2017）と同じく線形回帰モデルである[18]。従属変数についても善教・坂本（2017）の設定に鑑み、いくら寄付したか（寄付ポイント）と寄付したかどうか（寄付確率）の2つとした。なお善教・坂本（2017）は標準誤差を回答者でクラスター化した頑健標準誤差としているが、本研究は繰り返し型の実験ではないため通常の標準誤差を推定する。また上述したようにエントロピーバランシングスコアを用いて推定結果を補正し、妥当な結果を得ようとする点も善教・坂本（2017）とは異なる。

---

16）実験に用いたヴィネットの詳細は善教・坂本（2017）のそれとほぼ同一のものであることから、本論中での詳細な説明は割愛する。詳しくはOnline Appendixに記しているのでそちらを参照のこと。

17）善教・坂本（2017）のような繰り返すタイプの実験とすると実験1回につき約25万円を用意しなければならない。そのような研究資金を持ち合わせていなかったため、1回限りの実験とした。

18）線形回帰モデルで因果効果を推定することの妥当性については、Hainmueller et al.（2014）で証明されている。

## 3　実験結果

### （1）予測

善教・坂本（2017）では前節で述べた3要因の因果効果について、以下の推定結果が示されている。まずNPO法人は、ボランティア団体の場合と比較して、寄付確率に対しては有意な影響を与えない一方（coef. = −0.023, s.e. =0.016）、寄付金額に対しては有意な負の影響を与える（coef. = −43.05, s.e. =21.00）。次に30％の管理運営費は、0％の場合と比較して寄付金額と寄付確率の両者に有意な負の影響を与える。具体的には寄付確率に対しては約6％ポイントの低下、寄付金額に対しては約101円の低下をもたらす。最後に寄付者の名前を記した活動報告書についてである。回帰係数の符号は従属変数によって異なり、寄付率の場合は負（coef. = −0.034, s.e. =0.017）、寄付金額の場合は正となる（coef. =1.654, s.e. =21.426）。ただし、ともに統計的に有意ではない。

善教・坂本（2017）の推定結果が信頼性と妥当性を兼ね備えたものであるならば、実際に寄付してもらうことを想定している本稿の実験においても、同様の結果を得ることができるだろう。寄付確率に関しては2値変数であるため、実験結果の直接的な比較検証が可能である。まずNPO法人（ボランティア→NPO法人格）の影響については、本稿の実験では統計的に有意ではないか、有意であったとしても小さな負の効果となることが予測される。次に管理運営費については（0％→30％）、有意な負の影響を与えるものと予測される。最後に寄付者の名前については（手書きのメッセージカード→寄付者の名前を記した報告書）、統計的に有意ではないか、もしくは有意だが小さな負の効果だと予測される。

寄付ポイントを従属変数とする場合については正確な予測は難しい。善教・坂本（2017）では寄付金額となっていたのに対して、本稿の実験では上限が100の（実験実施者によって配布された）寄付用の謝礼ポイントだからである。ただしどちらも金銭であることから、傾向としては善教・坂本（2017）における寄付金額を従属変数とする場合の推定結果に近似する結果が得られるものと予測される。したがってNPO法人と管理運営費については、本稿の実験でも統計的に有意な負の影響を与えるだろう。しかし回帰係数値は同じではなく、管理運営費の方がNPO法人よりも相対的に大きくなるだろう。一方の報告書の名前については、統計的に有意でないか、有意だとしても極めて小さな効果になるだろう。

### （2）実験結果

図1は、筆者らが実施したオンラインフィールド実験の結果を整理したものである。前節で述べたように、各要因ないし水準の効果は線形回帰モデルによって推定している。図中の黒い丸は、各要因ないし水準の平均因果効果の推定値であり、その横の棒は、点推定値の信頼区間である。黒色の横棒が95％信頼区間であり、灰色が99％信頼区間である。これらが0値に引かれている破線に重なっていない場合、その推定値は統計的に有意だと判断できる。左側の図は寄付ポイントを従属変数とする場合の推定結果であり、右側の図は寄付の有無（寄付確率）を従属変数とする

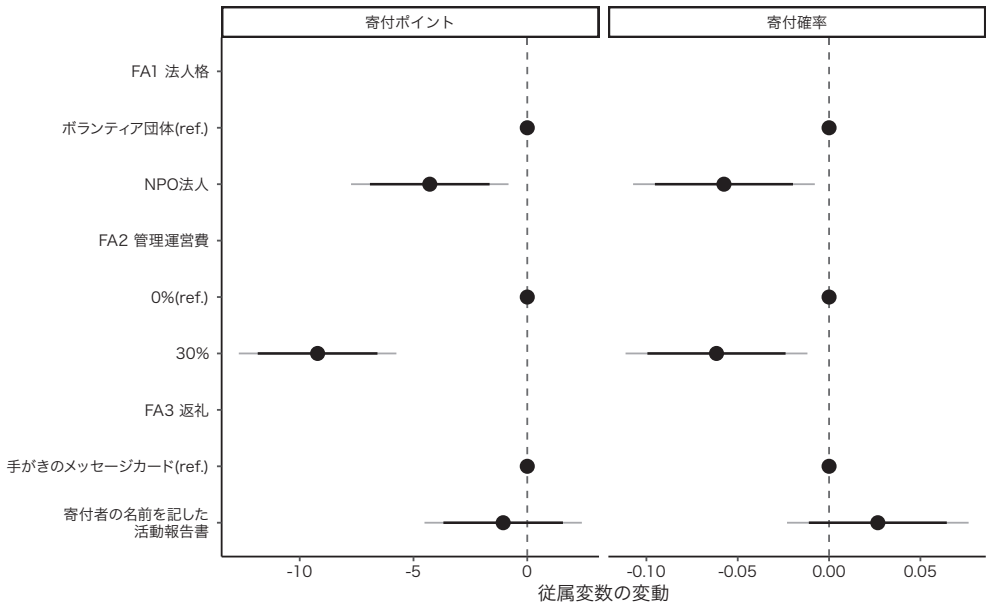


**図1　オンラインフィールド実験による因果効果の推定結果（ウェイト補正済）**

注）左図は寄付ポイントを従属変数とする推定結果であり、右図は寄付の有無を従属変数とする推定結果である。ref. は各要因の基準となる水準、図中の黒丸は各水準の平均因果効果の点推定値、黒色の横棒は推定値の95%信頼区間であり、灰色の横棒は99%信頼区間である。縦の破線は0値である。推定の際には共変量として性別、年齢、これまでの寄付経験、他者への同情に関する態度の4つを投入している。

場合の推定結果である[19)]。

まず、寄付ポイントを従属変数とする場合の推定結果から確認する（図1左図）。第1にボランティア団体からNPO法人へと団体表示が変化した場合、寄付ポイントは有意に低下する（coef. = −4.286, s.e. = 1.341）。第2に管理運営費が30%である場合、0%の場合と比較して寄付ポイントが有意に低下する（coef. = −9.214, s.e. = 1.342）。第3に寄付者の名前を記した活動報告書については、寄付ポイントに有意な影響を与えない（coef. = −1.059, s.e. = 1.340）。回帰係数の符号の向きも統計的に有意な影響かどうかという点も、善教・坂本（2017）と一致していることが、これらからわかる。寄付ポイントに与える効果量の相対的な大きさという点でも、管理運営費の影響力が法人格のそれの2倍程度であるなど、本稿の結果は善教・坂本（2017）の結果と一致している。

次に寄付確率を従属変数とする場合の推定結果を確認する（図1右図）。結論を先取りすれば、寄付ポイントの場合と比較して善教・坂本（2017）の結果と完全に一致するとはいえない結果であった。第1に、本稿ではNPO法人はボランティア団体と比較して有意に寄付確率を低下させるという結果となっており（coef. = −0.058, s.e. = 0.019, $p < 0.01$）、これは係数の符号の方向こそ

19）従属変数の記述統計量、つまり寄付ポイントと寄付確率の調査結果を予め述べておくと、寄付ポイントについては平均が29.93（S.D. = 34.78）、寄付確率は58.94%（S.D. = 24.20）であった。いずれもウェイト補正済みの推定結果である。

同一であるものの、善教・坂本（2017）と完全に一致する結果とはいえない。第2に管理運営費は、それが高いと実験参加者に寄付をさせない方向で有意な影響を与える。6％ポイント程度の変動という点も含め、この点については善教・坂本（2017）に近似する結果が得られた。第3に寄付者の名前を記した活動報告書については、回帰係数の符号は正だが統計的に有意ではなかった（coef. ＝0.027, s.e. ＝0.166）。

なお、因果効果の異質性（heterogeneity）についてここで簡単に述べておきたい。筆者らがこの点について分析したところ[20]、性別、過去の寄付経験、他者への同情の3共変量に関しては寄付行動に与える効果をほとんど左右しないという結果となった。一方、年齢に関しては因果効果を有意に左右する共変量であることが明らかになった。具体的には、寄付ポイントを従属変数とする場合は、すべての要因の因果効果と年齢の交互作用項が有意となった。また寄付確率を従属変数とする場合は、報告書に名前を記すかどうかという要因との交互作用項のみ、統計的に有意という結果になった。ただし後者の回帰係数はそれほど大きくない。従属変数の相違と異質性の推定結果の相違の因果関係を、本稿の分析から判断することはできないが、寄付行動に与える処置の効果は一様ではなく、年齢により異なる可能性がある。寄付行動の規定要因を分析する場合、被験者ないし調査対象者の選択バイアスに注意しなければならないことを、ここでの分析結果は示唆している。

### （3）実験結果の検討

本稿の実験結果は、1）寄付ポイントを従属変数とする場合、善教・坂本（2017）とほぼ一致する推定結果が得られたが、2）寄付確率を従属変数とする場合完全に一致するとはいえず、一部は善教・坂本（2017）と異なっていた。表1は善教・坂本（2017）と本稿の実験結果を整理し、比較したものである。先行研究の実験結果を再現できたかという疑問に対しては、概ね成功していると評価できるのではないだろうか。本稿の実験設計は仮想的なヴィネットを繰り返し提示するものではなく、実際に寄付用の謝礼ポイントを与え、それを用いて寄付してもらうというものであった。そのようなセッティングの違いを考慮しても、善教・坂本（2017）の結果をおおよそ再現できたといえる結果だと要約できるだろう。

ただし先に述べたように、寄付確率については、本稿の実験結果と善教・坂本（2017）の間に相違が見られる。したがってこの点について、改めてここで詳しく検討しておきたい。善教・坂本（2017）では寄付の有無に対する主体特性について、その因果効果は統計的に有意ではないとの結果が示されていた。しかしNPO法人の回帰係数の符号は負であり、この点は本稿の実験結果と一致する。有意か有意ではないかだけ見ると異なるが[21]、両者の差は大きな差ではなく、信頼

20）異質性の分析結果はOnline Appendixに詳述しているのでそちらを参照のこと。

21）異なる結果となった理由はとしては、偶然そのような結果となった可能性のほか、ウェイト変数により推定結果を補正したからというものも考えられる。細かな結果は省略するが、NPO法人格の効果だけ推定結果を補正しない場合回帰係数の点推定値が小さくなり、また寄付確率を従属変数とする場合、5％水準だと有意ではないという推定結果となった。NPO法人表記の効果については他の要因以上に不明瞭な点があり、さら

表１　善教・坂本（2017）とオンラインフィールド実験の結果の比較

| | 善教・坂本（2017） | | オンラインフィールド実験 | |
|---|---|---|---|---|
| | 寄付金額 | 寄付確率 | 寄付ポイント | 寄付確率 |
| 寄付主体 | 負で有意 | n. s. | 負で有意 | 負で有意 |
| NPO 法人 | （－50円） | （－2%） | （－4ポイント） | （－6%） |
| 管理運営費 | 負で有意 | 負で有意 | 負で有意 | 負で有意 |
| 30% | （－100円） | （－6%） | （－9ポイント） | （－6%） |
| 返礼 | n. s. | n. s. | n. s. | n. s. |
| 寄付者名を記載 | （０円） | （－4%） | （－1ポイント） | （３%） |

注）n. s. は５%水準で統計的に有意な結果を得られなかったもの。括弧内は小数点第１位を四捨五入したおよその因果効果の点推定値。

区間を考慮すると誤差の範囲内におさまっている。再現に失敗したというよりも、おそらく真の効果が小さいために、実験ごとに有意になったりならなかったりしてしまうのだという解釈がもっとも適切だろう[22)]。

また、本稿の実験結果は、仮想的なシチュエーションに基づくサーベイ実験の結果であっても、それが現実の行動と明確にリンクする場合があることを示すものである。換言すれば本稿は、善教・坂本（2017）の結果が、現実的妥当性という意味での外的妥当性の高い結果であったことを示すものである。本稿の実験では、寄付用に追加した謝礼ポイントのすべてを実験参加者全員に支払ったが、それはあくまで事後的に行った措置である。仮想的な状況をイメージさせる実験だからといって、そこから得られた知見に妥当性がないわけでは決してないことを、本稿の分析結果は明らかにしている。

## 4　結論

本稿では、これまでほとんど行われてこなかった実験結果の概念的追試に取り組んだ。具体的には善教・坂本（2017）の実験結果について、著者自身が先行研究の実験結果の不備を指摘し、それを解決するために考案したオンラインフィールド実験に基づき、結果の再現性を検証した。本稿の実験結果は善教・坂本（2017）の結果と概ね一致するものであった。この結果は、寄付行動の規定要因を分析した善教・坂本（2017）の知見の頑健性を示すと同時に、仮想的な状況を設定するサーベイ実験の結果であっても、外的妥当性を兼ね備えたものになりえることを明らかにしている。

誰がどのような理由で寄付するかという疑問に対しては、これまで多くの実証研究がその解答を提示してきた。RCT の知見は妥当性と信頼性を兼ね備えたものといえるが、他方で費用等の都

なる実証分析を蓄積する必要がある。

22）類似の実験を複数回繰り返したからこのような判断ないし解釈が可能になる。この点を明らかにしたことも本稿の意義の１つだといえるだろう。

合上、全ての研究者がRCTを行えるわけではない。本稿で実施したサーベイ実験は、RCTなどと比較すれば安価に実施できる。もちろん知見の外的妥当性という点で批判を受ける可能性はあるが、RCTの実施に係る費用などを考慮しつつ、サーベイ実験が有力な選択肢になりうるのであれば積極的に実施することが望ましいだろう。

本稿の実験結果は、寄付金額のうちのいくらかを管理運営費として利用することと、「NPO」という呼称に否定的な評価が下されている実態を明らかにするものである。もっとも、だから寄付を募る際にこれらの情報を開示すべきではないと主張するものでは決してない。本稿は管理運営費として用いることへの理解を一層深めることや、NPOに対する否定的なイメージを払拭する必要性を、改めて示すものとして位置づけられる。

本稿の課題は以下の通りである。第1は操作チェックである。本稿では操作チェックを行えていない。今後はこの点についても確認していく必要がある[23)]。第2は事前登録（pre-registration）である。多くの実験研究に対して、現在、出版バイアスやp-hackingを防ぐための事前登録が推奨されている。本稿は残念ながら、自己追試・再現研究であるにもかかわらず、この点に大きな課題が残されている。第3は直接的追試の必要性である。厳密な意味でいうと、再現性を検証するには直接的追試が望ましく、概念的追試は「新規」の研究成果の報告だという指摘がある（Wilson et al. 2020）。概念的追試ではなく直接的追試を行い、本研究の信頼性について検証する必要がある[24)]。

最後に繰り返しとなるが、日本の社会科学では追試や再現性を確認する研究がほとんど行われておらず、それゆえに日本には先行研究の信頼性等を確認する研究に対して、その価値を理解しないまま不当に評価する研究者が少なからず存在する。そのような認識を改めると同時に、一つでも多くの再現研究の成果を、心理学者以外の研究者が蓄積していくことも残された課題であることを、ここに記しておきたい。

## 参考文献


Alpizar, Francisco, Fredrik Carlsson, and Olof Johansson-Stenman（2008）“Anonymity, Reciprocity, and Conformity: Evidence from Voluntary Contributions to a National Park in Costa Rica.” *Journal of Public Economics* 92(5–6) : 1047–1060.

Barabas Jason and Jennifer Jerit（2010）“Are Survey Experiments Externally Valid?” *American Political Science Review*, 104（2）: 226–242.

DellaVigna, Stefano, John A. List, and Ulrike Malmendier（2012）“Testing for Altruism and Social Pressure in Charitable Giving.” *Quarterly Journal of Economics* 127（1）: 1–56.

Hainmueller, Jens（2012）“Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies.” *Political Analysis* 20: 25–46.

Hainmueller, Jens, Daniel J. Hopkins, and Teppei Yamamoto（2014）“Causal Inference in Conjoint


23）操作チェックおよびその重要性などについてはKane and Barabas（2018）を参照のこと。

24）Wilson et al.（2020）では再現性は有意かどうかではなく効果量（と信頼区間）から確認すべきとされている。本稿は概念的追試であるが直接的追試と同じく再現できるかを確認するものでもあることから、効果の大きさなども検討している。

Analysis: Understanding Multidimensional Choices via Stated Preference Experiments." *Political Analysis* 22 (1) : 1-30.
Hainmueller, Jens, Dominik Hangartner, and Teppei Yamamoto (2015) "Validating Vignette and Conjoint Survey Experiments Against Real-world Behavior." *PNAS*, 112 (8) : 2395-2400.
Ishida, Yu and Naoko Okuyama (2015) "Local Charitable Giving and Civil Society Organizations in Japan." *Voluntas* 26 (4) : 1164-1188.
伊藤公一朗 (2017)『データ分析の力：因果関係に迫る思考法』光文社。
Jackson, Kristoffer (2016) "The Effect of Social Information on Giving from Lapsed Donors: Evidence from a Field Experiment." *Voluntas* 27 (2) : 920-940.
James, Russell N. and Deanna L. Sharpe (2007) "The Nature and Causes of the U-Shaped Charitable Giving Profile." *Nonprofit and Voluntary Sector Quarterly* 36 (2) : 218-238.
Kane, John V. and Jason Barabas (2018) "No Harm in Checking: Using Factual Manipulation Checks to Assess Attentiveness in Experiments." *American Journal of Political Science*, online version.
King, Gary (1995) "Replication, Replication." *PS: Political Science and Politic*s, 28 (3) : 444-452.
Laitin, David D. and Rob Reich (2017) "Trust, Transparency, and Replication in Political Science." *PS: Political Science and Politic*s, 50 (1) : 172-175.
Maniaci, M. R. and R. D. Rogge (2014) "Caring about Carelessness: Participant Inattention and Its Effects on Research." *Journal of Research in Personality* 48: 61-83.
Mason, Dyana P. (2016) "Recognition and Cross-Cultural Communications as Motivators for Charitable Giving: A Field Experiment." *Nonprofit and Voluntary Sector Quarterly* 45 (1) : 192-204.
Muise Daniel and Jennifer Pan (2019) "Online Field Experiments." *Asian Journal of Communication*, 29 (3) : 217-234.
三浦麻子・小林哲郎 (2015)「オンライン調査モニタの Satisfice に関する実験的研究」『社会心理学研究』31：1-12。
Open Science Collaboration (2015) "Estimating the Reproducibility of Psychological Science." *Science*, 349 (6251).
Oppenheimer, Daniel M., Tom Meyvis, and Nicolas Davidenko (2009) "Instructional Manipulation Checks: Detecting Satisficing to Increase Statistical Power." *Journal of Experimental Social Psychology*, 45(4): 867-872.
Parigi, Paolo, Jessica J. Santana, and Karen S. Cook (2017) "Online Field Experiments: Studying Social Interactions in Context." *Social Psychology Quarterly*, 80 (1) : 1-19.
Shadish, William R., Thomas D. Cook, and Donald T. Campbell (2002) *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. New York: Houghton Mifflin Company.
Shah, Anuj K., Sendhil Mullainathan, and Eldar Shafir. (2019) "An Exercise in Self-replication: Replicating Shah, Mullainathan, and Shafir (2012)." *Journal of Economic Psychology*, 75 (102127).
友永雅己・三浦麻子・針生悦子 (2016)「心理学の再現可能性　我々はどこから来たのか我々は何者か我々はどこへ行くのか ── 特集号の刊行に寄せて ── 」『心理学評論』59 巻 1 号: 1-2。
Wilson, Brent M., Christine R. Harris, and John T. Wixted (2020) "Science is Not a Signal Detection Problem." *PNAS*, online published.
善教将大・坂本治也 (2017)「何が寄付行動を促進するのか：Randomized Factorial Survey Experiment による検討」『公共政策研究』17：96-107。

**[謝辞]**

本稿の実験は、筆者らとの共同研究プロジェクトのメンバーである岡本仁宏先生（関西学院大学）および三浦麻子先生（大阪大学）のご助力を得ながら設計・実施した。ここに記して感謝申し上げる次第である。無論、残された誤りはすべて筆者らの責にある。

［付記］

本稿は「関西学院大学2016年度大学共同研究（学長指定研究）熊本地震関連共同研究（公募型）」による研究成果の一部である。