測定・評価・研究法

『教育心理学研究』における統計改革の現状

——サンプルサイズ設計を中心に——

山 内 香 奈

(成城大学)

Statistical Reform in *The Japanese Journal of Educational Psychology*:
Focusing on Sample Size Planning

KANA YAMAUCHI

(SEIJO UNIVERSITY)

Descriptions about planning sample sizes using a test of significance in 29 studies published in *The Japanese Journal of Educational Psychology* from July 2019 to June 2020 were reviewed and compared with those in *The Japanese Journal of Psychology*, *Japanese Psychological Research (JPR)* and *Journal of Personality and Social Psychology (JPSP)*. The percentage of papers describing the statistical basis of sample sizes was the highest by approximately eight times in *JPSP* (93%) compared to the other three journals (7–15%). The author also examined if the effect sizes and confidence intervals, which are other crucial statistical reforms, were described in the results. Results indicated that all four journals described these more often than sample sizes and that all the papers in *JPSP* mentioned effect sizes and confidence intervals. On the other hand, 44% of *The Japanese Journal of Educational Psychology* papers mentioned neither effect sizes nor confidence intervals. Finally, the author discusses measures to promote statistical reforms in *The Japanese Journal of Educational Psychology*, especially in relation to sample size planning.

Key Words: statistical reform, sample size planning, power analysis, effect size, confidence interval

本稿では、2019年7月から2020年6月までの1年間に『教育心理学研究』に掲載された29編のうち、有意性検定を用いた研究論文のサンプルサイズ設計に関する記述状況について概観した。その際、『心理学研究』、*Japanese Psychological Research (JPR)*、*Journal of Personality and Social Psychology (JPSP)*のそれと比較し論じた。研究のサンプルサイズの根拠について何らかの記述がみられた論文の割合は*JPSP*が最も高く(93%)、他の3誌(7–15%)の約8倍であった。また、検定力分析と同様、統計改革の柱である効果量や信頼区間が分析結果に併記されているかを調べたところ、概して4誌ともサンプルサイズ設計の記載より実践度は高く、*JPSP*では対象論文全てにいずれかの記載がみられた。一方、『教育心理学研究』では効果量、信頼区間のいずれも記載がない論文が全体の44%を占め、実践度が最も低かった。最後に、『教育心理学研究』における統計改革の促進について、特にサンプルサイズ設計の実践の促進に向けた方策について筆者なりの見解を述べた。

キーワード：統計改革、サンプルサイズ設計、検定力分析、効果量、信頼区間

はじめに

本稿では、2019年7月から2020年6月末までの1

年間に刊行された『教育心理学研究』における実証的な研究論文について、研究の標本の大きさ(以降、サンプルサイズと表す)がどのように決められ、それが論文中

にどのように記されているのか、研究のサンプルサイズ設計の観点から概観する。

教育心理学をはじめとする心理学の多くの研究において、統計的仮説検定は広く用いられているが、この統計的仮説検定には、長く指摘されてきた問題点がある。それは、検定の結果が有意となるかどうかサンプルサイズが決定的な影響を与えているが、そのサンプルサイズは多くの場合、恣意的に決められており、検定の結果が恣意的判断に依存し、ひいては研究の結論が恣意的な判断に依存してしまうことである（南風原, 2014b）。このような問題を孕む統計的仮説検定に過度に依存することの限界を認識し、そこから脱却する必要性を強く叫ぶ動きは、「統計改革」（Finch et al., 2004; 大久保・岡田, 2012）と呼ばれ、注目されて久しい。その統計改革では、効果量、信頼区間、検定力を主な柱とし、それらを活用することが推奨される。具体的には、（1）統計的仮説検定の結果と併せて効果量や信頼区間を報告することや、（2）検定力分析に基づくサンプルサイズ設計を行うことなどである。これらにより、統計的データ分析をより現実的な意味のあるものにするのが目指される。

上述の（1）の効果量や信頼区間を検定結果と併せて表記することについては、大久保（2009）が海外の論文誌において効果量や信頼区間を報告する論文が増えている一方、日本における統計改革が全く進んでいないことを、1982年から2008年までに刊行された『基礎心理学研究』を対象に調査し、述べている。

上述の（2）の検定力分析を用いたサンプルサイズ設計については、昨今、我が国においてもその重要性が論文や著作物、学会のシンポジウムなどの場で論じられる機会が少しずつではあるが増えてきている（井関, 2019; 村井・橋本, 2018; 山内, 2017）。井関（2019）は、認知心理学領域の実験研究を対象に、*Psychonomic Bulletin & Review*（以降、PBL）と*Journal of Experimental Psychology: General*（以降、JEPG）の過去1年分の実験報告論文について、サンプルサイズ設計の有無やそのサンプルサイズ設計に関する記述内容について文献調査を行い、両誌を総合すると、検定力分析に限らず何らかのサンプルサイズ設計の報告を行った論文は、全体の41.8%を占めることを述べている。このように海外誌の論文においてサンプルサイズ設計に関する記述が充実してきた背景には、アメリカ心理学会の論文作成マニュアル（APA Publication Manual）の影響が大きいと考えられる。APA Publication Manual 第7版（American Psychological Association, 2020）では、サンプルサイズの決定方法（検

定力分析、母数の推定精度に基づく方法）について述べることで、意図したサンプルサイズと実際に収集したサンプルサイズが異なる場合は後者も記すこと、また、中間分析や停止規則を用いて意図したサンプルサイズの修正を行った場合は、その方法と結果を記すことが明記されている。

サンプルサイズを事前に決めることは、研究の信頼性や妥当性、効率性、倫理性¹を担保する上で欠かせない。また、サンプルサイズを事前に決めず、データを追加したり、途中で取るのをやめたりすることは、近年、研究の不正行為（Questionable Research Practices）として認識されるようになってきている。そのため、サンプルサイズを事前に決め、その根拠を論文に記すことは、昨今の研究の信頼性や再現可能性を高めようとする世界的潮流の中で、ますます重要視されていくと考えられる。

一方、国内に目を転じてみると、『教育心理学研究』が準拠する日本心理学会の「執筆・投稿の手びき」では、該当する記載はみられない。これまで国内誌を対象に、心理学の研究のサンプルサイズの報告状況に着目し、詳細に調べたものはないが、恐らく上述の井関（2019）が示した海外誌での報告割合よりも大きく下回ることが予想される。

近年、心理学の研究において、脱検定やベイズ的アプローチを推奨する動きなどがみられるが、今すぐに統計的仮説検定を用いる研究がなくなるとは考えにくく、サンプルサイズの恣意性の問題をこのまま放置しておくことは賢明ではない。また、統計的仮説検定の限界を超えるための方法として近年、注目されているメタ分析を促すためにも、統計的仮説検定を用いた研究におけるサンプルサイズ設計や効果量の提示といった研究実践の重要性を、研究者コミュニティの中で広く共有することが有効であると考えられる。

今後、『教育心理学研究』における統計的仮説検定のより意味のある利活用を促すには、まずは『教育心理学研究』に掲載された論文での報告状況（報告件数、報告内容）を国内誌や海外誌のそれと比較し、量的に把握した上で、望ましいとされる実践の具体例について知ることが有効と考えられる。

以上のような問題意識から、冒頭で述べたように、本稿では2019年7月から2020年6月末までの1年間

¹ 例えば、欺瞞手続きによる実験条件の統制を行う研究や、有効性が確認されていない教育法や治療法を試験的に適用する実践研究などでは、必要以上に多くの参加者の協力を要請することは、倫理的に問題があると考えられる。

に『教育心理学研究』で発表された論文のうち、統計的仮説検定を用いた研究のサンプルサイズ設計の報告状況について概観する。具体的には、『心理学研究』とその発行元である日本心理学会が発行している英文誌 *Japanese Psychological Research* (以降、*JPR*)、海外誌では APA マニュアルが採用されている *Journal of Personality and Social Psychology* (以降、*JPSP*) を取り上げる²。その際、論文誌に掲載されている論文の特徴、特に採用されている研究法について調べ、その傾向を把握する。その際、研究法の分類方法については、2017 年 7 月から 2018 年 6 月末までの 1 年間に『教育心理学研究』に掲載された論文で採用された研究法について調査した登藤 (2019) を踏襲する。

また、検定力分析と同様、統計改革の重要な柱である効果量や信頼区間についても、それらの論文での報告状況について調べ、サンプルサイズ設計の報告状況と比較し論じる。それらを踏まえ、本稿の最後に、サンプルサイズ設計の実践やその報告の促進に向けた方策について筆者の見解を述べる。

なお、本年報には、日本教育心理学会第 62 回総会で発表された測定・評価・研究法部門の研究について報告するという役割もあるため、以下、『教育心理学研究』に掲載された論文について論じる前に、簡単にそれについて述べる。

日本教育心理学会第 62 回総会における 「測定・評価・研究法」部門の研究

登藤 (2019) は、日本教育心理学会第 60 回総会の測定・評価・研究法部門の研究で採用された研究法を、「調査研究」か「実験研究」か「実践研究」かに分類している。各分類の定義は、下記に示す南風原 (2002,

2011) を参考に定義している。「調査研究」とは、研究者の側から現実に対して特に手を加えることなく、現実をそのまま把握しようとする研究、「実験研究」とは、研究者の側で意図的に条件を操作して、その効果を調べる研究、「実践研究」とは、対象者の現実を改善することを主目的として現実介入しながら行う研究のことをそれぞれ指す。なお、「調査研究」、「実験研究」、「実践研究」のいずれにも該当しない研究は、「その他」に分類している。

以上の定義を用いて、日本教育心理学会第 62 回総会において「測定・評価・研究法」部門の研究として発表された 21 件の研究について分類した。その結果、「調査研究」は 8 件 (約 38%)、「実験研究」は 0 件、「実践研究」は 2 件 (約 10%)、「その他」は 11 件 (約 52%) であった。この結果と、日本教育心理学会第 60 回総会での測定・評価・研究法部門の研究について調べた登藤 (2019) の結果を比較すると、まず、総件数は今回 10 件減っていた。次に、研究法の分類では、調査、実験、実践、その他の順に 18 件 (約 58%)、0 件、3 件 (約 10%)、10 件 (約 32%) であったことから、「実験研究」が 0 件、「実践研究」が全体の約 10% であることは変わらないが、「調査研究」の割合は約 20% 減り、「その他」の割合は 20% 増えていた。「その他」に分類される研究には、尺度開発研究、シミュレーション研究、項目分析研究など、測定・評価・研究法の部門に特徴的な研究が多くみられた。総発表件数の減少が、COVID-19 の影響によるものかは明らかではないが、少なからず影響を与えたと考えられる。

調査対象誌および論文

以降、本稿で概観するのは、2019 年 7 月上旬から 2020 年 6 月の 1 年間に発行された『教育心理学研究』(67 巻 3 号—68 巻 2 号) に掲載された 32 編のうち、方法論の開発に関する論文と展望論文の各 1 編、掲載が取り下げられた 1 編を除く 29 編 (安達・安達, 2019; 芦田, 2020; 石山他, 2020; 神谷, 2020; 河合, 2019; 木村・小泉, 2020; 吉良他, 2020; Kuramoto, 2019; 楠見, 2019; 牧, 2019; 永井, 2019; 長峯・外山, 2019; 長峯他, 2019; 中野他, 2020; 庭山, 2020; 則武他, 2020; 小野田・大澤, 2020; 尾之上・井口, 2020; 酒井・窪田, 2019; 酒井・相川, 2020; 杉本他, 2019; 竹下・藤田, 2019; 外山, 2020; 外山・湯, 2019; 脇・須藤, 2020; 山田, 2020; 山田・小泉, 2020; 吉田他, 2019, 2020) である。その際、比較する 3 誌については、下記の論文を概観する。『心理学研究』は、2019 年 8 月から 2020 年 6 月に発行された 90 巻 3 号から 91 巻 2 号までに掲載された 47 編のうち、展望論文 2 編、メタ分析

² 心理学研究の再現性を巡り、とりわけ大きな影響力があったと考えられる Open Science Collaboration (2015) では、心理学の各分野を代表する 3 誌 (*Psychological Science*, *Journal of Personality and Social Psychology*, *Journal of Experimental Psychology: Learning, Memory, and Cognition*) に掲載された 100 の研究について追試した結果を報告している。そこでは、5% 水準で有意だった研究の数は、元の論文と比べて追試では半分以下であったこと、効果量の中央値も追試では半減以下になったこと、4 分の 1 の追試では負の効果量となり、元の論文とは逆方向の効果を示したことが述べられた。この結果は、「心理学の研究結果は 6 割以上が再現不可能」というセンセーショナルなタイトルで広く報じられ、その後、研究の再現可能性を検討する大規模プロジェクトがいくつか発表された。これらから、Open Science Collaboration (2015) が取り上げた 3 誌に着目し、中でも『教育心理学研究』との比較のしやすさから、*Journal of Personality and Social Psychology* を本稿での調査対象誌とした。

1編を除く44編を、*JPR*は、2019年1月から2019年10月までに発行された61巻1号から61巻4号までに掲載された24編のうち、論説、2次分析研究、メタ分析研究の各1編を除く21編をそれぞれ対象とした。*JPSP*は、2019年7月から2019年12月までに発行された117巻1号から117巻6号までに掲載された53編のうち、メタ分析、2次データ分析（大規模プロジェクト研究含む）の10編を除く43編を対象とした。

研究法

これらの論文について、採用された研究法について調べた。まず「量的研究」か「質的研究」かを、次に論文の種別として、「調査研究」か「実験研究」か「実践研究」かを分類した。なお、「調査研究」、「実験研究」、

「実践研究」の各定義については、上述の日本教育心理学会第62回総会の研究の分類に用いた定義を用いた。また、「量的研究」か「質的研究」かは、南風原（2011）を参考にした登藤（2019）と同じ定義とした。「量的研究」とは、数量で表されている数量データを取得し、結果に相当する節において、これらを適切な手法により分析した結果を報告している研究を、「質的研究」とは、数量化されていない質的データを取得し、結果に相当する節において、これらを適切な手法により分析した結果を報告している研究を、それぞれ指す。

量的研究・質的研究

まず、『教育心理学研究』の対象論文29編について、「量的研究」か「質的研究」かを分類した結果をTable 1の2列目に示す。Table 1から質的研究は2編（約7%）、

Table 1

『教育心理学研究』の対象論文の研究法・統計的仮説検定の利用と結果の報告状況

筆者	量／質	研究の種別			統計的仮説 検定の利用	Nの根拠の 記述	信頼区間の 記述	効果量の 記述
		調査	実験	実践				
杉本・遠藤・飯田・青山・中井	量	○			あり	経験	なし	なし
長峯・外山・湯・三和・肖・相川	量	○			あり	なし	あり	なし
外山・湯	量	○			あり	なし	なし	なし
長峯・外山	量		○		あり	なし	あり	あり
Kuramoto	質	○			—	—	—	—
牧	量	○			あり	なし	なし	なし
酒井、窪田	量		○		あり	なし	あり	あり
吉田、吉澤、浅野、玉井、吉田（2019）	量	○			あり	なし	あり	あり
竹下、藤田	質	○			—	—	—	—
永井	量	○			あり	なし	なし	あり
河合	量	○			あり	なし	あり	あり
安達・安達	量			○	あり	経験	なし	あり
楠見	量			○	なし	—	—	—
石山・鈴木・及川・及川	量		○		あり	なし	なし	あり
吉良・尾形・上手	量	○			あり	なし	なし	なし
山田	量		○		あり	なし	なし	なし
脇・須藤	量			○	なし	—	—	—
小野田・大澤	量			○	あり	なし	あり*	あり*
中野・鋤柄・村・中川・大野	量			○	あり	なし	なし	なし
庭山	量			○	あり	なし	なし	なし
酒井・相川	量	○			あり	なし	あり	あり
尾之上・井口	量		○		あり	なし	なし	なし
則武・武井・寺崎・門田・竹内・湯澤	量		○		あり	なし	なし	あり
吉田・吉澤・浅野・玉井・吉田（2020）	量	○			あり	なし	なし	なし
神谷	量	○			あり	経験	あり	あり
外山	量	○			あり	統計	あり	なし
木村・小泉	量			○	あり	なし	なし	なし
芦田	量			○	あり	なし	なし	なし
山田・小泉	量			○	あり	なし	なし	あり

注1) 表中の著者は『教育心理学研究』への掲載順に記載。

注2) 「調査」「実験」「実践」については当該研究の研究法として該当するものに「○」と表記。

注3) 「Nの根拠の記述」については「経験」、「統計」に分類されたものに「経験」、「統計」と表記。

注4) 表中の「*」は母集団効果量の信頼区間を含むことを示す。

量的研究は27編(約93%)であり、これらの割合は登藤(2019)で報告された数値と同じであった。『教育心理学研究』に掲載された論文における量的研究と質的研究の割合はここ数年、安定していることが窺える。

なお、比較対象の論文誌の結果は、『心理学研究』とJPSPでは全て量的研究、JPRは1編が質的研究であった。

調査研究・実験研究・実践研究

次に、対象論文について、それが調査研究か、実験研究か、実践研究かという観点から分類した。調査研究、実験研究、実践研究の各定義は、上述で述べた日本教育心理学会第62回総会の研究の分類で用いたものと同じである。なお、『教育心理学研究』において実践研究に分類されている研究は全て実践研究とした。ただし、複数の研究から構成された論文については、井関(2019)の調査結果と比較可能性を考え、1つ目の研究で分類した。ただし、井関(2019)と同様に、1つ目の研究にサンプルサイズ設計に関する記載がない論文の場合は2つ目以降の研究の方法も確認した。分類結果をTable 1の3列目から5列目に示す。調査研究は14編(約48%)、実験研究は6編(約21%)、実践研究は9編(約31%)であった。登藤(2019)では、順に約59%、約17%、約24%であり、実験研究の割合は大きく変わらず、調査研究が約11%減り、実践研究は約7%増えていた。

なお、『心理学研究』では、調査研究は32編(約73%)、実験研究11編(約25%)、実践研究1編(約2%)であった。JPRは、調査研究は7編(約32%)、実験研究14編(約64%)、実践研究1編(約5%)であった。JPSPは、調査研究11編(約21%)、実験研究41編(約79%)であった。『教育心理学研究』と『心理学研究』では調査研究の割合が高く、JPRとJPSPでは実験研究の割合が高い。『教育心理学研究』は実践研究の割合が他誌に比べて高いという特徴も有す。

統計的仮説検定の利用状況

量的研究に分類された『教育心理学研究』の27編について統計的仮説検定が使用されているかどうかを調べ、結果をTable 1の6列目に示した。Table 1から、統計的仮説検定が使用されていなかったのは、楠見(2019)と脇・須藤(2020)の2編の実践研究であり、これらでは記述統計が使用されていた。『教育心理学研究』に掲載された量的研究のうち約93%は統計的仮説検定が用いられていたことになる。『心理学研究』、JPR、JPSPでは統計的仮説検定が用いられていない論文の数

は順に1編、0編、1編であり、その使用割合は極めて高いと言える。

サンプルサイズ設計に関する記述状況

『教育心理学研究』

『教育心理学研究』の統計的仮説検定を用いた27編を対象にサンプルサイズ設計の記述状況を調べた。その際、井関(2019)を参考に、まず、論文の「方法」、「参加者」のセクションにおいて、サンプルサイズの決定方法、目標サンプルサイズ、効果量、検定力に関する記述があるかを調べた。同一論文の中に複数の調査や実験がある場合は、全ての研究について確認した³。井関(2019)では、サンプルサイズ設計に関する記述について、「統計」、「経験」、「なし」と、事前の決定はしていないが、事後的に観測検定力を報告しているものを「事後」として大きく4つに分類している。「統計」とは、効果量や検定力の具体的な数値に言及しているもの、「経験」とは、それらの記載はないが、何らかの設計や事前のサンプルサイズ決定を行ったことを述べたもの、「なし」とは、「統計」および「経験」のいずれにもあてはまらないものを指す。なお、井関(2019)を参考に⁴、論文中に「検定力分析を行った」と記述している論文でも、効果量も検定力も記載していないものは「経験」に分類した。この分類に沿って『教育心理学研究』のサンプルサイズ設計に関する記述内容を調べた結果をTable 1の7列目に示した。「統計」は外山(2020)の1編、「経験」は杉本他(2019)、安達・安達(2019)、神谷(2020)の3編、残りは全て「なし」に分類され、サンプルサイズ設計に関する何らかの記述があった論文は全体の約15%であった。記述がみられた4編は、調査研究が3編、実践研究が1編であった。

論文に記述されていた内容について順にみていく。まず、「統計」に分類された外山(2020)では、方法のセクションに、関東圏内にある大学に在学する大学生375名を対象にすることが書かれており、脚注に検定力分析に関する下記の記述がみられた。

研究1、研究2ともに事前にサンプルサイズの設計を行った($f^2=0.15$ (中程度の効果量)、 $\alpha=.05$, power=

³ 井関(2019)では、同一論文の中に複数の実験がある場合は、基本的には第一実験の方法を参照し、第一実験にサンプルサイズ設計に関する記載がない論文の場合には第二実験の方法も確認していた。そのため、本稿のように論文中の全ての実験を調査対象とはしていない。

⁴ 井関(2019)では、「検定力」ではなく「検出力」の用語が用いられているが、本稿では「検定力」と表記している。

0.80で設定)ところ、サンプルサイズは68となったが、実際は、設計されたサンプルサイズよりも多くのサンプルを収集した。(外山, 2020, p. 176)

当該箇所に明示的には書かれていないが、階層的重帰帰分析を想定し、Cohen (1992) の効果量の定義に基づき、検定力分析を行っていると考えられる。ただし、中程度の効果を想定した根拠については述べられていない。目標数に対して実際に収集したデータ数はかなり多いことが窺える。

次に、「経験」に分類された3編は、本文中に目標人数と実際に収集した人数が記されていたが、目標人数の根拠については述べられていなかった。なお、杉本他 (2019) は予備調査において、安達・安達 (2019) は授業時間内でのプレポストデザインによる介入効果を調べる調査において、それぞれ目標人数と実際に収集した人数が下記のように記されていた。

インターネット調査会社が保有する425,589人のモニターを対象に、第一子に小学生の子を持つ保護者と第一子に中学生の子を持つ保護者に、調査の告知を行い各50名程度を目標に募集し、小学生の第一子を持つ保護者52名、中学生の第一子を持つ保護者52名(男性=39名、女性=65名、年齢30代=27名、40代=73名、50代=4名)の回答を収集した。(杉本他, 2019, p. 151)

年度当初、30名程度を対象としてATの授業を対象として、…を6月に開講する予定であった。それに合わせて、pre-postモデルでATの効果を検証する予定であった。しかし、ガイダンスを行ったところ、100名近くの学生が受講を希望した。…最終的な受講者は88名であった(筆者らはいずれの学生とも面識はなかった)。(安達・安達, 2019, p. 319)

神谷 (2020) では、調査上のデザインの観点から人数を決めたことが下記のように述べられていた。

(株)クロスマーケティングのモニタ(2016年2月時点で全国170万人登録)を、性別、年齢の偏りを防ぐために、20歳代から60歳代の10歳刻みで5群、性別で2群の計10群に各群100名ずつのデータを収集することとした。…数値の丸めの関係で協力者の総数は1,004名となった。(神谷, 2020, p. 163)

『心理学研究』

同様に、『心理学研究』で統計的仮説検定が用いられた43編を分類した結果、「統計」はターン他 (2019) と岩本 (大久保) 他 (2020) の2編、「経験」は増田他 (2019) の1編、「事後」は岩本 (大久保) 他 (2020) 1編、残りは全て「なし」であった。サンプルサイズ設計に関する記述がみられた論文の割合は約7%であった。「統計」と「経験」に分類された3編は、いずれも調査研究であった。

以下、具体的な記述内容について順にみていく。まず、ターン他 (2019) では、方法のセクションの中に、「サンプルサイズの決定と実際の調査参加者」という小節が設けられ、下記の記述があった。検定力分析を行う際、想定する統計分析手法や、効果量とその大きさについて根拠と共に述べられていた。データロスを見越して目標数より多くのデータを収集したことも明記されていた。

40項目を用いて確証的因子分析を行うことを前提に、必要となるサンプルサイズを計算した。自由度を722、有意水準 α を.05、検出力を.80としたとき、高程度の効果量($w = .50$; Cohen, 1992)を検出できるサンプルサイズは399名以上であった。調査の途中棄権や不適切な回答など分析の対象とならないデータの存在を考慮して、それよりも多い人数を募集したところ、495名が調査に参加した。(ターン他, 2019, pp. 507-508)

次に、岩本 (大久保) 他 (2020) では、研究3の結果について、観察検定力を計算し、.20と低い値を示したことから、適切なサンプルサイズ設計を行った上で再調査を実施するとし、研究4の方法において調査参加者のサンプルサイズ設計について記している。下記に、研究3の観察検定力に関する記述と研究4のサンプルサイズ設計に関する記述を順に示す。

本研究の検定力を分析するため、R言語 (ver. 3.5.0) のpwrパッケージを用いた検定力分析を実施した。分析では、効果量はstep2のSMM投入によるモデル上昇率である0.01、パラメータ数7、有意水準5%、自由度306と設定した。その結果、本研究の検定力は.20であった。…本研究の検定力は.20と低く、適切なサンプルサイズ設計を行った上で再調査を実施する必要がある。(岩本 (大久保) 他, 2020, p. 599)

サンプルサイズ設計にあたり、効果量に Crum et al. (2013) の研究 1 の人生満足感を目的変数とした階層的重回帰分析におけるストレスマインドセット投入によるモデルの説明率の上昇値 2% を設定した検定力分析 (パラメータ数 7, 有意水準 5%, 検出力 80%) を実施した。その結果、必要なサンプルサイズは 710 人と算出された。なお、計算には、研究 3 と同様の方法を用いた。以上に基づき、20 代から 50 代の男女の就労者 800 人が調査に回答した。

(岩本 (大久保) 他, 2020, p. 600)

観察検定力は、標本効果量を用いて算出されるため、検定結果と観察検定力は基本的に同じことである。小野寺・山本 (2004) が観察検定力について、「その 1 回限りのデータについての記述的な統計量にしかすぎず、将来の推定値とみなしたり、検定力の高低を議論するためには使えない」と述べているように、検定力と観察検定力は異なることには留意が必要である。

「経験」の増田他 (2019) では、調査デザインの観点からサンプルサイズ設計について述べており、データ収集における各条件のストップングルールについて述べているが、設定した人数の根拠は書かれていない。

回答者は、後述する 4 条件のいずれかにランダムに割り当てられ、20 代から 60 代までの各年代の男女別に 25 名に達した時点で調査を打ち切った。したがって両社共に、各条件の回答者は性別と年代が同数の 250 名ずつであり、…性差や年代差は相殺されていると言える。(増田他, 2019, p. 465)

JPR

同様に JPR の 20 編については、「統計」は Ishii et al. (2019) の 1 編、「経験」は Tanaka & Ikegami (2019) と Tayama et al. (2019) の 2 編、残りはすべて「なし」に分類され、サンプルサイズ設計に関する記述がみられた論文の割合は全体の約 15% であった。「統計」の 1 編は実験研究、「経験」の 2 編は実験研究と調査研究が各 1 編ずつであった。

「統計」の Ishii et al. (2019) では、Method のセクションに 22 人の実験参加者のうち、20 人分のデータを分析対象としているとし、脚注で検定力分析によるサンプルサイズ設計について述べている。当該研究のサンプルサイズが一般的な社会心理学の研究のそれに比べ小さいと考えられることに対し、関連研究から比較的大きな効果を想定して検定力分析を行い、目標人

数が 25 人になったことを述べている。また、検定力分析に G*power を使ったことが明記されている。後述の JPSP も含め、海外誌では検定力分析の計算に何を使ったかを明記する論文がみられる。

The sample size in the present study ($N=20$) might be thought to be small compared to general social psychological research standards. Therefore, we conducted a statistical power analysis for sample size estimation. We adopted a very large effect size in terms of the following two points. (a) Based on data from Adam and Galinsky (2012), the effect sizes (η_p^2) in their study were .09, .11, .16, and these scores were considered to be medium or large using Cohen's (1988) criteria. (b) We measured prior scores and used them as covariates. We conducted a priori power analyses using G*power 3.1.9.2 (Faul, Erdfelder, Lang, & Buchner, 2007) with power ($1-\beta$) set at 0.80, $\alpha = .05$, and $f = .06$ (very large effect size). The result indicated that a sample size of 25 would be sufficient to detect a significant effect. (Ishii et al., 2019, p. 126)

「経験」の Tanaka & Ikegami (2019) と Tayama et al. (2019) では、Method のセクションにおいて、先行研究をサンプルサイズ決定の根拠としていることを明記している。

We determined the sample size so that it would be almost equal to those of DeWall et al. (2009) and Tanaka and Ikegami (2015). (Tanaka & Ikegami, 2019, p. 135)

In accordance with a previous study (Arnett, 2000), we defined the participants in this study as young adults aged 18–25 years. (Tayama et al., 2019, p. 264)

JPSP

JPSP では、統計的仮説検定を用いた論文 42 編のうち、サンプルサイズ設計に関する記述がなかった「なし」に分類されたのは 2 編で、それ以外は、「統計」、「経験」、「事後」に分類された。ただし、JPSP は 1 編の論文に含まれる研究数が平均 4.3 個 (標準偏差 2.6) と、他の 3 誌に比べて多く (『教育心理学研究』, 『心理学研究』, JPR の順に平均は 1.2, 1.7, 1.4, 標準偏差は 0.5, 0.9, 0.7), 1

編の論文の中に複数の研究がある場合は、「統計」、「経験」、「事後」の複数のカテゴリを付与した。カテゴリの分類結果は、「統計のみ」は24編、「経験のみ」は6編、「統計と経験」は5編、「統計と事後」は3編、「統計と経験と事後」は1編、「事後のみ」は1編であった。これらから、検定力分析に基づくサンプルサイズ設計に関する記述を含む論文は33編（約79%）あり、*JPS*では検定力分析に基づくサンプルサイズ設計を行う研究が多いことが明らかになった。検定力分析に限らず、なんらかのサンプルサイズ設計に関する記述がみられた論文の割合は約93%を占め、上述の井関（2019）が2017年に刊行された*PBL*と*JEPG*に掲載された論文の割合（41.8%）と比べてもさらに高い値を示した。井関（2019）の結果との比較は調査対象の論文誌が異なるため、慎重に行う必要があるが、海外誌でのサンプルサイズ設計に関する記述の充実化は急速に進んでいると考えられる。

記述内容の下位分類

「統計」または「経験」に分類された論文について、記述内容をさらに細かく14のカテゴリ（「目標人数」、「予備実験」、「文献」、「慣習」、「デザイン」、「外部事情」、「追加」、「ストップングルール」、「効果量」、「信頼区間（Confidence Interval: CI）に基づく設計」、「検定力」、「効果タイプ」、「Gpower」、「その他」）に分類した結果をFigure 1に整理した。14のカテゴリのうち、「ストップングルール」、「信頼区間」

に基づく設計」、「その他」以外は井関（2019）の下記のカテゴリとその定義を踏襲した。

「目標人数」とは、実際に収集したサンプルサイズとは別に、具体的な数値目標に言及している。数値が先行研究や分野内の慣習に基づく場合もあれば、検定力分析に基づく場合もあった。「予備実験」とは、論文の著者または同じグループによる予備実験に基づくことに言及している。効果量の数値に言及している場合もあれば、単に予備実験に基づくことのみを述べた場合もある。「文献」とは、目標人数や効果量の根拠となった具体的な先行研究を挙げている。Cohen（1992）に基づく中程度の効果量といった一般論的な参照は含めず、論文ごとのテーマに即した先行研究が引用されている場合のみ、この項目に該当するものとした。「慣習」とは、当該研究分野の慣習に広く基づくことに言及している。「デザイン」とは、実験デザインによる必要性を考慮してサンプルサイズを決めたことに言及している。カウンターバランスを成立させるための調整などが含まれたが、多くの場合、デザイン上の必要性に触れることなく、セルごとに何名になることを目指したなどの記述がなされていた。「外部事情」とは、実験設備の利用可能性や募集期間などの外部事情による制約に言及している。「追加」とは、何らかの理由から経験や検定力分析に基づくサンプルサイズよりも多く募集したことに言及している。例えば、検定力やデータの安定

Figure 1

*JPS*の論文のサンプルサイズ設計に関する記述内容の下位分類

性・信頼性を高めるため、実験や調査に際してデータロスが予想されるためのなどの理由が述べられた。「効果量」とは、Cohen の d や相関係数などについて具体的な効果量の値に言及している。「検定力」とは、具体的な検定力の値に言及している。「効果タイプ」とは、特定の比較や交互作用など、検定力分析の対象となった効果について限定的に言及している。「Gpower」とは、G*power を使ったことに明示的に言及しているか、Faul らの文献 (Faul et al., 2007, 2009) を引用していることから G*power を用いたことが推測できる場合とした。新たに追加した「ストッピングルール」では、データ収集をどこで止めるかに関して明示されているもの、「信頼区間に基づく設計」では、あらかじめ信頼区間の幅を定め、必要なサンプルサイズを求めているもの、「その他」では、R パッケージを使ったことなどを記しているものとした。

Figure 1 から、記述頻度が最も高いのは「目標人数」、その次に多いのは、「効果量」、「検定力」の順であった。これは、JPSP では検定力分析に基づくサンプルサイズ設計を行う研究が多いことを反映したものと考えられる。「Gpower」も検定力分析に関連して記されることが多かった。「文献」、「デザイン」に関する記述も比較的多くみられた。一方、「信頼区間に基づく設計」や「慣習」に関する記述の頻度は少ないことが明らかになった。

サンプルサイズの報告例

「経験」に分類されたものと、「統計」に分類されたもののなかからいくつか具体的な記述例を下記に示す。

「経験」(検定力分析以外) に基づく記述 Huang et al. (2019) は、データ収集のための被験者プール・方法、研究デザインに関わる制約に基づいてサンプルサイズを実験の実施前に決めたことが脚注に書かれている。しかし、各条件につき少なくとも 50 人とした根拠は書かれていない。

Target sample sizes for each experiment were determined in advance of data collection based on consideration of participant availability (e.g., subject pool size), study design, and collection method. We aimed for at least 50 observations per condition (Simmons, Nelson, & Simonsohn, 2011) for all studies. (Huang et al., 2019, p. 608)

Hudson et al. (2019) は、授業の総履修者数によってサンプルサイズが決定されたことが書かれている。

A total of 377 participants provided at least one wave of data (data were collected for only one semester and sample size was determined by the total enrollment in participating classes). (Hudson, et al., 2019, p. 841)

Jackson & Gray (2019) は、関連する先行研究の知見がなく、検定力分析で特定の効果量を推定することが困難だったことから、500 名のデータを取り、その人数で中効果量と小効果量を想定したときの検定力を求め、述べている。

The lack of prior relevant research made it difficult for us to estimate a specific effect size in our power analysis. We recruited 500 participants from Amazon Mechanical Turk, which gave us 99% power to detect a medium-sized effect of $\hat{f}^2 = .15$ and 71% power to detect a small effect of $\hat{f}^2 = .02$. (Jackson & Gray, 2019, p. 1206)

「統計」(検定力分析) に基づく記述 Wolf et al. (2019) は、検定力分析のために、どの分析手法を想定し、どのような効果量について、どのくらいの効果の大きさを想定したか、関連研究から理論的、実践的意味があると思われる具体的な大きさを述べ、サンプルサイズを計算している。また、除外した被験者の分、目標人数を満たすために、追加で被験者の収集を行ったことが記されている。計算に G*power を使っていることも記されている。

Previous research testing the link between value dissimilarity and prejudice has generally found relatively large effect sizes between $\beta = .42$ and $\beta = .45$ (Dunbar et al., 2000; Haddock et al., 1993; Struch & Schwartz, 1989). However, there is also evidence of weaker effects between $\beta = .00$ and $\beta = .21$ (Schwartz et al., 1990), and importantly, we expected the current polynomial regression approach to provide a more conservative estimate than previous research. Hence, we based the power analysis on a medium effect size ($\beta = .30$), which we reasoned would reflect a theoretically and practically meaningful value dissimilarity effect. Using G Power (Faul, Erdfelder, Lang, & Buchner, 2007), the required sample size was 82 participants for a two-tailed linear regression analysis, a medium effect, a power of .80, and an alpha level of .05. This

power calculation is consistent with Barranti, Carlson, and Côté's (2017) recommended sample size of 77 participants to detect a medium effect in polynomial regression analyses. We recruited additional participants in order to meet this sample size requirement after participant exclusion. (Wolf et al., 2019, pp. 931-932)

信頼区間・効果量の記載状況

統計的仮説検定が用いられた『教育心理学研究』の25編の論文について、検定結果に加え信頼区間や効果量の記載があるかを調べた結果をTable 1の8列目、9列目に示した。その結果、信頼区間・効果量が共に記載されていたのは7編（約28%）、信頼区間のみ記載されていたのは2編（約8%）、効果量のみ記載されていたのは5編（約20%）、どちらも記載がなかったのは11編（約44%）であった。

同様に『心理学研究』、*JPR*、*JPSP*について調べた結果をTable 2にまとめた。『教育心理学研究』は、信頼区間、効果量共に記載がない論文の割合が最も多く（約44%）、信頼区間、効果量を結果に併記するという実践は他誌に比べ浸透していないことが窺える。一方、サンプルサイズ設計に関する記述の実践度が最も高かった*JPSP*では、全ての論文に信頼区間もしくは効果量が結果に併記されており、その両方が記載されている論文の割合（約76%）も高く、他の3誌に比べ統計改革の影響が最も顕著に表れていた。また、概して信頼区間よりも効果量を報告する論文が多いことも明らかになった。APA マニュアル第7版では、3.7 節「量的結果の標準」において、統計的仮説検定をどの程度重視するか（重視しないか）は、異なる分野、個々の学術誌

Table 2

4誌に掲載された論文における信頼区間・効果量の記述状況（論文数と割合）

雑誌名	教心	日心	<i>JPR</i>	<i>JPSP</i>
共にあり	7 (28%)	11 (26%)	5 (24%)	38 (76%)
信頼区間のみ	2 (8%)	0 (0%)	0 (0%)	2 (4%)
効果量のみ	5 (20%)	28 (67%)	15 (71%)	10 (20%)
共になし	11 (44%)	3 (7%)	1 (5%)	0 (0%)
対象論文数	25	42	21	50

注1) 表中の数値は論文数を表す。

注2) 表中の「教心」は『教育心理学研究』を、「日心」は『心理学研究』を表す。

注3) 「共にあり」は信頼区間と効果量の記載がある場合、「共になし」は信頼区間と効果量の記載がない場合を表す。

の編集長の決定によるが、妥当な効果量や信頼区間について、全ての検証された仮説や推定について完全に報告することを、APA 発行の学術誌では必須の要件とすると明記している。また、信頼区間（母数の推定値、平均値差のような母数の関数、効果量などについての信頼区間）を報告することは、研究結果の報告に極めて有効な方法であると述べ、全般的にみて信頼区間は、結果報告において最良の手法であり、その使用を強く薦めると明記されている。*JPSP*には、結果に信頼区間や効果量が高い割合で報告されており、これはAPAの方針が遵守されていることの証左といえる。一方、信頼区間の報告の実践度については、APAの推奨に対し、未達であることが窺える。

4誌を通して、信頼区間や効果量を結果に併記している論文において、それらの値について本文の中で解釈し、述べているものは極めて少なかった。しかし、当該の研究の中で、併記された値がすぐに有効に活用できなかったとしても、先行研究など他の研究の効果量との比較を通して有用な知見が得られる可能性もあるため（南風原, 2018; 石井他, 2013）、効果量や信頼区間を結果の基本的な情報として報告することは望ましいと考えられる。

標本効果量に基づく判断

標本効果量をもとに何らかの判断をすることについて述べておく。正田他（2019）は、B-MMRの各尺度得点を従属変数、性別を独立変数とした対応のない t 検定を行い、データから計算された標本効果量と先行研究で示された基準を比較し、下記のように述べている。

効果量は最大でも $d=0.34$ と、Ferguson（2009）の基準に照らしても大きくないものの、「回復」、「強烈な感覚」、「気晴らし」および「心内作用」の4因子において、男性よりも女性の方が有意に得点が高いことが示された。（正田他, 2019, p. 404）

標本効果量は基本的にサンプルサイズに依存しない値であり、統計量の精度や安定性を反映していない。そのため、その値について判断や解釈をするより、下記に述べる母集団効果量をその信頼区間と共に計算し、それらを判断や解釈に活用することがより有用であると考えられる。

母集団効果量の推定とその信頼区間

小野田・大澤（2020）は、平均値差の t 検定および一要因分散分析の結果として、効果量 d' および ε^2 の95%信頼区間を算出して明記している。これらの値につい

て、本文中で解釈はなされていないが、母集団効果量を推定し、その信頼区間を併記することの意義は大きいと考えられる。APA マニュアル第7版においても、研究で得られた知見の重要性を完全に理解するには、結果のセクションに効果量を加える必要があり、効果量は推定値のため、その精度を示すために、可能な限り“効果量の信頼区間”を示すことと書かれている。効果量の推定値の信頼区間の情報を示すメリットには、大きく下記の2点がある。一つは、信頼区間が0をまたぐか否かで統計的仮説検定の結果も判断でき、かつ、効果の大きさがどれくらいかを、その不確実性の程度（信頼区間の幅）と共に示すことができることである。この性質を利用すれば、望ましい信頼区間の幅からサンプルサイズを決定することができる。もう一つは、効果量の大きさの判断や解釈に関わることである。現在はCohenの基準（Cohen, 1992）に照らして効果の大きさを判断している研究が多くみられるが、今後、効果量の信頼区間と共に算出して報告する論文が増えてくると、効果量の大きさを、実験の要因操作における刺激の差異の大きさと対比で評価したり、他の対象者集団での結果との対比でみたりするところから、効果の大小・強弱についての相対的あるいは絶対的な判断が徐々に可能になっていくことが期待でき（南風原, 2014b）、効果の大きさに関してより意味のある考察や判断ができるようになっていくことである。単に標本効果量を示すだけでなく、母集団効果量を推定し、その信頼区間と共に論文に記すことの有用性は、国内では南風原（2014a, 2014b）や清水（2020）が述べてきたが、その実践度はまだ低く（例えば、山内・菊地, 2016）、今後より広く実践されるようになることが望まれる。

今後に向けた提言

以上をまとめると、まず、『教育心理学研究』と他の3誌について概観した結果からは、下記が明らかになった。1点目は、統計的仮説検定を用いた研究では、サンプルサイズ設計に関するなんらかの記述があった論文の割合は、JPSPは93%、他の3誌では7—15%を示し、報告状況に国内外で顕著な差がみられた。2点目は、サンプルサイズの決定の根拠が多く記述されていたJPSPでは、検定力分析に基づく記述が信頼区間や慣例に基づく記述に比べて多いことが示された。3点目は、JPSPでは大半の論文にサンプルサイズ設計に関する記述があり、反対にそれ以外の3誌では記述があった論文が極めて少なかったため、論文で採用されている研究法とサンプルサイズ設計に関する記述の実

践度との間に特定の傾向を見いだすことができなかった。4点目は、概して、サンプルサイズ設計に関する記述の実践度よりも信頼区間や効果量を結果に併記する実践度の方が高かった。また、信頼区間と効果量の報告状況を比べると、効果量の方が信頼区間よりもより報告されている割合が高かった。5点目は、『教育心理学研究』では、信頼区間や効果量を結果に併記した論文の割合が他誌に比べて最も低く、結果に併記することが広く定着していないことが窺えた。

サンプルサイズの根拠を示すことの意義とその促進策

長年にわたり指摘されてきた統計的仮説検定の恣意性の問題に対する解決策として、「研究のサンプルサイズを根拠をもって決める」ということがある。他にも、統計的仮説検定をそもそも用いない脱検定というより、抜本的な解決策や、昨今、耳目を集めているオープンサイエンス⁵なども考えられる。

今後、それらを積極的に推進していくことが、より一層重要である一方、本稿でみてきた通り、『教育心理学研究』をはじめ、本邦の心理学の論文誌に掲載された研究は、統計的仮説検定を使ったものが圧倒的多数を占めており、その現実を鑑みると、研究者のサンプルサイズ設計への意識を高め、統計的仮説検定を使ってできることを模索し、研究実践を改善していくことが現実的かつ有用であると考えられる。

サンプルサイズ設計とは、必ずしも検定力分析に基づくもののみを指すわけではないと考えられる。JPSPでもみられたように、サンプルサイズの決定の根拠が、当該研究領域の慣習に倣ったものや、研究資金や研究にかけられる労力や時間といった現実的な制約によることもあり、それが実際に論文に書かれていた。そのような根拠でも論文の中に記すことは一定の価値があると考えられる。その理由は、そうした記述があることで、研究の結果やそれに基づく結論に内包される偶

⁵ オープンサイエンスとは、「科学的な研究を市民（非専門家）により開かれた活動へと変革する運動」とされるが、狭義には、より多くの人々が科学研究の基礎となるデータや成果にアクセス可能とすること（オープンデータ）を指すことが多い（矢守他, 2021）。ここでは狭義の意味で用いている。近年、心理学の研究領域においてもこのオープンサイエンスを志向する動きが活発化している。その背景には、著名な研究者による重大な研究不正の発覚（Callaway, 2011）や、研究結果の再現性の低さが問題にされ（Open Science Collaboration, 2015）、科学としての信頼性を再構築するための積極的な施策が急務になったことがある（三浦, 2018）。そのうちの重要な1つが、研究の透明性（transparency）と開放性（openness）の確保であり、具体的には、オープンデータとオープンアクセスの試みである（三浦, 2018）。

然性の影響の大きさについて読者が考えることができ、当該研究のデータの根拠の強さを判断するのに役立つからである。さらに検定力分析に基づいた根拠が示せると、その報告の仕方や記述内容にもよるが、母集団においてどのくらいの大きさの効果をどのような根拠のもとに想定し、それをどのくらいの検定力で検出しようとしていたのかが読者に伝わり、より有用である。ただし、検定力分析で想定する効果については、その定量的な判断が難しいことが多く（村井・橋本, 2018）、検出すべき母集団での効果の大きさを設定することが難しいという問題もある。そのため、検定力分析をしさえすれば、統計的仮説検定の恣意性の問題が完全に解決できるというわけではないが、その点を考慮したとしても、サンプルサイズの根拠を何も記さなくてよいという積極的な理由にはならないと考えられる。

ではなぜ、『心理学研究』や『教育心理学研究』においてサンプルサイズの決定の根拠を論文に記すという実践が広がっていないのか、その要因を探ることは意義があると考えられる。考えられる要因として、(1) そもそもサンプルサイズの決定の根拠を論文に書く必要性（意味や意義）が理解されていない、(2) サンプルサイズの決定は検定力分析で行われなければならないという思い込みがある、それに関連するものとして(3) 検定力分析の具体的なやり方（プログラムの使い方を含め）、計算の仕方がわからない、(4) その重要性は認識しつつも、論文誌に掲載された論文の多くに書かれていない状況を見て同調行動をとっている、(5) 論文の紙面の制約から（書きたくても）書けない、(6) 執筆・投稿の手引きなどに書くことが推奨されていない、などが考えられる。

上記(1)から(3)については、教育や啓発が有効であると考えられ、それらは学会でのシンポジウムや著作物の刊行などによってすでに行われている部分もある。しかし、それらを今後も続け、さらに質、量共に充実化していく必要がある。特に、(3)に関する著作物は医学領域などに比べるとかなり少ない。また、対象者となる読者や聴衆の層を広げていくことも肝要である。心理測定や統計に馴染みのない人にもその必要性や重要性が伝わるような具体的でわかりやすい解説や説明が求められる。心理学を専攻する学生を対象とする場合は、心理統計教育の場で、それらについて学ぶ時間を設けることや、それらを教えることができる人材を確保したり、育成したりすることも重要であるだろう。また、昨今、大学教育の中で注目されているデータサイエンス教育においても、統計的仮説

検定の結果をより深く理解するために有効な情報として、サンプルサイズ設計や信頼区間・効果量などに関しても扱われることが望ましい。(4)については、(1)から(3)と同様に、教育や啓発が有効であるだろうが、同時に、『教育心理学研究』が学術誌として、この問題をどのように考えるのか、スタンスや考え方を明示することも重要であるだろう。(5)については、本文に限らず脚注を活用することも一案である。また、より抜本的には、論文誌の紙面の制約を緩和することや、海外誌でみられるようにオンライン補足資料の検討も有効である。(6)については、(4)にも関わるが、日本心理学会の「執筆・投稿の手引き」において、APAマニュアルのように、サンプルサイズに関する記述として何を記すべきか、推奨事項を明確化することも検討に値すると考えられる。現在、「執筆・投稿の手引き（2015年度改定版）」には、「仮説検定の適用にあたっては、前提とするデータの性質（データの分布の正規性や、標本相互の独立性など）が成立していることを確認する。」と述べられているだけであり、サンプルサイズの根拠を求める記述はみられない。研究者の研究実践を変容させることを考える場合、個人の主体性、内発的動機づけにアプローチするだけでは必ずしも十分とは言えず、研究者の研究活動を取り巻く社会的構造にも目を向け、介入していくことが肝要であると考えられる。今後は、このような研究者コミュニティレベルへのアプローチについても積極的な検討が求められる。日本教育心理学会として上述の(4)から(6)について検討していくことは、掲載される論文の質の向上に寄与すると考えられる。

研究知見のより効果的な蓄積に向けた工夫

本稿で概観してきたように、国内誌においても統計改革の影響は、「はじめに」で述べた大久保（2009）の調査がなされた頃に比べ、徐々にではあるが現れてきており、効果量や信頼区間を結果に併記する論文は増えてきている。これをさらに進め、母集団効果量の推定とその信頼区間と共に結果に併記する実践がより広く定着していくことを期待する。これらを実践する研究が増えていくと、より研究の知見の蓄積が効果的、かつ、効率的に行えるようになって考えられる。母集団効果量の信頼区間の情報は、後続の研究のサンプルサイズ設計に活用することができ、また、メタ分析を促すことにも繋がる。1編の論文から明らかにできることの限界に真摯に向き合い、研究者コミュニティ全体で知見を積み上げ、信頼性や再現性の高い、データの根拠のより強い知見を生み出していく仕組みづくりを積極的に進めていく必要があるだろう。そのため

には、研究者コミュニティの個々のメンバーがその価値を共有し、研究のルーティーンの改善を目指すことが不可欠である。本稿がそうした研究者コミュニティのメンバーの価値共有化の一助となれば幸いである。

引用文献

- 安達知郎・安達奈緒子 (2019). 大学新入生に対するアサーション・トレーニングの効果—適応感とアイデンティティ, 自己受容に注目して 教育心理学研究, 67(4), 317-329. <https://doi.org/10.5926/jjep.67.317>
- American Psychological Association (2020). *Publication manual of the American Psychological Association* (7th ed.). American Psychological Association.
- 芦田祐佳 (2020). 低学年児童の情動面に対する教師の個別支援—単一学級における児童間差の検討 教育心理学研究, 68(2), 202-215. <https://doi.org/10.5926/jjep.68.202>
- Callaway, E. (2011). Report finds massive fraud at Dutch universities. *Nature*, 479(7371), 15. <https://doi.org/10.1038/479015a>
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155-159. <https://doi.org/10.1037/0033-2909.112.1.155>
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41, 1149-1160. <https://doi.org/10.3758/BRM.41.4.1149>
- Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175-191. <https://doi.org/10.3758/bf03193146>
- Finch, S., Cumming, G., Williams, J., Palmer, L., Griffith, E., Alders, C., Anderson, J., & Goodman, O. (2004). Reform of statistical inference in psychology: The case of Memory & Cognition. *Behavior Research Methods, Instruments, & Computers*, 36, 312-324. <https://doi.org/10.3758/bf03195577>
- 南風原朝和 (2002). 心理統計学の基礎—統合的理解のために 有斐閣
- 南風原朝和 (2011). 臨床心理学を学ぶ 7 量的研究法 東京大学出版会
- 南風原朝和 (2014a). 続・心理統計学の基礎—統合的理解を広げ深める 有斐閣
- 南風原朝和 (2014b). 分散分析を基礎から見直す—有意性検定による「推測革命」と近年の「統計改革」(閉じられたANOVAとその先—心理統計の現状と将来を考える, 2013年度日本基礎心理学会第1回フォーラム) 基礎心理学研究, 32(2), 217-222. <https://doi.org/10.14947/psychono.KJ00009351487>
- 南風原朝和 (2018). 心理統計の新しい展開と今後の統計教育 心理学評論, 61(1), 142-146. <https://doi.org/10.24602/sjpr.61.1.142>
- Huang, S. C., Lin, S. C., & Zhang, Y. (2019). When individual goal pursuit turns competitive: How we sabotage and coast. *Journal of Personality and Social Psychology*, 117(3), 605-620. <https://doi.org/10.1037/pspi0000170>
- Hudson, N. W., Briley, D. A., Chopik, W. J., & Derringer, J. (2019). You have to follow through: Attaining behavioral change goals predicts volitional personality change. *Journal of Personality and Social Psychology*, 117(4), 839-857. <https://doi.org/10.1037/pspp0000221>
- 井関龍太 (2019). 実験心理学における例数設計の周辺 基礎心理学研究, 37(2), 180-187. <https://doi.org/10.14947/psychono.37.27>
- 石井秀宗・吉田寿夫・岡田謙介・南風原朝和 (2013). 心理学研究における効果量の活用と報告—APAの指針をふまえて 教育心理学年報, 52, 234-237. <https://doi.org/10.5926/arepj.52.234>
- Ishii, K., Numazaki, M., & Tado'oka, Y. (2019). The effect of pink/blue clothing on implicit and explicit gender-related self-cognition and attitudes among men. *Japanese Psychological Research*, 61(2), 123-132. <https://doi.org/10.1111/jpr.12241>
- 石山裕菜・鈴木直人・及川昌典・及川 晴 (2020). 表現筆記が防衛的悲観主義者のパフォーマンスに及ぼす影響—認知方略による調整効果の検討 教育心理学研究, 68(1), 1-10. <https://doi.org/10.5926/jjep.68.1>
- 岩本 (大久保) 慧悟・竹橋洋毅・高 史明 (2020). ストレスマインドセット尺度の邦訳および信頼性・妥当性の検討 心理学研究, 90(6), 592-602. <https://doi.org/10.4992/jipsy.90.18229>
- Jackson, J. C., & Gray, K. (2019). When a good god makes bad people: Testing a theory of religion and immorality. *Journal of Personality and Social Psychology*, 117(6), 1203-1230. <https://doi.org/10.1037/pspp>

- 0000206
 神谷哲司 (2020). 成人用ファイナンス効力感尺度の作成と信頼性・妥当性の検討 教育心理学研究, 68(2), 160-173. <https://doi.org/10.5926/jjep.68.160>
- 河合輝久 (2019). 大学生における身近な友人の抑うつ症状への情緒的巻き込まれの恐れ—楽観的見通し, 深刻度評価, 専門家への援助要請の必要性との関連 教育心理学研究, 67(4), 289-303. <https://doi.org/10.5926/jjep.67.289>
- 木村敏久・小泉令三 (2020). 中学校におけるいじめ抑止の意識向上に向けた社会性と情動の学習の効果検討—教師による実践及び生徒の社会的能力との関連 教育心理学研究, 68(2), 185-201. <https://doi.org/10.5926/jjep.68.185>
- 吉良悠吾・尾形明子・上手由香 (2020). 青年のソーシャルスキルにおける汎状況的なスキルと具体的な対人場面でのスキルとの関連性の検討 教育心理学研究, 68(1), 11-22. <https://doi.org/10.5926/jjep.68.11>
- Kuramoto, M. (2019). Cultural identity disequilibrium experienced by intercultural parents during their children's elementary school selection and adaptation: A qualitative analysis of interviews with Japanese mothers married to non-Asian fathers. 教育心理学研究, 67(3), 203-220. <https://doi.org/10.5926/jjep.67.203>
- 楠見友輔 (2019). ダウン症児の文章理解の学習—社会文化的アプローチによる一生徒の学習過程の分析から 教育心理学研究, 67(4), 330-342. <https://doi.org/10.5926/jjep.67.330>
- 牧 郁子 (2019). 保護者との情動交流が小学生の無気力感に与える影響—構造方程式モデルによる分析 教育心理学研究, 67(4), 223-235. <https://doi.org/10.5926/jjep.67.223>
- 増田真也・坂上貴之・森井真広 (2019). 調査回答の質の向上のための方法の比較 心理学研究, 90(5), 463-472. <https://doi.org/10.4992/jipsy.90.18042>
- 三浦麻子 (2018). 心理学におけるオープンサイエンス—「統計革命」のインフラストラクチャー 心理学評論, 61(1), 3-12. <https://doi.org/10.24602/sjpr.61.1.3>
- 村井潤一郎・橋本貴充 (2018). 統計的仮説検定を用いる心理学研究におけるサンプルサイズ設計 心理学評論, 61(1), 116-136. <https://doi.org/10.24602/sjpr.61.1.116>
- 永井 智 (2019). 援助要請スタイル間の差異に関する探索的検討—援助要請過剰型・回避型の特徴 教育心理学研究, 67(4), 278-288. <https://doi.org/10.5926/jjep.67.278>
- 長峯聖人・外山美樹 (2019). ノスタルジアが時間的態度に与える影響—本来性を媒介要因として 教育心理学研究, 67(3), 190-202. <https://doi.org/10.5926/jjep.67.190>
- 長峯聖人・外山美樹・湯 立・三和秀平・肖 雨知・相川 充 (2019). 制御焦点とライバル関係との関連—ライバルによる理想自己の顕在化と動機づけの生起を考慮して 教育心理学研究, 67(3), 162-174. <https://doi.org/10.5926/jjep.67.162>
- 中野有美・鋤柄増根・志村尚理・中川敦夫・大野 裕 (2020). こころのスキルアップ教育が新入生に及ぼす影響—Q-Uを用いた中学校での探索的研究 教育心理学研究, 68(1), 66-78. <https://doi.org/10.5926/jjep.68.66>
- 庭山和貴 (2020). 中学校における教師の言語賞賛の増加が生徒指導上の問題発生率に及ぼす効果—学年規模のポジティブ行動支援による問題行動予防 教育心理学研究, 68(1), 79-93. <https://doi.org/10.5926/jjep.68.79>
- 則武良英・武井祐子・寺崎正治・門田昌子・竹内いつ子・湯澤正通 (2020). ハイプレッシャー状況が引き起こすワーキングメモリ課題成績低下に対する短期筆記開示の効果 教育心理学研究, 68(2), 134-146. <https://doi.org/10.5926/jjep.68.134>
- 小野田亮介・大澤和仁 (2020). 受け手に合わせた情報発信を促す指導方法—中学校国語科のタブレットを用いた授業による検討 教育心理学研究, 68(1), 50-65. <https://doi.org/10.5926/jjep.68.50>
- 小野寺孝義・山本嘉一郎 (2004). SPSS 事典—BASE 編 ナカニシヤ出版
- 尾之上高哉・井口 豊 (2020). ブロック練習と交互練習の単独効果と複合効果の比較検討—学習内容の定着度, 及び, 確信度判断の正確性に着目して 教育心理学研究, 68(2), 122-133. <https://doi.org/10.5926/jjep.68.122>
- 大久保街亜 (2009). 日本における統計改革—基礎心理学研究を資料として 基礎心理学研究, 28(1), 88-93. <https://doi.org/10.14947/psychono.KJ00005878671>
- 大久保街亜・岡田謙介 (2012). 伝えるための心理統計—効果量・信頼区間・検定力 勁草書房
- Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716. <https://doi.org/10.1126/science>

- aac4716
 酒井麻紀・窪田由紀 (2019). 小学校教師の職場における援助要請に関連する要因の検討—被援助志向性、問題に対する内的な帰属、協働的風土に着目して 教育心理学研究, 67(4), 236-251. <https://doi.org/10.5926/jjep.67.236>
- 酒井智弘・相川 充 (2020). 感謝表出スキルの実行が孤独感の低減に及ぼす効果 教育心理学研究, 68(2), 111-121. <https://doi.org/10.5926/jjep.68.111>
- 清水裕士 (2020). 再現可能性にかかわる統計的問題と新しい論文のカタチ 第84回日本心理学会シンポジウム 2020.11.1 再現可能な心理学研究入門 <https://norimune.net/wp/wp-content/uploads/2020/12/61d1f3d7cf9c9eefac5412999ccf08b2.pdf>
- 正田 悠・安田晶子・中原 純・田部井賢一・伊坂忠夫 (2019). 短縮版音楽による気分調整尺度 (B-MMR) の日本語版の作成および信頼性・妥当性の検証 心理学研究, 90(4), 398-407. <https://doi.org/10.4992/jjpsy.90.18207>
- 杉本希映・遠藤寛子・飯田順子・青山郁子・中井大介 (2019). 保護者による教師の信頼性認知尺度の開発とその関連要因の検討 教育心理学研究, 67(3), 149-161. <https://doi.org/10.5926/jjep.67.149>
- 竹下 浩・藤田紀勝 (2019). 就労移行支援員による利用者の就労スキル発達支援過程 教育心理学研究, 67(4), 265-277. <https://doi.org/10.5926/jjep.67.265>
- Tanaka, H., & Ikegami, T. (2019). Social exclusion and disengagement of covert attention from social signs: The moderating role of fear of negative evaluation. *Japanese Psychological Research*, 61(2), 133-141. <https://doi.org/10.1111/jpr.12197>
- Tayama, J., Schaufeli, W., Shimazu, A., Tanaka, M., & Takahama, A. (2019). Validation of a Japanese version of the work engagement scale for students. *Japanese Psychological Research*, 61(4), 262-272. <https://doi.org/10.1111/jpr.12229>
- ターン有加里ジェシカ・橋本剛明・シュミット, マンフレッド・唐沢かおり (2019). 公正感受性尺度日本語版 (JSI-J) の作成 心理学研究, 90(5), 503-512. <https://doi.org/10.4992/jjpsy.90.18234>
- 登藤直弥 (2019). 『教育心理学研究』における測定・評価・研究法の現状と今後に向けての提言 教育心理学年報, 58, 119-130. <https://doi.org/10.5926/arepj.58.119>
- 外山美樹 (2020). 楽観性は適応的なのか?—下位目標が達成されたときに焦点を当てて 教育心理学研究, 68(2), 174-184. <https://doi.org/10.5926/jjep.68.174>
- 外山美樹・湯 立 (2019). 大学運動部活動における部員の自律的動機づけが部活動への適応感に及ぼす影響—主将のリーダーシップを調整変数として 教育心理学研究, 67(3), 175-189. <https://doi.org/10.5926/jjep.67.175>
- 脇 貴典・須藤邦彦 (2020). 特別支援教育コーディネーターへの集中トレーニングによるコンサルテーションスキルの獲得と維持 教育心理学研究, 68(1), 33-49. <https://doi.org/10.5926/jjep.68.33>
- 山田真世 (2020). 幼児における絵の表現調整の発達—他者とのミスコミュニケーション場面での検討 教育心理学研究, 68(1), 23-32. <https://doi.org/10.5926/jjep.68.23>
- 山田洋平・小泉令三 (2020). 幼児を対象とした社会性と情動の学習 (SEL-8N) プログラムの効果 教育心理学研究, 68(2), 216-229. <https://doi.org/10.5926/jjep.68.216>
- 山内香奈 (2017). 社会心理学研究におけるサンプルサイズ設計 村井潤一郎・橋本貴充 (編著) 心理学のためのサンプルサイズ設計入門 (pp. 107-128) 講談社
- 山内香奈・菊地史倫 (2016). 鉄道従業員向けアナウンス研修の転移促進手法に関する実験的検討 教育心理学研究, 64(1), 131-143. <https://doi.org/10.5926/jjep.64.131>
- 矢守克也・飯尾能久・城下英行 (2021). 地震学のオープンサイエンス—地震観測所のサイエンスミュージアム・プロジェクトをめぐる 実験社会心理学研究, 60(2), 82-99. <https://doi.org/10.2130/jjesp.2009>
- 吉田琢哉・吉澤寛之・浅野良輔・玉井颯一・吉田俊和 (2019). 社会化エージェントが社会的認知バイアスに及ぼす影響—親の養育, 教師の指導, 友人の非行, 地域の集合的有能感を指標とした検討 教育心理学研究, 67(4), 252-264. <https://doi.org/10.5926/jjep.67.252>
- 吉田琢哉・吉澤寛之・浅野良輔・玉井颯一・吉田俊和 (2020). 地域に根ざした学校運営におけるチームワーク尺度の作成—保護者の視点からの検討 教育心理学研究, 68(2), 147-159. <https://doi.org/10.5926/jjep.68.147>
- Wolf, L. J., Weinstein, N., & Maio, G. R. (2019). Anti-immigrant prejudice: Understanding the roles of (perceived) values and value dissimilarity. *Journal of Personality and Social Psychology*, 117(5), 925-953. <https://doi.org/10.1037/pspi0000177>