心理学の再現可能性

友永雅己・三浦麻子・針生悦子

2016 年（心理学評論）

1　友永・三浦・針生 (2016)

Open Science Collaboration (2015)

過去の心理学研究論文について追試を行ったところ、結果が統計的に再現されたものは追試実験全体の 40%に満たない。

Trafimow & Marks (2015)

Basic and Applied Social Psychology 誌のエディトリアルで、今後一切統計的検定に関する記載を行わないと宣言。

Wasserstein & Lazar (2016)

アメリカ統計学会 (ASA) における p 値に関する見解の表明

2　池田・平石 (2016)

2.1　研究法上の問題点

p hacking (Simmons, Nelson, & Simonsohn, 2011)

どのような荒唐無稽な仮説であっても、支持する報告をすることができる。

- 行った条件や測定した変数の一部しか報告しない
- 参加者を少しずつ足しながら分析を行い、有意差に至ったところで止める
- 様々な共変量を用いて分析を行い、有意になった組合せのみを報告する

p hacking を行うことで、少なくとも何らかの分析で有意差が見つかる可能性（Type I Error 率）は 61%に上がる。

Questionable Research Practices (QRPs) (John, Loewenstein, & Prelec, 2012))

統計的妥当性の疑われる研究手法を行っている心理学者は、回答者の半数以上[1]。

2.2　なぜ問題のある研究実践が行われるのか？

心理学の理論の「弱さ」[2]

前提となる知見の根拠が乏しく、また知見ないし仮説間の相互依存性も少ないため、厳密な事前の予測が難しい。

[1] 心理学者 5,000 名を調査し、2,000 名から回答。
[2] Eysenck (1985)

審美的な観点から研究を見る[3]

- 結果の一貫性
  研究内での一貫した結果が望ましいとされがちであるが、検定力が低い場合はむしろ有意になった実験のみを報告している可能性を示唆する [4]

- 物語性
  仮説検証に伴う処々の前提がしっかりと裏付けられていないので、仮に仮説を支持しない結果が得られたとしても、それが仮説を反駁したのか明確にわからない（仮説検証式の物語に耐えられない）。

- 新規性
  「弱い理論」においては新規な真の仮説を生み出すことが難しい

現状の心理学においては、仮説検証型物語よりも、記述的研究によって新規性を探索すべき[5]。

メモ

物語性について 再現府可能性を手続き的不備に求めるのは、実証主義に共通した限界である。心理学に限ったことではない。

新規性について 新規な真の仮説を生み出すことの難しさは、仮説の「弱さ」とは関係ないのでは？強い理論であれば、それだけそこから予測できることは限られてくるわけで、それこそ新規性を生み出すのは簡単ではない。

メモここまで

2.3 HARKing

データを得た後に、それに適合する仮説を構築すること（Hypothesizing After the Results are Known: HARKing[6]）

第 1 種の過誤が増大してしまう[7]。

2.4 偽発見率（false discovery rate: FDR）

陽性と判断されたもののうち、偽陽性の割合[8]

$$FDR = \frac{[\text{False Hypothesis Rate}] \times \alpha}{[\text{False Hypothesis Rate}] \times \alpha + [\text{True Hypothesis Rate}] \times \text{Power}} \tag{2.1}$$

$$= \frac{[\text{False Significance Rate}]}{[\text{False Significance Rate}] + [\text{True Significance Rate}]} \tag{2.2}$$

$$\left.\begin{array}{l} \text{Weak Theories} \rightarrow \text{High FDR} \\ \text{Aesthetic Standards} \rightarrow \text{Publication Bias} \end{array}\right\} \rightarrow \text{QRPs, HARKing} \tag{2.3}$$

[3] Giner-Sorolla, 2012
[4] Schimmack (2012)
[5] Rozin (2009)
[6] e.g., Bones, 2012; Kerr, 1998
[7] 検定の多重性の問題が見えてこない。つまり、有意差が報告された結果が、あたかも当初から予定されていた唯一の分析であるかのように見える。
[8] Benjamini & Hochberg, 1995; Sterne & Smith, 2001

Ioannidis (2005)

- こうした構造により加算的に FDR が引き上げられていくと、科学研究のほとんどが偽陽性であるかもしれない
- 歴史的にも、存在しない現象を真実だと信じ、証明しようと努力をし続けた事例は数多くある（「**無の領域」null field**）

メモ

差や相関が全くないということはそもそもあり得ない。つまり、誤って有意となることは考えられないので（偽陽性はゼロ）、FDR は常にゼロであると思う。

メモここまで

2.5 解決策

論文ガイドライン

Society for Personality and Social Psychology, Psychonomic Society, Psychological Science で

- 統計分析の結果報告
  - 検定力分析、サンプルサイズ決定法、効果量、信頼区間の報告
- QRPs の禁止
- 事前登録やデータ公開

がガイドライン化されている。

その他、Transparency and Openness Promotion (TOP) Guideline の制定による学会誌の格付け（どれくらい厳しい基準を採用するかを学術誌が選択）[9]

研究の事前登録制度

- 目的、方法（サンプルサイズ、デザイン、分析手法）を投稿し、査読
- 得られた結果がどのようなものであろうと掲載される

Open Science Framework[10]による事前登録制度が用いられるようになっている。

直接的追試

- Psychfiledrawer
- Curate Science

3 山田 (2016)

3.1 追試に対するインセンティブ

追試を Psychfiledrawer に登録されるだけでは、インセンティブがわかない。

→ **被追指数（replicitations）**指標の提案[11]

追試研究の投稿を明確に奨励しているか、あるいは追試専用のセクションを用意している雑誌[12]

[9] Nosek et al., 2015

[10] Center for Open Science (COS)

[11] Maniadis, Tufano, & List (2015)

[12] *Journal of Experimental Psychology: General, Archives of Scientific Psychology, Perspectives on Psychological Science, PLoS ONE, PeerJ, Quantitative Methods for Psychology* 等

3.2 チームでの研究推進

個別の技能に優れたものによって構成されたチームで研究を行う。特に、統計家による認知研究への参加を。

4 森口 (2016)

- 乳幼児研究では、結果の再現性よりも結果の解釈が問題になることが多い
- 長期縦断研究を追試することは可能なのか？[13]

5 大久保 (2016)

5.1 心理学における再現可能性

- *Perspectives on Psychological Science* (2012) における特集
- Brian Nosek らによる大規模な追試 [14]

5.2 第 1 種の誤り

Simmons, Nelson, & Simonsohn (2011)

本来差がないデータセットでも、データの取得ごとに検定を行うと 22%のケースで有意差が見られた[15]

Cramer, van Ravenzwaaij, ... , & Wagenmakers (in press)

多要因の ANOVA で探索的に検定を行うと、第 1 種の誤りは 14%になる。

5.3 効果量と p 値

過去の心理学研究について調査したところ、

- 報告されていた p と d には負の相関があるが、$d = .5$ のときは $p = .001 \sim .05$ まで、$p = .01$ のときは $d = .02 \sim 1.0$ までばらついていた [16]
- p と効果量が一致しないケースは日本でも報告されている [17]

→ p 値に依存した解釈が効果量を無視したものになるのであれば、結果の再現性を低めるのではないか。

メモ

効果量を考慮したところで、その推定精度が考慮されていなければ、再現可能性は担保されないままではないのか？

メモここまで

5.4 p 値の分布

過去の心理学研究について調査したところ、

- 報告されていた p の分布は、.05 よりわずかに小さいところでスパイク状に増加する [18]

[13]「非認知的能力（社会情緒的能力）特に幼児期や児童期における自己制御能力が青年期における学力や友人関係、成人期における年収や社会的地位、健康状態、犯罪の程度を予測する」
[14]Open Science Collaboration (2015) in *Science*
[15]池田・平石 (2016) （1 ページ）でも言及。ただし、Murayama, Pekrun, & Fiedler (2014) によると、$.05 < p < .10$ のときのみデータ収集を継続すると、偽陽性率は 7.1%まで下がった。
[16]Wetzels et al. (2011)
[17]波多野・吉田・岡田 (2015); 本記事でも社会心理学研究を対象に同様の傾向を見いだしている。
[18]Masicampo & Lalande (2012); 本記事でも社会心理学研究を対象に同様の傾向を見いだしている。

→ 近年出版への圧力が高まってきていることによって生じているのでは？[19]

メモ

検定力分析が一般になってきたために生じてきた、とは解釈できないのか？

メモここまで

5.5　帰無仮説検定への過度の依存により再現可能性が低下する

- 標本サイズが p 値に与える影響
  標本の大きい研究で無理矢理有意にさせる → 標本の小さい追試では再現できない（有意にならない）
- 検定の繰り返し
- p 値と効果量の不一致

5.6　例数設計（サンプルサイズ決定）

APA のマニュアルでは、例数設計の手続きを明示することを求めている。

- 正確度分析
  事前に設定した CI に収まるよう [20]
- 適応的な停止規則
  事前に設定した規則に達した段階でデータ収集を停止

5.7　ベイズ統計学

p 値からベイズファクターを用いた判断への転換[21]

- ベイズファクターを用いた場合の判断基準は仮説検定に比べて極めて厳しい
- 有意水準を .005 あるいは .001 にするくらいの厳しさがないと十分な再現可能性が担保されない

メモ

そもそも、ベイズファクターの基準自体が定まっていないのでは？
また、例数設計に関して言えば、検定力分析に比べて CI にもとづく方が極めて厳しくなる（大きいサンプルが要求される）。

メモここまで

6　藤島・樋口 (2016)

- 再現性を示す基準については、議論の余地がある[22]
- 本論文では、「十分な検定力の下で元研究と同一手続きで追試をしたときに、元研究と同方向の有意な効果が認められれば、再現可能性があるとみなす」ことにする

[19] Leggett et al. (2013)
[20] Altman et al. (2013)
[21] Johnson (2013)
[22] Open Science Collaboration (2015)

メモ

Anderson & Maxwell (2016) in *Psychological Methods, 21*(1) でも、何をもって結果が再現されたとするかについては（検定結果だけでなく）より広い見方が可能で、何をめざすのかによって再現研究（追試）の方法や分析方法も異なってくる、と指摘している。

メモここまで

7 渡邊 (2016)

7.1 研究者効果や実験者効果

オリジナルの実験状況の中に実験結果に強く影響する要因があるが、それが研究者に認識されていないために論文には記載されない → 結果の再現性が低まる[23]

メモ

Makel & Plucker (2014) in *Educational Researcher, 43*(6) においても、研究グループが異なると追試の成功率が低いことが指摘されている。

> 教育分野の研究では、テーマの新規性にばかり注意が向けられて、先行研究の追試が極めて少ない。また、異なる研究グループによる追試は結果の再現性が低いことも指摘されている。他の研究分野では、研究データや手続きの開示など結果の再現可能性を高める様々な取り組みがすでに行われている。新規性を追い求めることよりも、より真実に迫ることに注意を向けるべきだ。

メモここまで

7.2 再現性の高さは何によって判断するのか？

- 39%という再現性[24]は、本当に不十分なのか？
  現象の重大性、申告制、対処可能性によっても大きく変わるのでは（つまりデータの外側にある何らかの基準に照らし合わせなければ判断できない）
- 心理学の研究では 30%前後の数値が「足りないもの、少ないもの」と解釈されることが多く、これを「心理学係数」と名づけてはどうか

7.3 心理学ではデータがエビデンスではなくデモンストレーションとして機能してきた

社会心理学黄金時代の実験

Milgram の実験、Asch の線分組合せ実験、Zimbardo のスタンフォード監獄実験

- 仮説検証のためではなく、研究者の理論（思想・イデオロギー）から予測されることが実際に起きるということをデモンストレーションするため
- 社会に対して何らかの示唆や警告を与えることを重視
- こうした仮説は、常識や様々な知識、過去の経験からデータ以前に想像がつくものであった
- 実験で実証されなくても、仮説自体が完全に否定されるわけでもない

→ 追試が余り行われず、再現可能性の問題に余り注意が払われなかった。

[23] 澤・栗原 (2016); 西阪 (2001)
[24] Open Science Collaboration (2015)

デモンストレーションからエビデンスへ

- 「大きな仮説」「大きな差」は一通り検証し終わり、「小さな仮説」「小さな差」を検証するようになってきた
- データがないと仮説の真偽の判断がつかなくなってきた

→ 心理学のデータが科学としてのエビデンスの役割を期待されるようになり、再現可能性問題を生み出した

7.4 再現性の高さを目的にすべきではない

- 必要な再現性の大きさは研究対象や目的、方法によって異なる
- 低い再現性が検出されることの方が現象を妥当に反映している可能性もある

→ 心理学が「科学」の特定の基準に自らを適合させる必要はない。

8 平井 (2016)

8.1 医学臨床研究

- プロトコールの作成が必須
  どんな人を対象に、どんな介入を行い、何と比較し、どのようなアウトカムをどのくらい改善するのか[25]
- プロトコールを作成した段階で、研究の知的作業の 8 割は終わっている

→ 事前に得られるべき効果量を設定、サンプルサイズの計算、PECO で定式化したシンプルで明確なリサーチクエスチョンの重要性、研究計画段階でのレビューの有用性

9 三中 (2016)

9.1 統計学の誤用

Fisher (1953)

王立統計学会の会長就任講演にて：

> 実験終了後に統計学者に相談を持ちかけるのは、統計学者に、単に死後診察を行って下さいと頼むようなものである。

HARKing の事例と弊害

9.2 統計学におけるパラダイムシフト

- 過去 80 年にわたって、統計理論は Neyman-Pearson 竜の意思決定パラダイムに支配されてきた[26]
  仮説の「真偽」を判断する「強確証／強反証」[27]
- データを仮説に対する「証拠」とみなす「尤度パラダイム」の提唱[28]
  データを証拠として仮説の相対的な「支持」の強弱を判定する「弱確証／弱反証」立場[29]

[25] Patients, Exposure, Comparison, & Outcomes: PECO （福原, 2008）
[26] Royall, 1997
[27] Sober (1988)
[28] Royall, 1997
[29] Sober (1988)

アブダクション

「真偽」ではなく、相対的な「支持」の順位を踏まえ、その時点で最もよい仮説を選ぶ。

- 前提 1：観察データ D がある
- 前提 2：ある仮説 H はデータ D を説明できる
- 前提 3：H 以外の全ての対立仮説 H' は H ほどうまく D を説明できない
- 結論：仮説 H を最良として受け入れる

→ 仮説に対する推論に終わりはなく、新たに追加されたデータや新たな仮説との比較により、推測が覆される可能性は常に残されている。

メモ

いかにもプラグマティズムな考え方である。

メモここまで

10　武田 (2016)

Open Science Collaboration (2015) については、Gilbert et al. (2016) が反論となるコメント論文を出している。

11　佐倉 (2016)

生命科学においても、*Nature* に掲載された医学／生命科学領域の論文は 70%以上が結果が再現できなかったという指摘がある[30]。

11.1　科学者の行動規範

CUDOS

科学者が無私の精神を発揮して真理を追究することが科学の駆動原理である[31]

Communalism（知識の公共性）, Universalism（普遍性）, Disinterestedness（利害への無関心）, Organized Skepticism（組織的懐疑主義）

PLACE

科学者の行動規範はもっと利己的である[32]

Proprietary（知識の独占）, Local（局所性）, Authoritarian（権威主義）, Commissioned（権力からの委託）, Expert work（専門家主義）

[30] Wadman, 2013
[31] Merton, 1973
[32] Ziman, 2000

11.2 再現性の高さ

- 生存確率が 40%であれば、その生命体（知識の集合体）が生き残るには十分
- 40%を不十分と考える判断は、再現性に重みを置きすぎているのではないか
- 複雑で微妙な現象を対象とする場合には、再現性の確保を優先しすぎると、知りたい事柄を見逃す可能性があるようだ（第 2 種の過誤）

メモ

再現性の確保を優先することと第 2 種の過誤が増加することとの関係についてはよくわからない。

メモここまで

12 特集号全体を通じての印象

12.1 押さえておくべき事実

読むべき論文

- Open Science Collaboration (2015) 心理学研究の再現可能性
- Gilbert et al. (2016) 上記への反論
- Wasserstein & Lazar (2016) ASA による p 値に対する見解
- Simmons, Nelson, & Simonsohn (2011) p hacking
  - Murayama, Pekrun, & Fiedler (2014)
- Wetzels et al. (2011); 波多野・吉田・岡田 (2015) 効果量と p 値の関係
- Masicampo & Lalande (2012) p 値のスパイク状の増加
- Altman et al. (2013) 正確度分析
- Anderson & Maxwell (2016) 結果の再現性に対する定義
- Makel & Plucker (2014) 教育実践研究における追試の少なさ、他グループの追試の成功率の低さ

確認しておくべき事項

- p_{rep} と再現可能性
- 検定力分析、正確度分析と再現可能性
- 交差妥当化による再現可能性のチェック

12.2 同意できる点

- 有意差がないという結果も事実ではあるのだから、事前登録制度の導入やデータ公開[33]はお蔵入り問題を避けるためにも極めて重要である
- 追試に対するインセンティブを高めるために、追試指数を導入したり、追試専用のジャーナルやセクションを設ける[34]ことも重要だろう

[33] as 池田・平石 (2016)
[34] as 山田 (2016)

- 高度な知識や技能が要求されるようになった今、研究を各専門家で分担すること[35]も有用だと考えられる
- 心理学研究におけるデータはデモンストレーションとして機能してきた[36]という主張は確かに同意できる

12.3 検証すべき点

12.3.1 「再現可能性」の定義が曖昧である

- 何をもって再現できたとするのか？[37]
- 再現可能性は何%であれば十分なのか？（40%は低すぎるのか？）[38]
- 再現可能であった論文の中にも、偽陽性（再現できないはずなのにできてしまった）ものが含まれているはず

そもそも「統計的有意性」自体が大した情報をもたらさないのであるから、それが繰り返されるかどうかを持って再現可能かどうかとするのは意味がないように思える。

12.3.2 統計的根拠の希薄さ

- p hacking、QRPs、HARKing、偽発見率の高さは再現可能性を低下させるのか？[39]
- 標本サイズ、検定の繰り返し、p 値への依存、効果量の無視は本当に再現可能性を低下させるのか？[40]

12.3.3 偽陽性と偽陰性のいずれを深刻だと考えるのか？

p hacking や HARKing、偽発見率など、基本的には偽陽性を問題としているようだ。

しかし、帰無仮説が真であることは実質的にあり得ないのだから、偽陽性よりも偽陰性の方が問題なのではないか？[41]

また、偽陽性は出版され追試が失敗することで、それが偽陽性だということが明らかになる。一方で、偽陰性はお蔵入りになる可能性が高く、そもそも追試の対象にすらならない。

あるいは、真偽ではなく「尤もらしさ」を議論するということであれば、QRPs によって「尤もらしさ」が意図的無意図的に左右されていると考える方がよい。

偽陽性 本来ないはずの効果を QRPs によって無理矢理引き出す
（予知能力の存在、血液型と性格の関連）

- 大きすぎる標本、検定の繰り返しによってたまたま有意となったものが報告された
  → 偶然得られたものであれば、追試は（正しく）失敗する可能性が高い
- 標本の系統的偏りや妥当性の低い測度、独立性からの逸脱の無視、層別相関などによって見いだされた
  → 直接的追試では、同じ分析上の誤りを犯すことで、（誤って）成功するかもしれない

偽陰性 本来あるはずの効果が信頼性の低い測度、検定力不足によって見いだせない
（非認知的能力と認知的能力の関係、心理療法の効果の差）

- 効果があることを主張したく、またそれがインパクトを持つなら、そもそも（有意でないという意味で）正しい結果であっても、研究として世に出ていない
  → 追試の対象とならない（まずは元研究の検定力を挙げよ）or なったとしても、（誤って）有意差なしとなる

[35] as 山田 (2016)
[36] as 渡邊 (2016)
[37] as 藤島・樋口 (2016)
[38] as 渡邊 (2016); 佐倉 (2016)
[39] To 池田・平石 (2016); 三中 (2016)
[40] To 大久保 (2016)
[41] 予知能力がある、というのは偽陽性かもしれないが。

- 等価であることを主張したく、検定力不足によって（有意差なしという意味で）誤った結論になった → 直接的追試であれば、同じく検定力不足によって（誤って）有意差なしとなり、その意味で成功するだろう

以下の表に示すように、元論文の結果が偽陽性もしくは偽陰性である場合、そもそも元論文の結果が再現されること自体が誤りである。再現可能性が高いことが望ましいのは、元研究が真陽性・真陰性の場合のみであり、そうかどうかは誰もわからない。

Table 12.1: 偽陽性・偽陰性と追試の成功・失敗

	追試	
元研究	陽性	陰性
偽陽性	元研究の結果が系統的誤差由来 再現はできるが誤った結果	元研究の結果が偶然誤差由来 再現はできないが正しい結果
	（※陰性であることが仮説でないとそもそも公表されない）	
偽陰性	追試の検定力が高い 再現できないが正しい結果	追試の検定力が低い 再現できているが誤った結果

再現性には 2 つの側面がある

- 偽陽性・偽陰性の論文がたくさん存在し、追試が再現されないことによってそれを正しく見ぬいている
- 元研究と同じ系統的誤差、低い検定力の追試が行われれば、結果が再現できたとしてもそれは誤りである

つまり、再現性の高さ自体を目的視することは渡邊 (2016) の述べるように危険である。

偽陽性・偽陰性の論文が多いこと自体は問題であったとしても、結果が再現できないことによって偽であることが明らかになるのは、コミュニティ全体としては健全。

直接的追試によって明らかになるのは、元研究が偶然誤差由来で偽陽性となっていたことというくらいであろう。

12.3.4 HARKing は QRP か？

- データを見てから仮説を構成することが誤りであるのであれば、適合度指標を用いたモデルの比較や探索は誤りと言うことになるのか？
- 正の相関関係を予想していて曲線的な相関が得られた場合、適切な手法で曲線性を示すことは好ましくないことなのか？[42]

池田・平石 (2016) の言うように心理学の理論が「弱い」もので、仮説がどうとでも立てられそうなものであるとするならば、データにモデルを当てはめる「記述」がむしろ適切な方法ではないのか？[43]

三中 (2016) の指摘するように科学がアブダクションによって進み、仮説は常に更新され続けるのであれば、データに仮説を適合させることは悪いことではないのではないか？

そもそも、心理学の仮説はただちに、そして常に、真偽が定まるようなものではないはず。目の前の現象（データ）をよりムリなく説明するものが理論であり、モデルである。

[42] 正の相関関係を予測していたが…と記載すれば済む話？

[43] これは、池田・平石 (2016); Rozin (2009) のいう記述的研究のこと？