

心理学における再現可能性危機：問題の構造と解決策¹⁾

池田 功毅¹, 平石 界²

¹ 中京大学大学院／日本学術振興会特別研究員 PD

² 慶應義塾大学

The reproducibility crisis in psychology: Its structure and solutions

Koki IKEDA¹ and Kai HIRAISHI²

¹Chukyo University/JSPS Research Fellow

²Keio University

Psychological science is now facing an unprecedented crisis of reproducibility. The field is becoming aware of the systematic problems embedded in its research practices that have been widely employed by most academic journals. An emphasis on aesthetic rather than scientific standards has led to a publication bias for positive results, which, in turn, has encouraged questionable research practices (QRPs), such as *p*-hacking and HARKing. These processes have potentially created “null fields” where many findings are mere products of false positives. This risk is especially large in fields where the prior probability of the hypotheses being true is low. In fact, a recent large-scale replication project reported that the reproducibility of psychological literature is less than 40%. The psychology community is starting to respond to this crisis by becoming aware of the importance of pre-registered replication, and by reforming the publication standards of many journals. In this paper, we provide an overview of the facts and solutions to the present problems.

Key words: reproducibility, replicability, questionable research practices, pre-registration, direct replication

キーワード：再現可能性, 疑わしい研究方法, 事前登録制度, 直接的追試

近年, 人間には未来予知能力が存在するという, 一聴すれば非科学的と思われる論文が社会心理学のトップジャーナルに掲載され (Bem, 2011), 社会心理学における重要な諸知見の追試が次々に失敗し (Doyen, et al., 2012; Galak, LeBoeuf, Nelson, & Simmons, 2012; Gomes & McCullough, 2015; Hagger et al., 2014; Johnson, Cheung, & Donnellan, 2014; Ritchie, Wiseman, & French, 2012; Shanks et al., 2013, 2015), また大規模な追試プロジェクトの結果, 社会心理で 25%, 認知心理で 50% ほどしか結果の再現ができないことが報告された (Open Science Collaboration, 2015)。これらの報告に代表されるように, 現在の心理学は, その方法論的基礎に大きな問題を抱えていると考えられている²⁾。本稿では, この問題の構造と, 現在までに提案されている解決策について, その概略を紹介する。

1. 問題の構造

1.1 *p* Hacking と QRPs

まず統計的有意水準と研究方法に関する問題点について述べる。なお以下では, 心理学で広く採用されている帰無仮説検定を基に議論するが, ベ

1) 文字数の制限のため本稿に記載できなかったが, 当該トピックについて重要だと思われる論点については, 以下で追加的に議論している。http://bit.ly/1T4RhWj

2) 確かに個々の追試の成否についての判断は難しい場合もあり得るが (Asendorpf et al., 2013; Maxwell et al., 2015; Simonsohn, 2015a, 2015b, 2016; Valentine et al., 2011; Verhagen & Wagenmakers, 2014), 近年盛んに行われ始めた追試研究を見渡せば, 全体として再現可能性危機が存在することは異論の余地がない (Schwarzkopf, 2016)。その原因を, 追試実験者の技量不足のみに求めるのは (Baumeister, 2016; Bloom, 2016; Lieberman, 2014; Mitchell, 2014), 議論として無理があり (Lakens, 2016; Neuroskeptic, 2014a, 2014c), 未知の変数が介在したためとする議論にも (Cesario, 2014; Dijksterhuis, 2014; Feldman Barrett, 2015; Locke, 2015; Stroebe & Strack, 2014), 明らかな限界がある (Gelman, 2015; Neuroskeptic, 2015; Simons, 2014)。

イズ統計でも同様の問題が生じると考えられている (Simonsohn, 2014)。

心理学の研究プロセスにおいて大きな意味を持つのが p 値, すなわち有意水準 α である。周知の通り, 有意水準 α (通常 5% 未満) は偽陽性 (false positive) すなわち第一種の過誤 (type I error) を防ぐために設けられているものである。しかし現状, p 値が研究の質を担保する客観的水準として機能しているとは言い難い。 p 値は容易に操作可能であるためである (藤島・樋口, 2016)³⁾。

Simmons, Nelson, & Simonsohn (2011) は, 現在多くの心理学ジャーナルで要請されている形式を満たしつつも, p 値を有意水準未満に導くこと, すなわち “ p hacking” が十分可能であり, その結果として, どのような荒唐無稽な仮説であっても, それを支持するデータを論文として報告できることを, 実験報告とシミュレーションを通じて示している。 p hacking の具体例としては, (1) 行った条件や測定した変数の一部しか報告しない, (2) 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, (3) 様々な共変量を用いて分析を行い, 有意になった組み合わせのみを報告する, といったものがある。彼らのシミュレーションによれば, これら「研究者の自由度 (Researcher degrees of freedom)」を組み合わせることにより, 少なくとも何らかの分析で有意差が見つかる可能性, すなわち第一種の過誤が生じる可能性は, 当初有意水準 α の設定によって意図されていた 5% から, 約 61% にまで上がってしまう。

このような結果が生じる理由は二つ考えられる。まず, 変数, 条件や共変量の選択などの可能性を網羅的に探索することによって, 検定あたり, ないしは実験あたりの多重比較の問題を無視して分析を行ってしまうためである。第二に, 検定力 (ある効果が真に存在したとして, その効果を当該研究によって検出できる確率) が低い小さなサンプルサイズの実験では p 値の変動が大きい, 検定を行いながらデータを足していった場合, 偶然によって効果量を実際よりも大きく見積

もり, 統計的有意差を発見する確率が高まるためである (Bakker, van Dijk, & Wicherts, 2012; Button et al., 2013; Ioannidis, 2008; Maxwell, 2004; Schimmack, 2012)。

John, Loewenstein, & Prelec (2012) は, これら統計的妥当性が疑われる研究手法に「問題のある研究実践 (Questionable Research Practices, あるいは QRP) s)」という名前を与え, さらにアメリカ主要大学の心理学者約 5,000 名 (うち回答者は 2,000 名) を対象とした調査を行った。結果, 回答者の半数以上の心理学者が QRP を行っており, また彼らの多くがそうした行為を特に問題だとは考えていないことが浮かび上がってきた。すなわち p hacking は可能であり, そして少なからぬ心理学者がそれを (意図の有無は別として) 実践してきたのである。先に p 値が客観的水準として機能しているとは言い難いと述べた所以である。

さらに, こうした QRP の遂行を「自覚」することは容易でない (Gelman & Loken, 2014)。たとえ研究者が良い結果を探そうという意図を明確に持たず, 実際にたったひとつの分析しか行わなかった場合でも, その方法選択がデータを見た後に行われたものであったとしたら, 期せずして多重比較の問題を侵犯してしまっている可能性があるからである。その理由は, 心理学理論の弱さにあると思われる (Eysenck, 1985; Gigerenzer, 2010; Rozin, 2009)。Eysenck によれば, Newton の重力理論や Einstein の相対性理論などのような「強い」理論では, 頑健な実験結果等の根拠に基づく多くの知見が仮説検証の前提となっており, さらに各仮説間には強い相互依存性があるため, ひとつの仮説 A は, 他の仮説 B や前提となる知見による強い制限を受けており, そのため厳密な予測を事前に行うことが可能である。それに対して, 心理学で見られるような「弱い」理論では, 前提となる知見の根拠が乏しく, また知見ないしは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。そうした状況の下では, 仮に研究者がデータを見た後にその分析方法を検討すると, 与えられたデータに合致する予測と, それに適した分析を, 知らず知らずのうちに選んでしまう可能性が高まると考えられる。

3) p 値の正しい解釈に関して様々な誤解が蔓延していることも事実であるが, ここでは, 後に述べる偽発見率の問題を除いて, 議論しない (Nuzzo, 2014; Schervish, 1996; Wasserstein & Lazar, 2016)。

1.2 審美的判断基準, 出版バイアス, HARKing

次に, 研究を取り巻く制度, 特に結果を論文として報告する過程に関する問題点を見てみよう。学術ジャーナルに掲載される論文のほとんどは, 仮説を支持する方向で統計的に有意な結果を報告している (Fanelli, 2010, 2012; Sterling, 1959)。その一方で, 心理学研究の多くは低い検定力しか持ち合わせていないと推定されている (Button et al., 2013; Sedlmeier & Gigerenzer, 1989)。検定力が低いと, 本当は存在する効果であっても, 稀にしか検出できないはずであるから, 本来であれば, 統計的に有意に至らない結果が多数報告されてしかるべきである。しかし現実には, ジャーナルでの報告のほとんどは仮説通りの有意差を報告しているわけだから, 論文が刊行されるに際して, 否定的結果 (null results) が報告されにくいという「出版バイアス (publication bias)」が存在すると考えられる。出版バイアスは, 研究者側, ジャーナル側いずれの要因によっても生じうる (Mahoney, 1977; Sterling, Rosenbaum, & Weinkam, 1995)。報告されなかった研究結果は, 研究室の引き出しにしまわれてしまい「お蔵入り問題 (file drawer problem)」を引き起こす (Rosenthal, 1979)。最も極端なケースを想像すれば, あるテーマについて行われたすべての研究のうち, 第一種の過誤によって肯定的結果が得られた5%のみがジャーナルに報告され, 残りの95%は引き出しに眠っている可能性すら考えられる。

こうした出版バイアスが発生する原因として, 心理学者の多くが, 審美的な視点から研究を見てしまうという点が指摘されている (Giner-Sorolla, 2012)。審美的判断基準には三種類あるとされる。

第一は, 結果の一貫性である。一本の論文の中で複数の実験が報告され, それらが一貫して有意な結果を示していたとしよう。そうした報告を無条件に信頼できるものと感じたとしたら, 一貫性の審美基準に囚われている危険がある。なぜなら, 個々の実験の検定力が十分でない場合, 一貫した結果はむしろ, 有意になった実験だけを報告している可能性を示唆するからである (Schimmack, 2012)。逆に言えば, 結果の一貫性を求めるためにQRPsが行われ, 基準をクリアできなかったデータはfile drawerに隠されてしまうことになる。

第二, 第三の審美的判断基準は, 物語性と新奇性である。科学論文ではまず仮説を提出し, それをデータによって検証するというストーリーを伴った仮説検証型の物語形式が好まれ, 多くの心理学ジャーナルでも, その形式を踏襲することが明確に要求されている (Kerr, 1998)。そしてこの仮説検証物語の内には, 新奇性を伴う仮説の提案と, データによるその裏付けという形式も含まれている (Nosek, Spies, & Motyl, 2012)。そのため, 新奇性のない直接的追試などは, これまでの心理学の歴史の中では重要視されてこなかった (Greenwald, 1975; Koole & Lakens, 2012; Makel, Plucker, & Hegarty, 2012)⁴⁾。

しかしながら, 心理学で仮説検証型物語と新奇性とを完全に両立させるのは非常に困難である。その背景にあるのは, 前述した心理学における理論の弱さだろうと思われる。第一に, Eysenck (1985) の言う「弱い理論」では, 知見ないしは仮説間の相互依存性が低くなってしまふ。言い換えれば, 弱い理論の下では, 仮にある知見/仮説Aの頑健性が証明されたとしても, それが同じ理論から導かれた新奇な仮説Bの頑健性を高めることにつながりにくい。すなわち, 新奇性があり, かつ真である仮説を大量に生み出すことが困難である。さらに, 仮説検証に伴う諸々の前提がしっかりと裏付けられていないために, 仮に仮説を支持しない結果が得られたとしても, それが真に仮説を反駁したのか明確には分からず, 厳密な仮説検証の形式に耐えられないことが多い。おそらくこうした環境で, 唯一仮説検証形式に則ることができるのは, 直接的追試だけだと思われるが (Frank, 2016), そこに新奇性は乏しい。それゆえ, むしろ Rozin (2009) が指摘するように, 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奇性を探索すべきだと考えられる⁵⁾。

しかしながら, 現状の審美的判断基準のもとでは, 新奇性を伴う仮説検証型物語形式から逃れる

4) ここで明確にしておく必要があるが, 新奇性を伴う発見が科学の本質であることは言うまでもない。ここで問題としているのは, 新奇性の過度な重視によって, 知見の頑健さが損なわれてしまう事態である。

5) ただし, こうした記述的研究においても, QRPs等を厳密に拒否するべきであるのは言うまでもない。

ことは難しい。そこで、この状況を打開する方法として、データを得た後に、それに適合する仮を構築すること、すなわち HARKing が登場する (Hypothesizing After the Results are Known ; Bones, 2012 ; Kerr, 1998 ; O'Boyle, Banks, & Gonzalez-Mule, 2014)。HARKing の弊害は様々に論じられているが、おそらく最大の問題点は、それが第一種の過誤を増大させてしまう点にあると思われる。仮に、得られたデータに対して、様々な共変量の投入などの分析を複数行ったとしよう。そして多重比較の修正を無視すれば、その分析のひとつで有意差が得られたとしよう。ここで HARKing を用いず、当初の理論的予測に忠実に従った報告を行うならば、多重比較問題の無視が明らかとなってしまうため、報告の信頼性は損われ、論文の出版につながらないかもしれない。ところがここで HARKing を使えば、当該の有意差が得られた検定が、あたかも当初から予測されていた唯一の分析であるかのように扱うことができるため、多重比較の問題を「隠蔽」することが可能になる。さらに、心理学の弱い理論のもとでは、事前に厳密な理論的予測を立てることが難しいという事実を考えると、こうして「隠蔽」された多重比較を、論文報告のみから見抜くことは非常に難しいと思われる。このように HARKing には、言わば「見かけの研究者の自由度」を低め、第一種の過誤の発生確率を高める危険性がある。

1.3. 偽発見率と仮説が真である事前確率

さらに再現可能性を低める要因として、帰無仮説検定の枠組み自体の問題も存在する (Colquhoun, 2014 ; Ioannidis, 2005 ; Sterne & Smith, 2001)。よくある誤解として指摘されることだが、 $p < .05$ は、得られた結果が偶然の産物である確率が 5% より小さいという意味「ではない」。そうした可能性を正しく知るためには、(1) 偽陽性の確率と共に、(2) 検定力と、(3) その研究領域ないしは研究テーマにおいて、提案された仮説が正しいものである事前確率、の三つが必要となる。仮に、ある研究分野で検証されるすべての仮説のうち、10% が真実を反映しているとしてみよう。さらに研究者コミュニティが、これらの仮説を 1,000 件テストするとしてみよう。ここで、本当に存在する効果を正しく検出できる力、すなわち

表 1 様々な p 値、検定力、仮説が正しい確率において、有意な結果が偽陽性である確率

統計 検定力 (%)	有意となった結果が偽陽性である 確率 (%)		
	$p=.05$	$p=.01$	$p=.001$
仮説のうち 80% が正しい場合			
20	5.9	1.2	0.1
50	2.4	0.5	0.05
80	1.5	0.3	0.03
仮説のうち 50% が正しい場合			
20	20	4.8	0.5
50	9.1	2	0.2
80	5.9	1.2	0.1
仮説のうち 10% が正しい場合			
20	69.2	31	4.3
50	47.4	15.3	1.8
80	36	10.1	1.1

註：Sterne & Smith (2001), Table 3 の一部を改変して引用。

検定力が 80% のテストを行うとし、また偽陽性の基準は $\alpha=5\%$ と設定する。1,000 件の仮説のうち 10% の 100 件が真実であるが、それを検出できる可能性は 80% しかないから、80 件が真実として検出される。また残りの 900 件のうち偽陽性によって正しいと判断される可能性が 5% あるので、45 件が真実として検出されてしまう。結果、 $45/(80+45)=0.36$ となり、すなわち真実と判断された研究のうち 36% が、実際には偽陽性である。このような確率を、偽発見率 (false discovery rate : FDR) と呼ぶ (Benjamini & Hochberg, 1995)。

表 1 に、仮説が真である事前確率と、検定力、 p 値がそれぞれ変化した場合の FDR についてまとめた (Sterne & Smith, 2001)。ここで明らかのように、場合によって FDR は 5% から大きく離れてしまう。仮説が真である事前確率がどの程度であったとしても、心理学における平均的な検定力が 35% 前後であるという報告に基づけば (Button et al., 2013 ; Sedlmeier & Gigerenzer, 1989)、広く信じられている 5% 水準による偽陽性のコントロールと、実際の FDR との間に大きな開きがある可能性は否定できない。また、仮説が真である事前確率を正しく見極めることは難しく (Donnellan, 2014 ; Vazire, 2014)、また領域ごとの違いがあるとも考えられるが、先に述べたような心理学に



図 1

における理論の弱さに鑑みると (Eysenck, 1985 ; Gigerenzer, 2010 ; Rozin, 2009), 決して楽観視できないと考えるべきだろう。

以上述べてきた、心理学研究に関する根本問題の3側面について、ここで整理をしてみよう。最も基本的な問題は、おそらく心理学における理論の弱さと、研究への審美的基準であろう。理論の弱さは、真実である仮説を提案する事前確率を下げ、FDRを上昇させる。さらに審美的判断基準は出版バイアスを生み出し、それに対抗するためにQRPsが援用され、第一種の過誤の確率が増す。またQRPsの中には低い検定力の研究を行うことで、第一種の過誤による結果を有意と報告するという手段も含まれるため、さらにFDRを引き上げることにつながる。これらに加え、HARKingの入り込む余地を作り出し、FDRやQRPsによって生じる第一種の過誤に基づく結果を正当化する。そしてそのことは、さらに根拠のないランダムな心理学の理論を増大させ、真実である仮説に関する事前確率をより引下げる、という負のスパイラルを形成する (図1)。

1.4 Ioannidisの予測とNull Fieldの悪夢

Ioannidis (2005) は、こうした構造が加算的に偽発見率を引き上げていくことを考慮すれば、心理学を含む科学研究のほとんどが偽陽性である可能性すらある、という試算を行っている。さらに彼はまた、科学史を振り返れば、ある研究領域全体が、実際には存在しない現象を真実だと信じ、それを証明しようと努力を続けた事例は数多くあると指摘し、現在でもそうした「無の領域」(“null field”) が存在しないとは言いきれないとすら述べる。この5年ほどの間に報告された代表的

な追試研究の例を見てみると、心理学における再現可能性は、null fieldのように絶望的ではないものの、決して楽観的と言える状況でもない (Open Science Collaboration, 2015)。さらに、例えば社会的プライミングの研究領域は、多くの知見が再現可能性を持たないことが示されており、一部がnull fieldとなっている可能性も否定できない (Doyen et al., 2012 ; Gomes & McCullough, 2015 ; Johnson et al., 2014 ; Rohrer, Pashler, & Harris, 2015 ; Shanks et al., 2013, 2015)。事実、Inzlicht (2016) は、自分が専門としてきた、自我枯渇 (ego depletion) やステレオタイプ脅威 (stereotype threat) といった研究領域が、ひょっとしたらnull fieldだったのかもしれないという不安を表明している。追試プロジェクトが進行するにつれ、今後もこうした懸念が広がっていく可能性は否定できないだろう。

2. 解決策

それではこの危機的状況に対して、どのような解決方法が考えられるのだろうか。問題の構造を見直してみると (図1)、理論の弱さと審美的判断基準の存在が最も根源的な要因であり、他の要素はその二者から派生的に生じるものと思われる。このうちまず理論の弱さだが、これはまさに心理学の蓄積の進展が確保されなければ改善できないものと思われるため、これを直接的に是正するのは不可能である。次に審美的判断基準については、研究者各人に対してそうした諸基準を自主的に変更するように求めることは可能であるが、それだけでは有効性に疑問が残る。なぜならば、仮にある研究者が心理学全体の将来を憂い、審美的判断基準を変更し、QRPsやHARKingを止めようとしても、現状の体制の中でこれらの行為を他者より率先して止めることは、端的に当人のみ論文出版が困難になり、不利な状況に陥ることを意味している (Nosek et al., 2012)。すなわち、問題の誘因構造が典型的な社会的ジレンマに類似しているため、現状で自発的改善を望むのは現実的ではない (Engel, 2015 ; Everett & Earp, 2015 ; Koole & Lakens, 2012)。社会的ジレンマ解決のためには、構造的・制度的な変革が必要である (Engel, 2015)。

2.1 新しい論文ガイドラインの制定

既にいくつかの学会やジャーナルが、こうした制度的変革に着手し、論文投稿におけるガイドラインを修正している（Eich, 2014；Funder et al., 2014；The Psychonomic Society, 2012；表2）。こうしたガイドラインに従い、各問題を改善していくことが、結果の頑健性を保つ上で重要であることは言を俟たない。しかしこれらのガイドラインで触れられているポイント—検定力推定の励行、効果量や信頼区間の記載、QRPsの禁止など—は、先に行った整理（図1）によれば、問題の派生的な部分である。より本質的な問題である現行の審美的判断基準に変更を加えないまま、こうした厳しい諸条件を適用していけば、研究者は論文の出版を続けるため、より問題のある行動、例えば虚

偽報告などに手を染める可能性すら考えられる。

2.2 事前審査付き事前登録制度

以上のような議論から、現在再現性問題に対して最も有効な手段として提案されているのが、事前審査付き事前登録制度（pre-registration system）である（Chambers et al., 2013；Chambers, 2014；Neuroskeptic, 2008, 2011, 2012a, 2013a；Nosek & Lakens, 2014；Wagenmakers et al., 2012；Wolfe, 2013）。

事前審査付き事前登録制度では、まず研究の目的（論文の序論部分）、サンプルサイズ、研究デザイン、統計分析手法など（方法部分）が、データ取得以前にジャーナルに投稿され、掲載の可否についての査読が行われる。査読をパスした研

表2 頑健性確保のため各ジャーナル、団体が制定した新ガイドラインの例

	Society for Personality and Social Psychology (Funder et al., 2014)	Psychonomic Society (The Psychonomic Society, 2012)	Psychological Science (Eich, 2014)
統計報告	- 事前の検定力推定とサンプルサイズ決定方法の報告 - 効果量と95%信頼区間の報告	- 事前の検定力推定とサンプルサイズ決定方法の報告 - 効果量と95%信頼区間の報告と重視	- サンプルサイズ決定方法の報告 - 効果量と95%信頼区間の報告とメタアナリシスの奨励
QRPsの禁止	- 修正なしに同一データに対して複数の検定を行うことの禁止 - データ分析の結果を見てデータ収集を止めることの禁止 - データ分析の結果を見て観察、指標、項目、条件、参加者、行った実験などを除外して報告することの禁止	- 修正なしに同一データに対して複数の検定を行うことの禁止 - データ分析の結果を見てデータ収集を止めることの禁止 - データ分析の結果を見て観察、指標、項目、条件、参加者、行った実験などを除外して報告することの禁止 - 測定したすべての従属変数の報告 - 共変量分析は事前に計画し、探索的であると示す	- データ収集停止ルールの報告 - 測定したすべての独立変数、指標の報告 - 当該論文の研究目的のために分析したすべての従属変数の報告 - 観察を分析から除外した場合、理由とともにそれを示す
その他	- 付録での教示等の正確な報告 - データ公開 - 追試の奨励 - 柔軟で多様な研究評価軸の維持	- 帰無仮説が棄却されなかった時は第二種の過誤を疑う - ネガティブな結果報告の奨励 - ベイズ統計など帰無仮説検定以外の統計手法も受け入れる	- 論文の方法と結果の字数制限を撤廃 - データ公開の奨励 - マテリアル公開の奨励 - 事前登録の奨励

註：項目の翻訳と整理は著者が行った。

究は、内容をタイムスタンプと共に事前に登録し、その時点以降の改変が不可能なよう、内容が凍結される。その後実際にデータが収集されるが、得られた結果がどのようなものであろうと、報告はそのジャーナルに掲載されることになる (Neuroskeptic, 2012a)。

この制度のポイントは、論文掲載基準が事前審査のみとなっている点にある。すなわち研究報告が受理されるかどうかに関する審美的判断基準を、論文の序論・方法部分にのみ留め、結果と考察部分にまで波及させないことによって、そもそも研究者がデータや推論を事後的に補正しようとする積極的な動機を、根本的に消し去ることが可能になる。ここで注意が必要なのは、事前審査付き事前審査制度の導入は、必ずしもすべての審美的判断基準を排除するものではない、という点である。科学者が人間であり、人間の認知能力に限界がある以上、審美的基準を完全に排除した情報処理は容易ではないだろう。だが事前審査付き事前審査の導入は、それらの基準が、少なくとも研究の結果と考察を歪曲させることを防ぐ。それゆえ、現状のような出版バイアスが消滅するのはもちろんのこと、そもそも QRPs や HARKing を行う必然性も無くなる。また、この制度のもとでの論文出版が可能になれば、研究者は事前審査付き事前登録制度を活用した研究遂行に強い誘因を感じるようになるため、研究者コミュニティが自律的に、システムティックに心理学の信頼性を回復していくことにつながる。

なお、事前審査付き事前登録制度に関しては、その導入によって研究の探索的性格が奪われるのではないか、という懸念が示されているが (Mathôt, 2013; Scott, 2013)、これは杞憂である (Neuroskeptic, 2013a, 2013b, 2014b)。いかなる探索的な分析や考察も、そう明記した上であれば、論文にに記載することは妨げられないからである。それらが HARKing によって修飾されないよう、研究者の自由度に制限を設けることだけが、この制度の狙いである。

2.3 事前審査付き事前登録を伴う直接的追試

さらに、審美的判断基準にのみ捉われることなく、頑健な知見を蓄積していくためには、追試、特に先行研究の方法を忠実に再現する直接的追

試 (direct replication)⁶⁾ が不可欠である (Fisher, 1926; Popper, 1959)。データの取得には様々な物理的・経済的制限があり、ひとつの研究のみで十分な統計的検定力を達成すること難しい。その場合、直接的追試とメタ分析などによって知見の頑健性を確保していくことが必要となる。また、新しい研究を効率的に遂行させるためにも、直接的追試研究は重要である。先行研究の報告が虚偽であれば、それに基づく新しいアイデアや予測も間違ったものとなる可能性が高い。まずは直接的追試によって、先行研究の知見が十分信頼できるかを調べる必要がある。さらに、直接的追試は教育的にも重要である。教科書等に載っている知見の頑健性が、追試研究によって確保されることは、講師、学生の双方にとり望ましいことである。これらの重要性にもかかわらず、これまで直接的追試を出版することは、非常に困難であった (French, 2012; Greenwald, 1975; Koole & Lakens, 2012; Makel et al., 2012)。そのため、今後多くのジャーナルで追試を受理する制度を整備し、研究者が積極的に直接的追試に取り組めるための誘因を作る必要がある。

注意すべきことは、直接的追試それ自体は特に知見の再現可能性を担保するものではなく、頑健な直接的追試を行うためには、やはり事前審査付き事前登録制度を伴わなければならない、という点である (Neuroskeptic, 2012b)。たとえば、話題性の高いトピックを対象として直接的追試を行う場合、得られた結果が先行研究での報告に合致するものであれば、仮説検証型の物語に沿ったものになり、また先行研究の知見がインパクトのあるものであれば、それを否定する結果を示すことで、報告の新奇性が強調される。そのためこうした審美的判断基準が、直接的追試報告の方向性にバイアスを加える可能性は否定できない。加えて、低い検定力や研究者の自由度が許されていれば、QRPs や HARKing が入り込む余地も十分に出てくる。直接的追試の信頼性を保つためにも、事前審査・登録は必須である。

6) 同一の心理学概念や理論をテストする概念的追試 (conceptual replication) には新奇性を与えることが十分可能なので、ここでの議論の対象からは外れる。概念的再現研究はむしろ理論構築に寄与するものである (Earp & Trafimow, 2015)。

2.4 事前審査付き事前登録制度と直接的追試普及の現状

事前審査付き事前登録制度やそれを伴う追試の現状について見てみると、まず、タイムスタンプと内容の凍結手続きを伴う事前登録作業については、Center for Open Science (COS) が開発・運用する Open Science Framework (OSF; <https://osf.io/>) が広く用いられるようになってきた。さらに、事前審査付きの事前登録制度を採用するジャーナルの数も増加の一途をたどっている (<https://osf.io/8mpji/wiki/home/>; 常時受け付けているものが、2016年4月現在で21誌)。またよりカジュアルな追試報告の媒体としては、Psychfiledrawer も広く用いられている (<http://www.psychfiledrawer.org/>)。また、2016年4月現在ではまだ開発途中であるが、Curate Science (<http://curatescience.org/>) というページには、追試実験の結果がメタ分析と共にまとめられつつある。

さらに、学会やジャーナルが、各研究者に対して研究活動の透明性や再現可能性を求める際の基準として、Transparency and Openness Promotion (TOP) Guideline が提唱されている (Nosek et al., 2015)。このガイドラインは、データ、分析、研究に使用したマテリアル等の公開と、事前登録の有無、直接的追試報告の可否について、緩やかなものから厳しいものまで4つの段階を設定して、各学会やジャーナルが、自身の判断でそのいずれかを採用することができるように設計されている。また心理学のみを対象としたものではなく、科学界全体を対象としたものであり、2016年4月現在で、538のジャーナル、58の学会が採用を表明している。こうした動きは今後も広がりを見せていくものと予想される。

3. 結論と展望

本稿では、心理学における再現可能性危機について、その問題の構造と解決策の概要を述べてきた。そこで明らかになったように、現在の心理学研究の標準的プロセスには、研究報告の信頼性を損なう問題が構造的に存在している。既に事前登録制度や追試など、確実な効果が期待できる治療方法が確立されつつある以上、それらを取り入れ、積極的に心理学研究の根本構造を改革してい

くことが望まれる。

無論このことは、すべての心理学者が、これまで行ってきた研究手法をすべて改めなければならないという意味ではない。Open Science Collaboration (2015) で示されたように、心理学内部でも、認知心理学領域は、社会心理学領域と比較して、現行のままでもやや高い再現可能性を維持している。認知心理学でよく用いられるように、参加者内デザインを用い、正確に構成概念を測定するなどして、大きな効果量を手に入れる方法論を用いていけば、検定力は高くなり、それほど大きなサンプルサイズが必要ではなくなる上、追試も容易になる (Rouder, 2015)。高い検定力を伴う追試が容易になれば、あえて事前登録を用いずとも、同一の研究室内、あるいは同一の研究内で、追試を行いつつ探索的に研究を進めていくことも可能である (Frank, 2015a, 2015b)。だがそれでも、事前審査付き事前登録制度やそれを伴う直接的追試は、心理学研究の科学性を担保する最終防衛ラインであることは間違いないだろう。ある研究領域やトピックに関する信頼性に疑いが生じた場合には、一端ここに立ち戻り、確固たる基盤を確立してから、前に進んでいくことが、科学としての心理学を再構築していく近道となるだろう。

参考文献

- Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., ... Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, 27, 108–119. doi: 10.1002/per.1919
- Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, 7, 543–554. doi: 10.1177/1745691612459060
- Baumeister, R. F. (2016). Charting the future of social psychology on stormy seas: Winners, losers, and recommendations. *Journal of Experimental Social Psychology*. doi: 10.1016/j.jesp.2016.02.003
- Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, 100, 407–425. doi: 10.1063/1.3663724
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*.

- Series B (Methodological)*, 57, 289–300. doi: 10.2307/2346101
- Bloom, P. (2016). *Psychology's replication crisis has a silver lining*. *The Atlantic*. Retrieved from <http://www.theatlantic.com/science/archive/2016/02/psychology-studies-replicate/468537/>
- Bones, A. K. (2012). We knew the future all along: Scientific hypothesizing is much more accurate than other forms of precognition — A satire in one part. *Perspectives on Psychological Science*, 7, 307–309. doi: 10.1177/1745691612441216
- Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience *Nature Reviews Neuroscience*, 14, 365–376. doi: 10.1038/nrn3475
- Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on Psychological Science*, 9, 40–48. doi: 10.1177/1745691613513470
- Chambers, C. D. (2014). *Psychology's "registration revolution."* *The Guardian*. Retrieved from <http://www.theguardian.com/science/head-quarters/2014/may/20/psychology-registration-revolution>
- Chambers, C. D., Munafò, M., & more than eighty signatories. (2013). *Trust in science would be improved by study pre-registration*. *The Guardian [open letter]*. Retrieved from <https://www.theguardian.com/science/blog/2013/jun/05/trust-in-science-study-pre-registration>
- Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of *P* values. *Royal Society Open Science*, 1–15. doi: 10.1098/rsos.140216
- Dijksterhuis, A. (2014). Welcome back theory! *Perspectives on Psychological Science*, 9, 72–75. doi: 10.1177/1745691613513472
- Donnellan, M. B. (2014). *Things that make me skeptical ... The Trait-State Continuum [web log]*. Retrieved from <https://traitstate.wordpress.com/2014/03/26/things-that-make-me-skeptical/>
- Doyen, S., Klein, O., Pichon, C.-L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PloS One*, 7, e29081. doi: 10.1371/journal.pone.0029081
- Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, 6 (May), 1–11. doi: 10.3389/fpsyg.2015.00621
- Eich, E. (2014). Business not as usual. *Psychological Science*, 25, 3–6. doi: 10.1177/0956797613512465
- Engel, C. (2015). Scientific disintegrity as a public bad. *Perspectives on Psychological Science*, 10, 361–379. doi: 10.1177/1745691615577865
- Everett, J. A. C., & Earp, B. D. (2015). A tragedy of the (academic) commons: Interpreting the replication crisis in psychology as a social dilemma for early-career researchers. *Frontiers in Psychology*, 6 (May). doi: 10.3389/fpsyg.2015.01152
- Eysenck, H. J. (1985). The place of theory in a world of facts. In K. B. Madsen & L. Mos (Eds.), *Annals of Theoretical Psychology, Volume 3* (pp. 17–72). New York: Plenum Press. doi: 10.1007/978-1-4613-2487-4_2
- Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PLoS ONE*, 5. doi: 10.1371/journal.pone.0010068
- Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, 90, 891–904. doi: 10.1007/s11192-011-0494-7
- Feldman Barrett, L. (2015). Psychology is not in crisis. *The New York Times*. Retrieved from http://www.nytimes.com/2015/09/01/opinion/psychology-is-not-in-crisis.html?_r=0
- Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain*, 33, 503–513.
- Frank, M. C. (2015a). *A moderate's view of the reproducibility crisis*. *Babies Learning Language [web log]*. Retrieved from <http://babieslearninglanguage.blogspot.jp/2015/08/a-moderates-view-of-reproducibility.html>
- Frank, M. C. (2015b). *The slower, harder ways to increase reproducibility*. *Babies Learning Language [web log]*. Retrieved from <http://babieslearninglanguage.blogspot.jp/2015/08/the-slower-harder-ways-to-increase.html>
- Frank, M. C. (2016). *Should we always bring out our nulls?* *Babies Learning Language [web log]*. Retrieved from <http://babieslearninglanguage.blogspot.jp/2016/03/should-we-always-bring-out-our-nulls.html>
- French, C. C. (2012). Precognition studies and the curse of the failed replications. *The Guardian*. Retrieved from <https://www.theguardian.com/science/2012/mar/15/precognition-studies-curse-failed-replications>
- 藤島喜嗣・樋口匡貴 (2016) 社会心理学における“*p*-hacking”の実践例 心理学評論, 59, 84–97.
- Funder, D. C., Levine, J. M., Mackie, D. M., Morf, C. C., Sansone, C., Vazire, S., & West, S. G. (2014). Improving the dependability of research in personality and social psychology: Recommendations for research and educational practice. *Personality and Social Psychology Review*, 18, 3–12. doi: 10.1177/1088868313507536
- Galak, J., LeBoeuf, R. A., Nelson, L. D., & Simmons, J. P. (2012). Correcting the past: Failures to replicate psi. *Journal of Personality and Social Psychology*, 103, 933–948. doi: 10.1037/a0029709
- Gelman, A. (2015). *To understand the replication crisis, imagine a world in which everything was published*. *Statistical Modelling, Causal Inference, and Social*

- Science [web log]*. Retrieved from <http://andrewgelman.com/2015/09/02/to-understand-the-replication-crisis-imagine-a-world-in-which-everything-was-published/>
- Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, 102, 460. doi: 10.1511/2014.111.460
- Gigerenzer, G. (2010). Personal reflections on theory and psychology. *Theory & Psychology*, 20, 733–743. doi: 10.1177/0959354310378184
- Giner-Sorolla, R. (2012). Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. *Perspectives on Psychological Science*, 7, 562–571. doi: 10.1177/1745691612457576
- Gomes, C. M., & McCullough, M. E. (2015). The effects of implicit religious primes on dictator game allocations: A preregistered replication experiment. *Journal of Experimental Psychology: General*. doi: 10.1037/xge0000027
- Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*. doi: 10.1037/h0076157
- Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Birt, A., Brand, R., ... Cannon, T. (2014). A multi-lab pre-registered replication of the ego-depletion effect. *Perspectives on Psychological Science*, 25, 1227–1234. doi: 10.1177/0956797614526415.Data
- Inzlicht, M. (2016). *Reckoning with the past*. *Getting Better [web log]*. Retrieved from <http://michaelinzlicht.com/getting-better/2016/2/29/reckoning-with-the-past>
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2, e124. doi: 10.1371/journal.pmed.0020124
- Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology (Cambridge, Mass.)*, 19, 640–648. doi: 10.1097/EDE.0b013e31818131e7
- John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23, 524–532. doi: 10.1177/0956797611430953
- Johnson, D. J., Cheung, F., & Donnellan, M. B. (2014). Does cleanliness influence moral judgments?: A direct replication of Schnall, Benton, and Harvey (2008). *Social Psychology*, 45, 209–215. doi: 10.1027/1864-9335/a000186
- Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2, 196–217. doi: 10.1207/s15327957pspr0203_4
- Koole, S. L., & Lakens, D. (2012). Rewarding replications: A sure and simple way to improve psychological science. *Perspectives on Psychological Science*, 7, 608–614. doi: 10.1177/1745691612462586
- Lakens, D. (2016). *Where are all the competent researchers? The 20%Statistician [web log]*. Retrieved from <http://daniellakens.blogspot.jp/2016/02/where-are-all-competent-researchers.html>
- Lieberman, M. D. (2014). *Latitudes of acceptance*. *Edge*. Retrieved from <https://www.edge.org/conversation/latitudes-of-acceptance>
- Locke, E. A. (2015). Theory building, replication, and behavioral priming: Where do we need to go from here? *Perspectives on Psychological Science*, 10, 408–414. doi: 10.1177/1745691614567231
- Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research*, 1, 161–175. doi: 10.1007/BF01173636
- Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, 7, 537–542. doi: 10.1177/1745691612460688
- Mathôt, S. (2013). *The pros and cons of pre-registration in fundamental research*. *Cogsci.nl [web log]*. Retrieved from <http://www.cogsci.nl/blog/miscellaneous/215-the-pros-and-cons-of-pre-registration-in-fundamental-research>
- Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, 9, 147–163. doi: 10.1037/1082-989X.9.2.147
- Maxwell, S. E., Lau, M. Y., Howard, G. S., & Scott, E. (2015). Is psychology suffering from a replication crisis? What does “failure to replicate” really mean? *American Psychologist*, 70, 487–498. doi: 10.1037/a0039400
- Mitchell, J. P. (2014). *On the emptiness of failed replications*. Retrieved from http://web.archive.org/web/20150429064229/http://wjh.harvard.edu/~jmitchell/writing/failed_science.htm
- Neuroskeptic. (2008). *Registration: Not just for clinical trials*. *Discover [web log]*. Retrieved from <http://blogs.discovermagazine.com/neuroskeptic/2008/11/03/registration-not-just-for-clinical-trials/>
- Neuroskeptic. (2011). *How to fix science*. *Discover [web log]*. Retrieved from <http://blogs.discovermagazine.com/neuroskeptic/2011/05/24/how-to-fix-science/>
- Neuroskeptic. (2012a). *Fixing science – Systems and politics*. *Discover [web log]*. Retrieved from <http://blogs.discovermagazine.com/neuroskeptic/2012/04/14/fixing-science-systems-and-politics/>
- Neuroskeptic. (2012b). *Replication alone is not enough*. *Discover [web log]*. Retrieved from <http://blogs.discovermagazine.com/neuroskeptic/2012/08/25/replication-alone-is-not-enough/>
- Neuroskeptic. (2013a). *For preregistration in fundamental research*. *Discover [web log]*. Retrieved from <http://blogs.discovermagazine.com/neuroskeptic/2013/04/25/>

- for-preregistration-in-fundamental-research/
- Neuroskeptic. (2013b). *Preregistration ... Problem? Discover [web log]*. doi: <http://blogs.discovermagazine.com/neuroskeptic/2013/04/29/preregistration-problem/>
- Neuroskeptic. (2014a). *On "On the emptiness of failed replications."* *Discover [web log]*. Retrieved from <http://blogs.discovermagazine.com/neuroskeptic/2014/07/07/emptiness-failed-replications/>
- Neuroskeptic. (2014b). *Preregistration for data science? Discover [web log]*. Retrieved from <http://blogs.discovermagazine.com/neuroskeptic/2014/02/22/preregistration-reanalysis/>
- Neuroskeptic. (2014c). *The replication crisis: response to lieberman. Discover [web log]*. Retrieved from <http://blogs.discovermagazine.com/neuroskeptic/2014/08/31/replication-crisis-response-lieberman/>
- Neuroskeptic. (2015). *Psychology should aim for 100% reproducibility. Discover [web log]*. Retrieved from <http://blogs.discovermagazine.com/neuroskeptic/2015/09/07/100-percent-reproducibility/>
- Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... Yarkoni, T. (2015). Promoting an open research culture. *Science*, 348, 1422–1425. doi: 10.1126/science.aab2374
- Nosek, B. A., & Lakens, D. (2014). Registered reports. *Social Psychology*, 45, 137–141. doi: 10.1027/1864-9335/a000192
- Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, 7, 615–631. doi: 10.1177/1745691612459058
- Nuzzo, R. (2014). Scientific method: Statistical errors. *Nature*, 506, 1–19. doi: 10.1038/506150a
- O'Boyle, E. H., Banks, G. C., & Gonzalez-Mule, E. (2014). The chrysalis effect: How ugly initial results metamorphosize into beautiful articles. *Journal of Management*, published online before print. doi: 10.1177/0149206314527133
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349, aac4716–aac4716. doi: 10.1126/science.aac4716
- Popper, K. (1959). *The Logic of Scientific Discovery*. London: Routledge.
- Ritchie, S. J., Wiseman, R., & French, C. C. (2012). Failing the future: Three unsuccessful attempts to replicate Bem's "retroactive facilitation of recall" effect. *PLoS One*, 7, e33423. doi: 10.1371/journal.pone.0033423
- Rohrer, D., Pashler, H., & Harris, C. R. (2015). Do subtle reminders of money change people's political views? *Journal of Experimental Psychology: General*, 144, 1–13. doi: 10.1037/xge0000058
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86, 638–641. doi: 10.1037/0033-2909.86.3.638
- Rouder, J. N. (2015). *How many participants do I need? Insights from a Dominance Principle*. Retrieved from <http://jeffrouder.blogspot.jp/2015/04/how-many-participants-do-i-need.html>
- Rozin, P. (2009). What kind of empirical research should we publish, fund, and reward? *Perspectives on Psychological Science*, 4, 435–439. doi: 10.1111/j.1745-6924.2009.01151.x
- Schervish, M. J. (1996). P values: What they are and what they are not. *The American Statistician*. doi: 10.2307/2684655
- Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, 17, 551–566. doi: 10.1037/a0029487
- Schwarzkopf, D. S. (2016). *Why Gilbert et al. are missing the point. NeuroNeurotic [web log]*. Retrieved from <http://neuroneurotic.net/2016/03/08/why-gilbert-et-al-are-missing-the-point/>
- Scott, S. K. (2013). *Pre-registration would put science in chains. Times Higher Education [web log]*. Retrieved from <https://www.timeshighereducation.co.uk/comment/opinion/pre-registration-would-put-science-in-chains/2005954.article>
- Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, 105, 309–316. doi: 10.1037/0033-2909.105.2.309
- Shanks, D. R., Newell, B. R., Lee, E. H., Balakrishnan, D., Ekelund, L., Cenac, Z., ... Moore, C. (2013). Priming intelligent behavior: An elusive phenomenon. *PLoS ONE*, 8. doi: 10.1371/journal.pone.0056515
- Shanks, D. R., Vaddilo, M. A., Riedel, B., Clymo, A., Govind, S., Hickin, N., ... Puhlmann, L. M. C. (2015). Romance, risk, and replication: Can consumer choices and risk-taking be primed by mating motives? *Journal of Experimental Psychology: General*, 144, 142–158. doi: 10.1037/xge0000116
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22, 1359–1366. doi: 10.1177/0956797611417632
- Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, 9, 76–80. doi: 10.1177/1745691613514755
- Simonsohn, U. (2014). Posterior-hacking: Selective reporting invalidates bayesian results also. *SSRN Electronic Journal*, 1800, 1–10. doi: 10.2139/ssrn.2374040
- Simonsohn, U. (2015a). *Accepting the null: Where to draw*

- the line? *Data Colada [web log]*. Retrieved from <http://datacolada.org/42>
- Simonsohn, U. (2015b). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*. doi: 10.1177/0956797614567341
- Simonsohn, U. (2016). *Evaluating replications: 40%full ≠ 60%empty*. *Data Colada [web log]*. Retrieved from <http://datacolada.org/47>
- Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance — Or vice versa. *Journal of the American Statistical Association*, 54, 30–34. doi: 10.2307/2282137
- Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician*, 49, 108–112. doi: 10.2307/2684823
- Sterne, J. A. C., & Smith, G. D. (2001). Sifting the evidence—what's wrong with significance tests? *BMJ (Clinical Research Ed.)*, 322, 226–231. doi: 10.1136/bmj.322.7280.226
- Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, 9, 59–71. doi: 10.1177/1745691613514450
- The Psychonomic Society. (2012). *New Statistical Guidelines for Journals of the Psychonomic Society*. [web page]. Retrieved from <http://www.springer.com/psychology?SGWID=0-10126-6-1390050-0>
- Valentine, J. C., Biglan, A., Boruch, R. F., Castro, F. G., Collins, L. M., Flay, B. R., ... Schinke, S. P. (2011). Replication in prevention science. *Prevention Science*, 12, 103–117. doi: 10.1007/s11121-011-0217-6
- Vazire, S. (2014). *Unbelievable. Sometimes I'm wrong [web log]*. Retrieved from <https://sometimesimwrong.typepad.com/wrong/2014/03/unbelievable.html>
- Verhagen, J., & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology. General*, 143, 1457–1475. doi: 10.1037/a0036731
- Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J. & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, 7, 632–638. doi: 10.1177/1745691612463078
- Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, 1305 (March), 129–133. doi: 10.1080/00031305.2016.1154108
- Wolfe, J. M. (2013). Registered reports and replications in attention, perception, & psychophysics. *Attention, Perception, & Psychophysics*, 75, 781–783. doi: 10.3758/s13414-013-0502-5
- 2016. 3. 23 受稿, 2016. 4. 17 受理 —