【研究ノート】

心理学における効果量の活用とその具体例

津田　恭充*

Using effect size in psychology citing specific examples

Hisamitsu Tsuda

要　旨

多くの心理学系の学術誌では、帰無仮説検定の弱点を補うために効果量や信頼区間を報告することが推奨されている。本稿では、効果量の目安、Cohen の目安が適用できないとき、効果量の具体的な活用例をまとめた。

Abstract

Many psychological journals recommend reporting effect sizes and confidence intervals to compensate for the weaknesses of null hypothesis significance testing. This paper discusses the rough standards of effect size posited by Jacob Cohen for application when standards are not pertinent and cites specific examples of the use of effect sizes.

● ● ○ **Key words**　効果量 effect size／帰無仮説検定 null hypothesis significance testing

1　心理学における帰無仮説検定と効果量

この数十年、帰無仮説検定は心理学を含む多くの学問分野で研究の主張に統計学的根拠を与える道具として活用されてきた。例えば、日本における心理学の学術専門誌である「心理学研究」に掲載された論文の中での帰無仮説検定の使用率は、1948 年には 5% 前後であったが、1980 年代以降はおおよそ 90% 前後で推移している[1)]。現在でもこの傾向に変化はなく、2020 年の「心理学研究」に掲載された論文（90 巻 6 号から 91 巻 5 号までの計 37 本。展望論文は集計から除外した）では、そのすべてで何らかの形で帰無仮説検定が用いられている。しかし、有意水準という客観的かつ明確な基準で差や関連を検討できる帰無仮説検定は万能の道具ではないことは繰り返し指摘されている[2)3)4)]。本稿では、その帰無仮説検定の欠点を補うための有力な手段である効果量とその活用法について、いくつかの具体例を挙げながら論じる。

受付日　2020. 9. 11 ／ 受理日　2021. 1. 13
*関西福祉科学大学　心理学部　准教授

2 *d* 族の効果量とその目安

男女の身長と体重を比べたところ、男子のほうが女子よりも平均して10センチ身長が高く、体重は10キログラム重かったとする。このとき、男女の身長と体重の差はどちらも10であるので身長と体重の性差は同程度である、とは当然いえない。身長と体重の単位が異なるからである。こうした異なる単位をもつ数値を比較可能なように標準化したものが効果量である。これまでに数多くの効果量が考案されているが、それらは差の大きさを表す *d* 族と関連の強さを表す *r* 族とに大別できる。

d 族の代表は Cohen's *d*（以下，単に *d* とする）や Hedges' *g* であるが、いずれも基本的には2群の平均値の差を標準偏差で割ったもので、要するに平均値の差が標準偏差いくつ分あるかを示す。$d=1.0$ ならば標準偏差ひとつ分の差があるということである。サンプルサイズが小さい場合、*d* には無視できないバイアスがかかるが、それを補正したものが Hedges' *g* である。ただし、ややこしいことに、計算ツールやソフトウェアの中には、*d* という表記で Hedges' *g* を出力するものがあり、論文に *d* として記載されている効果量も中身は *d* ではなく Hedges' *g* であることが多い。つまり、*d* と Hedges' *g* はしばしば同一視ないしは混同されている。また、Hedges' *g* をさらに補正したものや、より簡潔な式で近似的にその補正値を求める方法もあり、それらも含めて名称に混乱が生じているが[5]、本稿では先行研究を引用する際にはその文献に記載されている名称をそのまま引用することとする。

ところで、標準偏差でいくつ分の差といっても具体的にそれがどれくらいなのかは実感しにくいだろう。そのような場合のために、Cohen（1988）は心理学を含む行動科学における効果量の目安を大中小で提案している[6]。それによると、小さな効果量が $d=0.2$、中程度の効果量が $d=0.5$、大きな効果量が $d=0.8$ である。*d* は2つの分布がどの程度重なるかの指標ともいえるが、それを視覚的に表したものが図1である。図1に示したのは、標準偏差が1、片方の群の平均得点を0にした正規分布（標準正規分布）である。図1(a) は $d=0.2$ のときの分布の重なりを表すが、このとき、ふたつの分布は85.2%重なっている。同様に、$d=0.5$ のときは67.0%、$d=0.8$ のときは52.6%の重なりとなる。参考として、表1に $d=3.0$ までの分布の重なりをまとめた。

分布の重なり具合から考えると、$d=0.2$ はわずかな差といってよいであろう。一方、大きな効果量とされる $d=0.8$ では分布の重なりは約半分である。これよりも *d* が大きければ、もはや重なる部分のほうが小さくなりそうだという意味では $d=0.8$ は大きな効果量の目安としてわかりやすい。$d=0.8$ の具体例として Cohen（1988）は、大学の新入生と博士学位取得者の IQ の差や、13歳と18歳の女子の身長差を挙げている。

日本人を対象としたデータとして、2019年度学校保健統計調査のデータ[7]を元に計算すると（各群の正確なサンプルサイズが不明であるため、便宜的に両群とも同じサンプルサイズであると仮定して計算した。非常に大規模な調査であることを考慮すると、両群のサンプルサイズに多少の違いがあっても計算結果にはほとんど影響がないと考えられる）、女子の12歳（平均身長151.8センチ、標準偏差5.86）と14歳（平均身長156.5センチ、標準偏差5.31）の身長差の効果量は Hedges' $g=0.84$ である。また、身長の性差は誰の目にも明らかであるが、17歳の身長（男子の平均身長170.6センチ、標準偏差5.87、女子の平均身長157.9センチ、標準偏差5.34）の性差は Hedges' $g=2.26$ である。これは一目瞭然の差を表す効果量として一応の目安にはなる。

医療分野ではしばしば $d=0.5$ 前後が臨床的に意味のある差として登場する。ある特定の領域においてQOLにどの程度の差があれば臨床的に意味があるといえるのかを示す概念として最小重要差（MID : Minimally Important Difference）という用語があるが、Norman et al.（2003）は38の論文のシステマティックレビューを行い、$d=0.5$ を慢性疾患における MID としている[8]。この基準を用いている研究は多い。また、コンピュータ認知行動療法について So et al.（2013）は16の論文についてシステマティックレビューを行い、その効果量は $d=0.48$ であったと報告している[9]。

心理学では"Many Labs" replication project[10][11]という大掛かりな研究が行われた。これは影響力のある心理学論文について36の研究機関が参加して国際的に大規模な再現実験を行ったもので、図2はその結果の

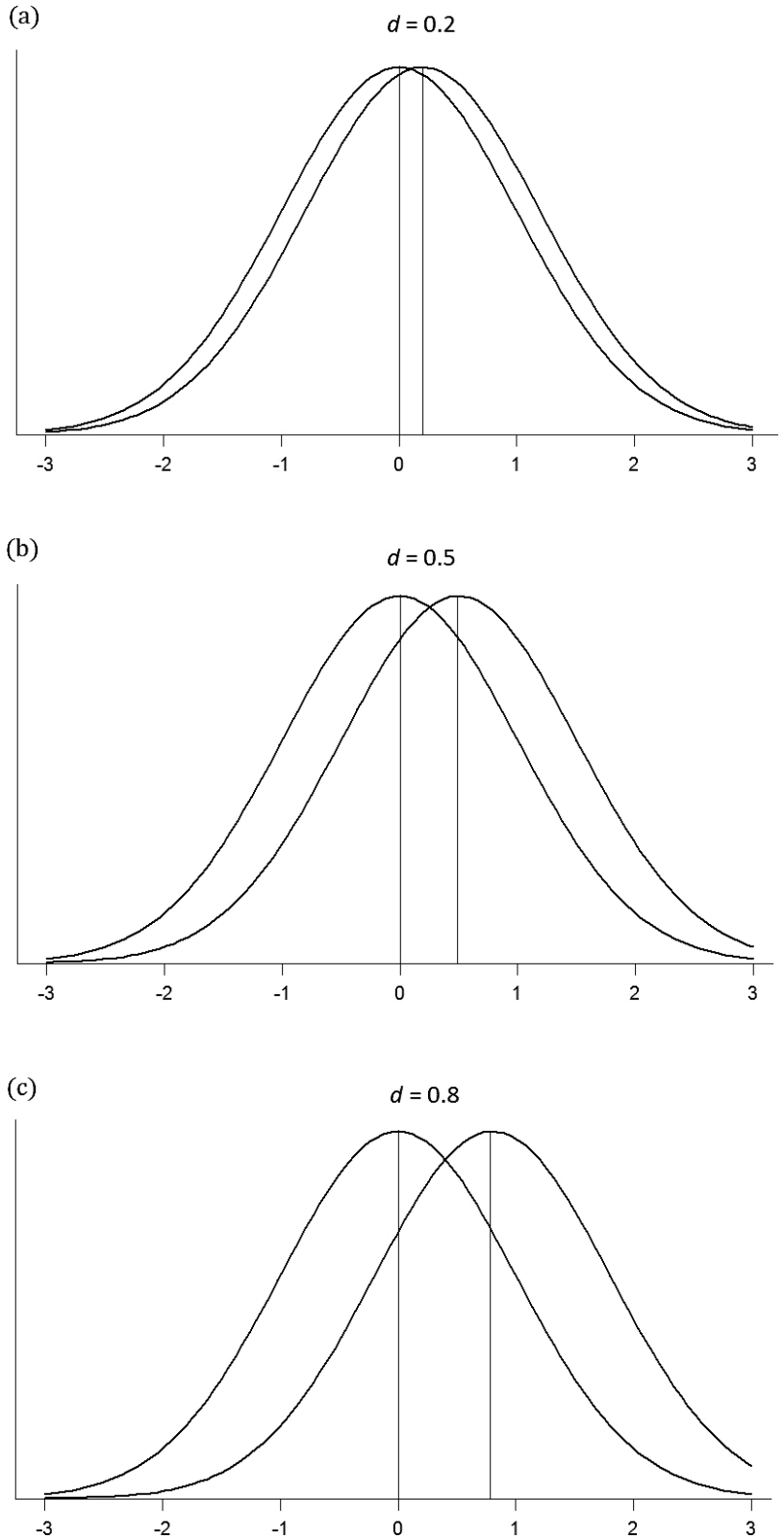

図 1　*d* の大きさと分布の重なり

表 1　*d* の大きさと分布の重なり

d	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
分布の重なり（%）	92.3	85.2	78.7	72.7	67.0	61.8	57.0	52.6	48.4	44.6
d	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
分布の重なり（%）	41.1	37.8	34.7	31.9	29.3	26.9	24.6	22.6	20.6	18.9
d	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9	3.0
分布の重なり（%）	17.2	15.7	14.3	13.0	11.8	10.7	9.7	8.8	7.9	7.2

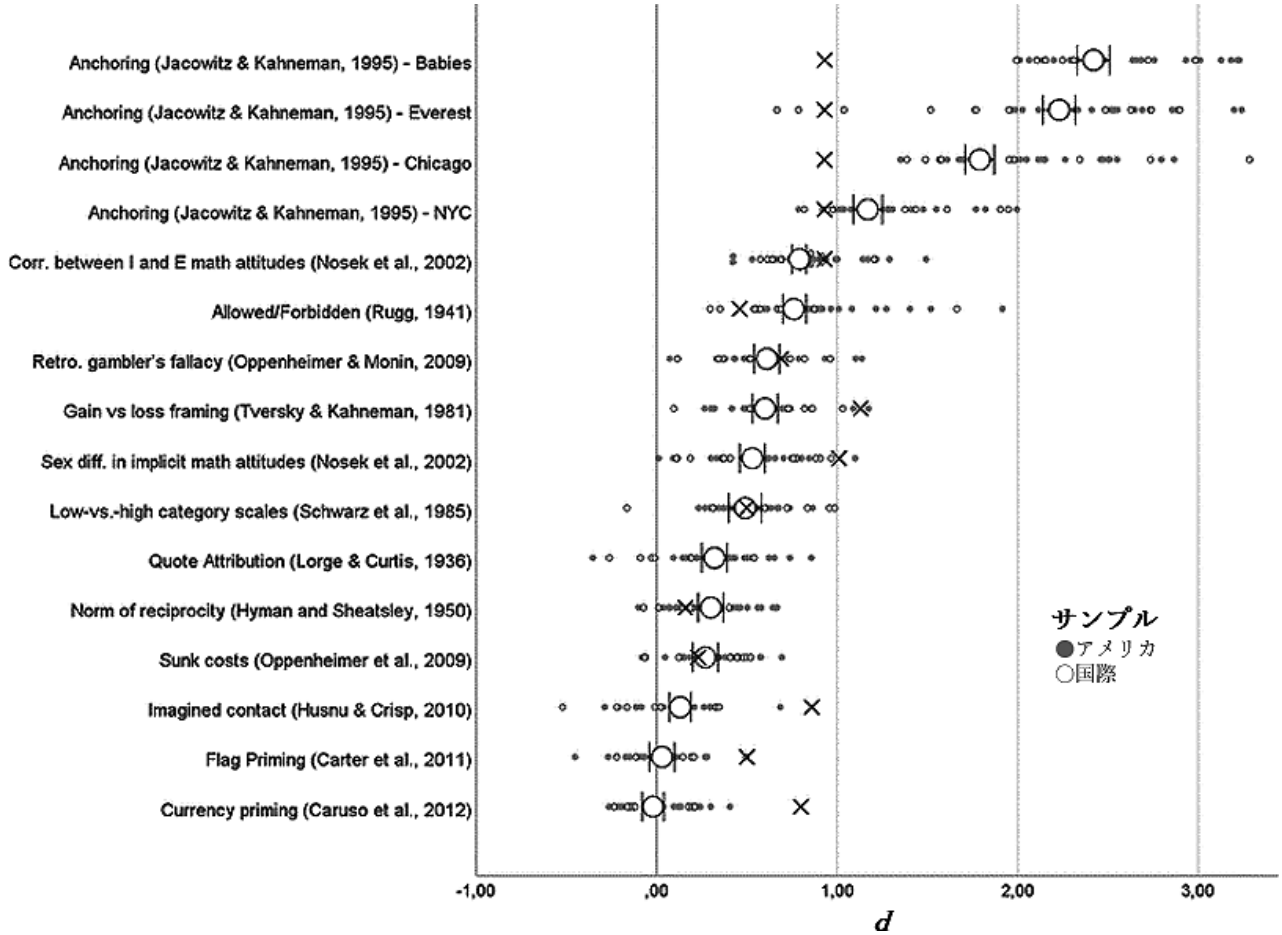

図 2　再現実験ごとの効果量（Klein et al.（2014）の図を一部改変）

注）×はオリジナルの論文で報告された効果量、小さな○は研究機関ごとの効果量、大きな○はすべての研究機関を総合した全体的な効果量を意味する。

概要の引用である。縦軸に再現実験の対象、横軸に d が描かれている。×はオリジナルの論文で報告された効果量、小さな○は研究機関ごとの効果量、大きな○はすべての再現実験参加者を総合した全体的な効果量を意味する。アンカリング（anchoring）に関してはいくつかの種類の再現実験が行われているが（図 2 の上から 4 つ分）、全体的な効果量はオリジナルの研究の効果量よりもかなり大きい。しかも、研究機関ごとの効果量もほとんどが $d=1$ 以上を示していることから、再現性が高いうえに効果量も非常に大きい現象であることがわかる。他の再現性の高かった実験の全体的な効果量はおおよそ $0.3<d<0.8$ に収まっている。その後、“Many Labs 2”[12]やさらに大規模な再現研究プロジェクト[13]が実施され、再現されない現象が多いことも明らかになったが、再現性の高い現象に焦点を当てると、$d>1.0$ の効果量を示した現象は少なく、その意味では $d>1.0$ という効果量は心理学研究の文脈では大きいものであるといえる。

3　r 族の効果量とその目安

r 族の代表は相関係数 r と決定係数 R^2 である。これらはおそらくもっとも身近な効果量で、どのような研究の文脈でどの程度の数値であればどの程度の意味をもつのかおおよそのイメージをもっている研究者は多いであろう。しかし、他の効果量と異なり、r や R^2 はそれ自体が効果量であるため、効果量と認識して報告する研究者は少ないという指摘[14]もある。

目安として、Cohen（1988）は $r=.10$ を小さな効果量としている。$r=.30$ は中程度の、$r=.50$ は大きな効果量とされる[15]。なお、d は t 値と自由度（df）を用いて以下の式で簡単に r に変換できる[6]。

$$r=\sqrt{\frac{t^2}{t^2+df}}$$

この変換のためのエクセルファイル[16]も公開されており、誰でも利用可能である。d は理論上は上限がない

のに対して、r は 0〜1 の範囲に値をとるため解釈しやすい。また、r に揃えることでメタ分析等が行いやすくなるため、r の使用を推奨する研究者[17]もいる。

R^2 は分散説明率とも呼ばれ、回帰分析ないしは重回帰分析において独立変数が従属変数のばらつきをどの程度説明できるかを表す。$R^2=0.25$ であれば、独立変数によって従属変数のばらつきを 25% 説明できることを意味する。大きさの目安は $R^2=0.02$ が小さな、$R^2=0.13$ が中程度の、$R^2=0.26$ が大きな効果量である[18]。

量的変数同士の相関を求める場合には上記の r を用いるが、質的変数と量的変数の相関を求める場合には η（イータ）を用いる。η は相関比と呼ばれ、本質的には r と同じものである。分散分析ではしばしば効果量として η^2 や η_p^2（偏イータ二乗）を報告するが、これは分散分析が質的変数と量的変数の関連を調べる分析であることによる。η^2 の解釈や大きさの目安は R^2 と同様に考えることができる。例えば、一要因の参加者間計画の分散分析であれば、η^2 はその要因によって従属変数のばらつきをどの程度説明できるかを表す。ただし、η^2 や η_p^2 は研究デザインによっては問題も生じるため、η_G^2（一般化イータ二乗）などの効果量も提案されており、状況に応じてこれらを使い分けるのが望ましい[19]。

その他にも、対応があるときの分析やノンパラメトリック検定で用いる効果量などもあり、それぞれに大きさの目安も提案されているが、それらについては他の文献[5][18]に詳しいので参照されたい。

4　Cohen（1988）などによる効果量の目安が適用できないとき

ここまで主な効果量とその目安についてまとめてきたが、ここで示した目安はあくまでも心理学を含む行動科学における目安である。態度がどの程度その後の行動を正確に予測するかを調べる（例：有権者のもつ候補者への態度を測定し、その後の投票行動との関連を調べる）とき、$r=0.99$ という予測精度はありえない高さであるが、化学や物理学ではそれでも不十分であることは珍しくなく、行動科学とは基準がまったく異なる。また Cohen 自身が注釈を加えているように、行動科学の中でも研究内容や目的に応じて効果量の目安は柔軟に運用すべきである。例えば、未開拓の分野で実験手法や介入方法が十分に洗練されていない場合には、$d=0.2$ も十分に意味がある効果量とみなせるかもしれない。ただし、臨床的な実用化を目的とした研究で、初期の研究よりも大きな効果を目指している場合には、はじめに $d=0.2$ を十分に意味のある効果量とみなしたからといってその基準をいつまでも守りつづけることは適切ではないだろう。MID に関しても「特定の領域における」という条件がついていることからわかるように、すべての疾患に同じ基準が適用できるわけではない。

ここで改めて図 2 をみると、報告されている効果量は小さなものから大きなものまで現象ごとに幅があることがわかる。したがって、関連する先行研究が十分に存在する場合には、機械的な目安よりも先行研究との比較を通じて効果量の解釈を行うほうが有用であることも多いだろう。例えばアンカリングに関する研究では、一般的な手続きを用いた場合には $d=0.5$ を中程度の効果量とみなすのは適当ではなく、アンカリングとしてはむしろ小さい効果量であると解釈するほうが正確であると考えられる。

個人を問題にしているのか、より大きなレベルを問題にしているのかどうかも効果量の解釈に影響を与えるように思われる。例えば、日本国民全員から 1 円寄付してもらえば 1 億円以上になるというように、ちりも積もれば山となるような集積的効果を狙う場合には、小さな効果量にも大きな意味があるといえるかもしれず、その場合は Cohen（1998）をはじめとした効果量の大きさの目安は妥当でない可能性がある。

他の効果量と異なり、r は効果量に関する議論が活発になる以前から頻繁に用いられてきた。そのため、前述したように、r に関しては同程度の効果量であっても研究の文脈によって解釈の仕方が異なることがすでに十分認識されているように思われる。例えば、抑うつとある種のパーソナリティの関連が $r=0.5$ であれば大きな関連といえそうであるが、時間的に安定的であることを仮定するパーソナリティ検査の再検査信頼性が $r=0.5$ であったら小さすぎる数値とみなされるであろう。これと同じように、r 以外の効果量についても、今後は機械的な目安の適用を超えた臨機応変な解釈がなされるのが一般的になっていくだろうと予

測される。

5 効果量の活用とその具体例

アメリカ心理学会[20]や日本心理学会[21]をはじめ、多くの学会や学術誌が論文の投稿規定に効果量の報告を推奨ないしは義務としたこともあり、効果量の記載はかなり一般的になってきた。実際に、冒頭に示した2020年の「心理学研究」に掲載された37本の論文を調べると、36本の論文で帰無仮説検定の結果とともに効果量が記載されている。その意味では効果量の記載はもはや当然のこととなっているといえる。ただし、効果量を仮説設定や考察に直接生かしていたのは、そのうちの8本のみであった。やや古いデータであるが、メジャーな教育心理学系の国際誌を対象にした調査[22]では、得られた効果量について考察していた論文は約半数であったとしている。これらの結果は、投稿規定に定められているからという消極的な理由で効果量を求めている場合が少なくないことを示唆している。たしかに、比較対象となる先行研究やデータがない、探索的な研究である、その効果量が現実世界においてどのような意味をもつのか解釈しにくいというような場合には効果量の扱いが困難であることも多い。また、後続の研究にとっては論文中に効果量が記載されているだけでも有益な情報となる。したがって、効果量の大きさについて言及していないからといって即座に論文の価値が低下するわけではないが、具体的にどのような研究でどのように効果量を活用するかについては整理された情報が少なく、それが効果量が仮説設定や考察に活用されにくい一因になっているように思われる。そこで、以下にいくつかの具体例を挙げながら効果量の活用方法についてまとめた。わかりやすさのためにいくつかのパターンに分類したが、これは便宜上の分類であり互いに排他的ではない。研究内容によっては重複する場合もありうる。

一つめは、現実場面への応用に際して、それが意味のある効果かどうかを把握するための研究である。例えば抑うつを題材とするとき、ある介入が1%水準で有意に抑うつを改善するのか5%水準で有意に抑うつを改善するのかよりも、効果量が大きいのか小さいのかという問題のほうが臨床的には重要である。こうした医学的治療や心理学的援助のほか、教育や政策の効果などに関する研究もここに含まれる。メタ分析（およびシステマティックレビュー）やMIDに関する研究もこの文脈に位置する。本邦における最近の研究として、岸田（2020）は、児童青年の不安症と抑うつ障害に対する、回避行動に焦点化した診断横断的介入プログラムのフォローアップの有効性を、Hedges' g を用いて先行研究と比較しながら論じている[23]。

二つめは、効果がないことを明らかにする研究である。帰無仮説検定における有意確率はサンプルサイズが大きいほど小さくなるので、大規模なデータであれば実質的に無意味な差や関連でも有意になる。また、帰無仮説検定はその原理上、「差や関連がある」という結論は導けても、「差や関連がない」という結論は導けず、厳密には「差や関連があるとはいえない」という結論を出せるにとどまる。したがって、帰無仮説検定は効果がないことを積極的に確かめる手段としては無力である。一方、効果量にはこうした問題はないため、差や関連があるかどうか疑わしい現象に対しては効果量を活用するのが有効である。具体例として、縄田（2014）は3つの大規模なサンプル（合計で一万人以上）を対象に調査を行い、血液型と性格の関連を検討した[24]。縄田（2014）はかなり多くの性格に関する項目を分析しているが、もっとも大きな効果量を示した項目でさえも $\eta^2 = 0.0027$ と極小であったことから、血液型と性格は無関連であると結論づけている。このように、非常に小さい効果量であることを根拠に差がないとか関連がないことを積極的に主張できるのも効果量の強みである。ただし、サンプルサイズが小さいと効果量の信頼区間が広がりやすく強い主張がしにくいため、差がないとか関連がないことを主張するには、サンプルサイズを大きくしたり効果量の点推定値だけでなく効果量の信頼区間も報告するといった工夫が必要である。

三つめは、再現性に関する研究である。帰無仮説検定では有意か有意でないかという2値データしか得られないが、効果量ならばより細かい情報を得られる。“Many Labs”のような大規模な追試を行う場合、この違いは特に大きなものとなる。例えば図2をみると、アンカリングはオリジナルの研究よりもむしろ追試のほうが大きな効果量を報告しているが、一方で、仮想接触（imagined contact）やプライミング（prim-

ing）では追試の効果量は小さく、オリジナルの研究で報告された効果量に届いているものはない。このように、複数の独立したサンプルを用いて追試を行い、その効果量を報告することで、単に現象が再現されたかどうかだけでなく、「当初言われていたよりも強い効果がありそうだ」とか「オリジナルの研究はたまたま高い効果が出ただけなのではないか」というように、より具体的な議論が可能になる。

四つめは、サンプルサイズ設計である。前述したように、帰無仮説検定ではサンプルサイズが大きすぎると実質的に無意味な差や関連でも有意となる。反対に、サンプルサイズが小さすぎると、差や関連があってもそれを正しく検出できなくなってしまう（第二種の過誤）。サンプルサイズが大きい場合、検定の結果のみに頼らず、効果量やその信頼区間も報告すれば問題ないといえるが、コストのかかる研究や研究参加者の負担が大きい研究である場合には倫理的に問題となる可能性がある。こうした理由から、ほどよいサンプルサイズを事前に設計したい場合が出てくるが、それは、効果量、有意水準、検定力を定めれば算出できる。絶対的な基準ではないものの、慣例[6]にしたがって、有意水準を 5%、検定力を 80% とすると、あとは効果量を決めれば必要なサンプルサイズを算出できる。効果量を定めるのはときとして難しいが、類似した先行研究からの推測、予備調査や予備実験の結果、MID、Cohen（1988）の目安などを元に設定することができる。計算には統計解析ソフトウェア R[25]の pwr パッケージなどが利用できるほか、G*Power[26]という GUI ベースのソフトウェアが多くの分析に対応しており便利である。

これら以外にも、異文化比較、国際比較、年齢や性別ごとの比較、先行研究との比較などにももちろん効果量は活用できる。また、同一研究内で複数の効果を比較する方法もある。津田（2020）は、「自分のことを自分の名前で呼ぶ女性は自己愛が強い」というステレオタイプが強くみられる（$p<.001$, $\eta^2_G=0.307$）ことを確認し、それが実際に正しいのかどうかを検証した。その結果、たしかにそのステレオタイプに合致した傾向がみられた（$p=0.031$, $\eta^2_G=0.037$）。しかし、両者を比較するとステレオタイプの効果量のほうがはるかに大きいことから、人々（特に自分のことを「わたし」と呼ぶ女性）は自分のことを自分の名前で呼ぶ女性の自己愛を実際よりもかなり誇張して認知していると結論した[27]。このような議論は有意か有意でないかのみを問題にする帰無仮説検定では不可能であり、効果量の強みを活用したものであるといえる。

6　効果量だけを報告すれば十分か？

ここまで述べてきたように、効果量にはさまざまな利点や活用方法があるが、効果量を報告しさえすれば十分なのかというとそうでもない。例えば、サンプルサイズが非常に小さければ偶然に大きな効果量を得られることがある。仮に実験群（20, 20, 30, 30, 60）、対照群（20, 20, 30, 30, 30）というデータを用意すると、Hedges' $g=0.49$ という効果量が得られる。ここでは何らかの理由で実験群のひとりだけおかしなデータになった可能性も考えられるが、効果量だけを提示すれば、両群に意味のある差があると強引に主張することもできてしまう。この問題を解決するための最適ではないかもしれないがもっとも簡単な方法は、効果量の分析と帰無仮説検定を併用することである。実際に Welch の t 検定を行ってみると、上記 2 群の差は 5% 水準で有意ではない（t（4.88）$=0.78$, $p=0.47$）。効果量の信頼区間を求めることも有効な方法である。R では compute.es パッケージや rpsych パッケージなどで効果量の信頼区間を算出できる。GUI ベースのソフトウェアとしては日本発の HAD[28]があり、多くの分析に対応している。

7　まとめ

具体的な効果の大きさを知りたい臨床家や研究者にとって効果量は非常に有益な情報である。また、差や関連がないことを主張したり、サンプルサイズ設計の際に用いるなど、効果量を活用できる場面は幅広い。一方で、機械的に効果量の大きさを解釈したり、効果量の点推定値のみに基づいて議論を進めることには危険も伴う。こうした危険を避けるため、近い将来、単に効果量を報告するだけでなく、何の目的でそれを報告するのか、どのようにそれを解釈するのかについても説明することが論文投稿者の義務となるかもしれな

い。筆者を含めた統計学を専門としない統計ユーザーもその時代に備える必要があるだろう。

謝辞

本研究は JSPS 科研費 20K03428 の助成を受けた。

【引用文献】

1）Omi, Y., & Komata, S（2005）. The evolution of data analyses in Japanese psychology. *Japanese Psychological Research, 47,* 137-143.

2）Johnson, D. H.（1999）. The insignificance of statistical significance testing. *Journal of Wildlife Management, 63,* 763-772.

3）Schmidt, F., & Hunter, J.（2002）. Are there benefits from NHST? *American Psychologist, 57,* 65-66.

4）Amrhein, V., Greenland, S., & McShane, B.（2019）. Scientists rise up against statistical significance. *Nature.* Nature Publishing Group.

5）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.

6）Cohen, J.（1988）. *Statistical power analysis for the behavioral sciences*（*2nd ed*.）. Hillsdale, NJ : Lawrence Earlbaum Associates.

7）文部科学省　2019 年度学校保健統計調査　https : //www.e-stat.go.jp/stat-search/file-download?statInfId = 000031925054&fileKind = 0（参照 2020-12-1）

8）Norman, G. R., Sloan, J. A., Wyrwich, K. W.（2003）. Interpretation of changes in health-related quality of life : The remarkable universality of half a standard deviation. *Medical Care, 41,* 582-592.

9）So, M., Yamaguchi, S., Hashimoto, S., Sado, M., Furukawa, T. A., & McCrone, P.（2013）. Is computerized CBT really helpful for adult depression? -A meta-analytic re-evaluation of CCBT for adult depression in terms of clinical implementation and methodological validity. *BMC Psychiatry, 13,* 113.

10）Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., . . . Nosek, B. A.（2014）. Investigating variation in replicability : A “many labs” replication project. *Social Psychology, 45,* 142-152.

11）“Investigating variation in replicability : A ‘many labs’ replication project” : Correction to Klein et al.（2014）.（2019）. *Social Psychology, 50,* 211-213.

12）Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., … Bahník, Š.（2018）. Many Labs 2 : Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science, 1,* 443-490.

13）Open Science Collaboration（2015）. Estimating the reproducibility of psychological science. *Science, 349,* aac 4716.

14）Alhija, F. N.-A., & Levy, A.（2009）. Effect size reporting practices in published articles. *Educational and Psychological Measurement, 69,* 245-265.

15）Hemphill, J. F.（2003）. Interpreting the magnitudes of correlation coefficients. *American Psychologist, 58,* 78-79.

16）http : //www.mizumot.com/stats/effectsize.xls（参照 2020-12-1）

17）Field, A.（2005）. *Discovering statistics using SPSS*（*2nd ed*.）. London : Sage Publications.

18）水本　篤・竹内　理（2008）. 研究論文における効果量の報告のために――基礎的概念と注意点――　関西英語教育学会紀要英語教育研究，*31,* 57-66.

19）井関龍太（2013）. 実験心理学者にとっての効果量　専修大学心理科学研究センター年報，*2,* 33-54.

20）American Psychological Association.（2009）. Publication Manual of the American Psychological Association（7th ed.）. Washington, DC : American Psychological Association.

21）日本心理学会機関誌等編集委員会（2015）. 執筆・投稿の手引き　日本心理学会.

22）McMillan, J. H., & Foley, J.（2011）. Reporting and discussing effect size : Still the road less traveled. *Practical Assessment, Research & Evaluation, 16,* article 14.

23）岸田広平・石川信一（2020）. 児童青年に対する診断横断的介入のフォローアップの有効性の予備的検討　心理学研究，*91,* 63-68.

24）縄田健悟（2014）. 血液型と性格の無関連性――日本と米国の大規模社会調査を用いた実証的論拠――　心理学研究，*85,* 148-156.

25）R Development Core Team.（2020）R : A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria.

26）Faul, F., Erdfelder, E., Lang, A., & 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.

27）津田恭充（2020）. 自分のことを自分の名前で呼ぶ女性は自己愛的か？――ステレオタイプと実際のパーソナリティ――日本パーソナリティ心理学会第 29 回大会発表論文集，78.

28）清水裕士（2016）. フリーの統計分析ソフト HAD：機能の紹介と統計学習・教育，研究実践における利用方法の提案　メディア・情報・コミュニケーション研究，*1,* 59-73.