Japanese Psychological Review
2016, Vol. 59, No. 1, 84–97

社会心理学における “*p*-hacking” の実践例[1)]

藤 島 喜 嗣[1]，樋 口 匡 貴[2]

[1] 昭和女子大学
[2] 上智大学

Case studies of “*p*-hacking” in social psychology

Yoshitsugu FUJISHIMA[1] and Masataka HIGUCHI[2]

[1]Showa Women’s University
[2]Sophia University

There is currently an ongoing debate about reproducibility in social psychology. One reason for low reproducibility is the excessive use of questionable research practices, called “*p*-hacking”. We present two direct replication studies of social priming and embodied cognition that failed to replicate the original findings under the circumstances of high statistical power. However, a variety of *p*-hacking attempts made it possible to obtain some false-positive findings based on the data from these two studies. We note that selectively reporting the results and deriving the hypothesis after the results are obtained may disguise the presence of *p*-hacking, and argue that pre-registration of studies and fair publishing of negative results could inhibit *p*-hacking.

Key words: reproducibility, *p*-hacking, direct replication, pre-registration, publication bias
キーワード：再現性，*p*-hacking，直接的追試，事前登録，出版バイアス

1. はじめに

1.1 社会心理学における再現性問題

社会心理学，特に社会的認知領域における研究知見の再現性（reproducibility）[2)]が問題となっている。たとえば，Bargh, Chen, and Burrows（1996）の高齢者プライミングを用いた知覚－行動リンク研究は社会心理学におけるプライミング効果研究で重要な地位を占めるが，追試に失敗しており再現性に疑義がもたれている（Doyen et al., 2012）。目標プライミング（Bargh et al., 2001）についても同様に追試失敗が報告されており（Harris et al., 2013），社会心理学におけるプライミング効果の存在自体への疑念となり，議論になっている（Bargh, 2012；Harris et al., 2013）。

近年，社会心理学領域では身体化認知（embodied cognition）に関する研究が多くなされている。身体化認知研究は，知覚システム，身体運動，情動が対人認知や社会的判断に及ぼす影響を検討しており，社会的認知過程に関して有用な知見をもたらしている（Barsalou et al., 2003）。その一方で，それらの研究知見の再現性が低いという批判もなされている（Lakens, 2014）。たとえば，Williams and Bargh（2008a）は冷たい袋を持つ場合と比較して温かい袋を持つ場合には向社会的選択がなされると報告しているが，Lynott et al.（2014）による追試ではその知見が再現されていない。また，Slepian et al.（2012）は重大な秘密を想起した者は軽微な秘密を想起した者と比較して坂を急勾配だと判断することを見いだしたが，こちらも同様

1) 本研究はJSPS科研費15K13122の助成を受けて行われた。また，本論文は，平石界氏，三浦麻子氏，池田功毅氏との議論に基づくものである。平石界氏と三浦麻子氏からは本論文で紹介した事例のデータ収集においても協力を得た。記して感謝する。

2) 再現性を示す基準については議論の余地がある（Open Science Collaboration, 2015）。本論文では，十分な統計検定力を確保した上で本研究と同一の手続きで追試をしたときに（検定力と例数設計に関しては大久保（2016）を参照），元の研究と同方向の有意な効果が認められたならば，再現性があるとみなす。これは近年の再現性に関する議論で素朴に用いられている基準でもある。

に追試失敗が報告されている（LeBel & Wilbur, 2014）。他にも本論文で後述する空間的距離プライミング（Williams & Bargh, 2008b），清浄プライミング（Schnall, Benton, & Harvey, 2008）についても再現性に関する疑義が示されている（Johnson, Cheung, & Donnellan, 2014；Pashler, Coburn, & Harris, 2012）。

このような再現性に関する疑義は，実験を多用する領域を中心に生じている。この理由として元の研究と同じ手続き，分析を用いる直接的追試が容易であることが考えられる。その意味で再現性問題は，再現可能な知見を元に理論を構築しようとする実験科学の方法論の中で起きている健全な議論かもしれない。他方で，再現可能な知見が日常的期待よりも少ない可能性が指摘されている。Open Science Collaboration（2015）は，2008 年に社会心理学の主要投稿先 3 誌に掲載された 100 研究の直接的追試を行った。その結果，元の研究の 97% が統計的に有意な効果を報告していたのに対し，直接的追試では 36% に留まった。このような社会心理学の理論と知見の信用性が疑われかねない状況が，社会心理学の主要な雑誌でも問題視されている（Association for Psychological Science, 2015；Varize, 2016）。中には，この問題をいち早く認識し対応した雑誌もある。たとえば，*Basic and Applied Social Psychology* 誌は，帰無仮説検定を破棄し，統計的仮説検定に関する記載を含めないこととした（Trafimow & Marks, 2015）。

1.2 *p*-hacking

研究知見の再現性を低める要因は複数ありうる。その中でも，帰無仮説検定において有意水準に達した分析結果を人為的に得る問題ある研究実践，通称 *p*-hacking が問題視されている（平石・池田，2015；池田・平石，2016；John, Loewenstein, & Prelec, 2012；Nosek, Spies, & Motyl, 2012；Simmons, Nelson, & Simonsohn, 2011）。*p*-hacking は，望まない結果をもたらすデータ・変数・実験条件の削除，事後的データ変換，共変量の使用，都合の悪い実験報告の抑制，選択的データ収集（都合の良い結果が得られた段階で実験を終える）など多岐にわたる。*p*-hacking の中には，事後的データ変換や共変量の使用など探索的分析で用いる実践が含まれている。このことは，*p*-hacking が日常の研究活動において意図せず行われる可能性を示唆する。

p-hacking が意図的，非意図的に行われることで研究知見の偽陽性（false positive）の可能性が高まり，再現性を低めることにつながる。Simmons et al.（2011）は，正規分布から無作為抽出して生成した複数の仮想データに対し，従属変数の選択，データの選択的追加，共変量の使用，実験条件の削除を行い，偽陽性が生じる可能性を検討した。その結果，“$p < .05$” の偽陽性が生じる可能性は 60.7% に到達した。

前節 1.1 における社会心理学研究が *p*-hacking を実践したかどうかは定かではない。その一方で，John et al.（2012）は 5964 名の心理学者に *p*-hacking を含む研究実践に関わる匿名調査を行っている。その回収率は 36% であったが，匿名性を保証しながら真実を話すことに誘因を設ける工夫がなされていた。その結果によれば，*p*-hacking を行ったことのある研究者の割合が高い傾向にあった。たとえば，うまくいった研究だけを報告したことのある研究者の割合はおよそ 5 割に達し，従属変数を選択的に報告した研究者の割合は 6 割を超えた。また，*p*-hacking は，認知，神経科学，社会領域で，さらには行動指標を用いた実験室実験を用いる場合に多く報告された（John et al., 2012）。これらの領域を中心として心理学界に *p*-hacking は普及していると考えられる。

次章以降，社会心理学におけるプライミング研究，特に身体化認知に関わる直接的追試研究を 2 事例紹介する。いずれの事例も，複数の研究者が協力して一定以上のサンプルを準備し，十分な検定力を備えるよう配慮した。結論からいえば，どちらの事例においても元の研究の知見は再現されなかった。それらのデータに対し付加的分析を行い，*p*-hacking を実践する。それぞれの事例において *p*-hacking を行うことによって一見仮説を支持するような分析結果が得られることを示す。

2. 事例 1：Williams and Bargh（2008b, Study 3）の直接的追試

2.1 目的

Williams and Bargh（2008b, Study 3）は，空間的距離プライミングが感情反応の強さに影響する

か検討したものである。空間的距離は，言語習得前の子どもも利用可能な概念とされ（Clark, 1973；Mandler, 1992），後に発達する心的距離概念の基礎をなし，評価的意味合いを持つとされる（Fauconnier & Turner, 2002；Lakoff & Johnson, 1980）。そのため，Williams and Bargh（2008b）は，空間距離の認識は感情反応と密接に関連すると主張した。具体的には，遠くにある物は自己との無関連が，近くにある物は自己との関連が示唆される。そのため，空間的距離が遠く知覚される事象には感情反応が弱まり，近くに感じられる事象には感情反応が強まるのである。Williams and Bargh（2008b）は，この議論をさらに推し進め，自分との距離でなくても，単なる空間上の距離について考えるだけでその後の感情反応が異なると主張した。彼らは，方眼紙上に2つの点を打たせた後に様々な判断を求める実験を行った。その結果，2点の距離が大きくなるほど，小説から感情を感じにくく（Study 1, 2），不健康食品の危険を感じず，カロリーを低く見積もり（Study 3），対人関係上の愛着を感じにくくなった（Study 4）。

Williams and Bargh（2008b）は，身体化認知の理論的枠組みである概念メタファー理論（Lakoff, 2012；Landau, Meier, & Keefer, 2010）を支持する重要な研究である。その一方，再現失敗報告もあり（Pashler et al., 2012），その理論構成に疑念を感じる部分もある。空間的距離の重要性を認めるにしても，自己からの距離でなくてよいとする理論的基盤が十分にない。また，Williams and Bargh（2008b）も認めるとおり，心的距離に関する理論である解釈レベル理論（Trope & Liberman, 2003, 2010）と整合しない。そこで，William and Bargh（2008b, Study 3）の直接的追試を複数の大学で行った。健康食品では空間的距離プライミング操作の影響はみられないが，不健康食品では近接プライミングに比べ，遠方プライミングを受けるとカロリー推定値が低くなるという交互作用効果の再現を試みた。

2.2　方法

実験参加者　Williams and Bargh（2008b, Study 3）は，健康食品，不健康食品のカロリー推定に空間的距離プライミングの効果を検討する際に参加者間1要因参加者内1要因の混合計画による分散分析を行い，η_p^2=.10の結果を得ている。この効果サイズで有意水準を5%としたとき，80%，90%，95%の検定力を得るためには，それぞれ42名，54名，66名の参加者が必要となる[3,4]。本研究では大学生396名が実験に参加した（男性124名，女性257名，不明15名，平均年齢20.10歳，*SD*=1.26）。各大学の内訳は，昭和女子大学106名，上智大学118名，慶應義塾大学172名であった。これらの実験参加者を近接（*n*=136），中庸（*n*=141），遠方（*n*=119）の3条件に無作為配置した。

実験素材[5]　Williams and Bargh（2008b）に記載されている手続きから質問紙を作成した。空間的距離プライミング操作に用いる方眼紙はWilliams and Bargh（2008b）のFig. 1を参考にA4用紙1枚大に作成した。元の論文では原点が示されていなかったが，操作上支障があるため本論文では原点を示した。教示は単純に「以下の方眼紙に次の2つの点を×で記入してください」とし，参加者の修学経験を考慮して，各点を “(x, y)=(2, 4)” の形式で示した。カロリー推定に用いる食品に関しては，食品の重さ（g）を記載し，写真を添付した。また，食品の提示順序は固定した。これら3点については元の論文に記載が無く，本研究独自の手続きである可能性が高い。健康食品，不健康食品の翻訳に際し，直訳を避け，参加者になじみやすく翻訳し，場合によっては類似品に変更した。不健康食品は，アイスクリーム，フライドポテト，ポテトチップス，チョコバー，チーズバーガーの5品目であった。その一方，健康食品はアロエヨーグルト，牛乳入りコーンフレーク，玄米，リンゴMサイズ，および芋料理となった。芋料理は各大学で用意したものが，じゃがバター，粉ふきいも，焼き芋と異なった。元の研究の “baked potato” を代替する食品の見解が異なったためである。しかしながら，およそ類似した食品であるので分析時には区別しなかった。また，各研究室で独自のターゲット食品を1～2品目を追加した

3)　事例1，事例2の検定力の算出にはG*Power 3.1（Faul et al., 2007, 2009）を利用した。
4)　元の研究では従属変数間の相関が報告されていない。検定力算出においてはこの相関を.50と仮定して算出した。
5)　実験素材（昭和女子大学版）がOpen Science Framework（https://osf.io/2ytz8/）から閲覧可能である。なお，研究実施に際しては研究実施機関の倫理審査を受け承認を得た。

が，これらは共通した測定ではないので本研究では用いなかった[6]。

手続き　Williams and Bargh（2008b）と同様に「標準化テスト新規作成のための調査」と偽りの研究目的を告げ，質問紙実験を実施した。回答に際しては倫理的配慮に関する事前説明を行い，同意した者だけに実験参加を求めた。その後，実験参加者各自のペースで質問紙への回答を求めた。最初に空間的距離プライミング操作である方眼課題に回答を求めた。近接条件では（2, 4）と（–3, –1）の座標に，中庸条件では（8, 3）と（–6, –5）の座標に，遠方条件では（12, 10）と（–11, –8）の座標に×印をつけさせた。その後，健康食品5品目，不健康食品5品目のカロリー量の推定を求めた。順番を無作為化したものを1パターン用意し，全ての実験参加者に同じ順番で提示した。回答は全体でおよそ5分間を要した。回答終了後，説明資料を配付し，デブリーフィングを行った。

2.3　結果と考察

Williams and Bargh（2008b, Study 3）ではカロリー推定の回答を健康食品5品目，不健康食品5品目で平均し指標化していた。論文中では10項目でのα係数（α=.75）は示されていたが，健康食品，不健康食品別のα係数は報告されていなかった。本研究ではいずれにおいてもα=.75となり，ある程度の信頼性が保証されていた。これら指標を従属変数とし，3（距離プライム：近接・中庸・遠方）×2（食品：健康・不健康）の参加者間1要因参加者内1要因の混合計画による分散分析を行った。各条件の平均値をTable 1に示す。その結果，食品の主効果が認められた（$F(1, 385)=747.23$, $p<.001$, $\eta_p^2=.66$）。健康食品（$M=192.68$）よりも不健康食品（$M=342.01$）でカロリーが高く見積もられていた。距離プライムの主効果（$F(2, 385)=0.04$, *ns*, $\eta_p^2=.00$），距離プライム×食品の交互作用効果（$F(2, 385)=0.11$, *ns*, $\eta_p^2=.00$）は認められなかった。ここまでの分析では空間的距離プライミング操作の影響はみられず，Williams and Bargh（2008b, Study 3）の結果は再現されなかった。これは，Pashler et al.（2012）と同様の結果であった。

6)　オリジナル項目は，昭和女子大学でカップ春雨，上智大学でカップヌードルと冷や奴，慶應義塾大学で蕎麦とラーメン二郎であった。

Table 1　事例1：条件別にみたカロリー推定値（全体データ，*N*=396）

距離プライム	近接（*n*=136）	中庸（*n*=141）	遠方（*n*=119）
健康	193.50（107.79）	189.83（70.69）	194.69（78.91）
不健康	339.71（174.60）	342.10（105.40）	344.23（131.19）

注　カッコ内は標準偏差

2.4　付加分析あるいは *p*-hacking の実践

本研究は複数の大学で実施しているため，実施大学の違いが結果に影響しているかもしれない。そこでもっとも大きな慶應義塾大学データ（*n*=172，男性45名，女性127名；平均年齢19.66歳，*SD*=1.38）にのみ着目し分析対象とした。最初に，食品毎に距離プライムの影響を検討するために参加者間1元配置の分散分析を行った（Table 2）。その結果，焼き芋においてのみ有意に近い効果が認められた（$F(2, 167)=2.43$, $p<.10$, $\eta_p^2=.03$）。近接条件（$M=297.21$）と比較して，中庸条件（$M=249.60$），遠方条件（$M=235.40$）の方がカロリーを低く見積もっていた。それ以外では有意な効果は認められず（Fs<1.12），効果サイズも小さかった（η_p^2s<.02）。

有意ではないものの平均値パターンに着目すると，健康食品5品目のうち，先述の焼き芋とリンゴはプライムされる距離が大きくなるほどカロリーを低く見積もっていた。これは不健康品目で認められるべきパターンである。その他3品目（ヨーグルト，コーンフレーク，玄米）はそのような影響は認められなかった。他方，不健康食品5品目に着目すると，チーズバーガーとフライドポテトでは明確なプライミング効果を読み取れなかったが，残り3品目（アイスクリーム，ポテトチップス，チョコバー）ではプライムされる距離が大きくなるほどカロリーを低く見積もっており，予測と一致する平均値パターンを示した。

総合して考えると，不健康食品のうちWilliams and Bargh（2008b）の予測通りの結果を示すのは，間食に相当する3品目であり，健康食品のうち予測通りの結果を示すのは主食に相当する3品目であった。元の研究では明確に見いだされなかった主食，間食という要因が影響しているかもし

Table 2　事例 1：条件別にみた各食品のカロリー推定値（慶應データ，*n*=172）

距離プライム	全体（*n*=172）	近接（*n*=58）	中庸（*n*=63）	遠方（*n*=50）	*F* 値	η_p^2
ヨーグルト	140.90（77.18）	136.50（98.41）	139.22（64.45）	148.12（63.76）	0.33	.00
コーンフレーク	283.18（173.44）	297.66（254.08）	267.86（98.93）	285.70（127.76）	0.45	.01
玄米	201.49（115.44）	185.28（106.87）	209.73（134.53）	209.90（97.86）	0.86	.01
りんご	119.61（78.32）	123.02（82.70）	123.38（87.55）	110.92（59.29）	0.43	.01
焼き芋	261.67（156.41）	297.21（205.67）	249.60（123.73）	235.40（117.16）	2.43†	.03
アイスクリーム	284.15（126.24）	289.79（129.86）	296.38（142.23）	262.20（96.71）	1.11	.01
フライドポテト	447.21（369.82）	424.31（258.70）	462.48（172.53）	454.84（596.51）	0.17	.00
ポテトチップス	408.23（254.77）	414.69（383.23）	432.22（164.09）	370.50（134.29）	0.85	.01
チョコバー	266.73（144.97）	270.66（150.45）	276.71（145.58）	249.80（139.04）	0.51	.01
チーズバーガー	422.40（223.94）	420.53（340.26）	431.43（131.08）	413.20（131.82）	0.09	.00

注　カッコ内は標準偏差。†：$p<.10$.

れない。そこで，該当する各 3 品目で平均して指標を作成し，あらためてこれらを従属変数とし，3（距離プライム：近接・中庸・遠方）× 2（食品：健康・不健康）の参加者間 1 要因参加者内 1 要因の混合計画による分散分析を行った。各条件の平均値を Table 3 に示す。その結果，食品の主効果が認められた（$F(1, 167)=157.76$, $p<.001$, $\eta_p^2=.49$）。健康食品（$M=208.89$）よりも不健康食品（$M=319.54$）でカロリーが高く見積もられていた。距離プライムの主効果は認められなかったが（$F(1, 167)=0.32$, *ns*, $\eta_p^2=.00$），有意に近い距離プライム × 食品の交互作用効果が認められた（$F(2, 167)=2.81$, $p=.06$, $\eta_p^2=.03$）。健康食品においては空間的距離プライミングの影響はみられなかったが，不健康食品においては近接条件（$M=325.05$），中庸条件（$M=334.85$）と比較して，遠方条件（$M=235.40$）の方がカロリーを低く見積もっていた。この結果は元の研究の仮説を支持するものであった。

さらに平均値パターン（Table 3）を考慮すると，近接条件の操作は影響力が小さく，本研究の結果は遠方条件操作によってもたらされているかもしれない。中庸条件を統制条件として考えた場合，遠方条件操作の影響は中庸条件と遠方条件との差によって検出されるはずである。そこで，近接条件を削除し，あらためて 2（距離プライム：中庸・遠方）× 2（食品：健康・不健康）の参加者間 1 要因参加者内 1 要因の混合計画による分散分析を行った。その結果，食品の主効果が認められた（$F(1, 110)=108.80$, $p<.001$, $\eta_p^2=.50$）。距離プライムの主効果は認められなかったが（$F(1, 110)=1.30$, *ns*, $\eta_p^2=.01$），有意な距離プライム × 食品の交互作用効果が認められた（$F(1, 110)=5.97$, $p=.02$, $\eta_p^2=.05$）。この結果は元の研究の仮説を支持した。

Table 3　事例 1：条件別にみたカロリー推定値（慶應データ，*n*=172）

距離プライム	近接（*n*=58）	中庸（*n*=63）	遠方（*n*=50）
健康（3 品目）	206.48（135.48）	206.55（73.68）	214.57（70.71）
不健康（3 品目）	325.05（191.82）	334.85（115.45）	294.17（97.13）

注　カッコ内は標準偏差

3. 事例 2：Schnall, Benton, et al.（2008, Experiment 1）の直接的追試

3.1 目的

Schnall, Benton, et al.（2008, Experiment 1）は，道徳判断における嫌悪感の影響を，清浄プライミングを用いることで検討した。Schnall, Haidt, et al.（2008）は，モラルジレンマ課題のような複雑な道徳判断において嫌悪感が関与すると主張し（Rozin, Haidt, & McCauley, 2000 も参照），嫌悪感を抱くほど不道徳と判断されやすいことを指摘した。Schnall, Benton, et al.（2008）は，この知見を拡張し，清浄感は嫌悪感を緩和すると考えられるので，清浄プライミングにより不道徳判断が緩くなると考えた。彼らの Experiment 1 は，単語課題を実施して清浄感や清潔感関連語を処理した後には，そのような処理を行わない場合と比較して不道徳行為が許容されやすくなることを示している。

Schnall, Benton, et al.（2008）の研究は，道徳的判断過程を考える上で重要な示唆を与えるが，その再現性に疑義がもたれ，議論となっている（Johnson et al., 2014；Schnall, 2014）。そこで，本研究では Schnall, Benton, et al.（2008, Experiment 1）の直接的追試を行った。統制プライミングと比較して，清浄プライミングをうけると道徳判断において不道徳行為が許容されるという条件差の再現を試みる。

3.2 方法

実験参加者　Schnall, Benton, et al.（2008, Experiment 1）は，道徳判断指標に対し 2 水準の参加者間 1 元配置分散分析を行い，$\eta_p^2 = .09$ の結果を得ている。この効果サイズで有意水準を 5% としたとき，80%，90%，95% の検定力を得るためには，それぞれ 82 名，110 名，134 名の参加者が必要となる。本研究では大学生 146 名が実験に参加した（男性 68 名，女性 77 名，不明 1 名，平均年齢 19.97 歳，*SD* = 1.47）。各大学の内訳は，一橋大学 67 名，関西学院大学 40 名，上智大学 39 名であった。これらの実験参加者は，統制条件と清浄条件に無作為配置された。その結果，統制条件 71 名，清浄条件 75 名となった。

実験素材[7]　実験は Schnall, Benton, et al.（2008）の直接的追試を行った Johnson et al.（2014）の公開マテリアルを翻訳，利用した。翻訳に際してはできるだけ原典に忠実であるよう心がけたが，一部，日本文化ならびに日本語の特徴に鑑みて軽微な修正を行った。

まず，清浄プライミング操作のための乱文構成課題を翻訳した。乱文構成課題は 4 単語から 3 単語を選択して文を構成するものであり，40 問あった。統制条件は，清浄さと関連しない語からなる課題であった。清浄条件では 40 問中 20 問が清浄関連語からなり，残り 20 問が無関連語からなった。この乱文構成課題を翻訳する際に 4 点留意した。第 1 に，定冠詞が課題に含まれる場合には，指示語にしたり，主語，目的語を用意したりした。第 2 に，直訳すると汚染語の否定になる表現はこれを避け，清浄語に直した。第 3 に，清浄に関する語は可能な限り訳し分けた。第 4 に，文にならない意味不明な項目は意味が通るように改編した。

次に，道徳判断課題 6 題をできるだけ忠実に翻訳した。各題は，飼い犬（事故死したペットの犬を食べる），トロッコ（5 人を救うため 1 人を犠牲にする），財布（見つけた財布の中身を盗む），飛行機（死にそうな生存者を殺害し，飢えをしのぐ），履歴書（履歴書に虚偽記載する），ネコ（ペットのネコを性の慰みものにする）であった。これらは，Schnall, Haidt, et al.（2008）が作成し，その後関連研究で利用されているものである。各場面の不道徳性を 0（完全に問題なし）から 9（非常に悪い）の 10 件法で回答を求めるものであった。

また，情動状態をたずねる 9 項目を用意した（Table 4 参照）。「全く感じない」と「非常に強く感じる」の両極を示した 11.5 cm の線分に印をつけることで回答を求めることとした。分析時には 5 mm 単位でカテゴリ化し 0 ～ 22 の値をつけた。Schnall, Benton, et al.（2008）では，10.5 cm の線分で 0–20 にカテゴリ化したが，技術的問題で本研究では 11.5 cm の線分で 0–22 のカテゴリとなった。

以上の翻訳から統制条件，清浄条件の 2 パター

7)　実験素材が Open Science Framework（https://osf.io/azukt/）から閲覧可能である。なお，研究実施に際しては研究実施機関の倫理審査を受け承認を得た。

Table 4　事例2：条件別に見た情動状態の平均値と標準偏差（全体データ，*N*=146）

	全体（*N*=146）	統制（*n*=71）	清浄（*n*=75）	*F* 値	η_p^2
くつろいだ	8.38（5.72）	9.41（5.81）	7.41（5.50）	4.55*	.03
怒った	3.25（4.87）	3.04（4.67）	3.45（5.08）	0.26	.00
しあわせな	5.62（5.25）	6.28（5.31）	5.00（5.14）	2.20	.02
かなしい	4.82（6.08）	4.73（5.92）	4.89（6.26）	0.03	.00
恐れた	4.94（5.85）	4.54（5.67）	5.32（6.02）	0.66	.01
沈んだ	7.42（6.70）	7.72（6.76）	7.13（6.67）	0.28	.00
吐き気のする	4.60（6.11）	4.25（5.75）	4.93（6.45）	0.45	.00
動揺した	7.55（7.01）	6.76（7.15）	8.29（6.84）	1.75	.01
混乱した	6.38（6.61）	6.07（6.71）	6.67（6.55）	0.30	.00

注　カッコ内は標準偏差。*：p<.05.

ンの質問紙を作成した。

手続き　集合状況で参加者に質問紙を配布，回答を求めた。回答に際しては倫理的配慮を行った。最初に不快と感じる内容が含まれる可能性を指摘し，いつでも実験参加を中止できること，回答拒否できることを保証した。また，座席を離す，質問紙回収は参加者がボックスに提出するなど，回答の匿名性を高める工夫を行った。これらを説明した上で，実験参加に同意した者だけが実験に参加した。

その後，実験者の指示に従い，質問紙への回答を進めた。最初に乱文構成課題に回答を求めた。回答には4分間設け，実験者が計時した。その後，実験参加者各自のペースで道徳判断課題，情動測定に回答した。全体でおよそ10分間を要した。回答終了後，説明資料を配付し，デブリーフィングを行った。

3.3　結果と考察

情動状態　清浄プライミング操作がこれらの情動測定に影響しているか検証するため，各項目に対して参加者間1元配置の分散分析を行った（Table 4）。その結果，「くつろいだ」で統制条件（M=9.41）よりも清浄条件（M=7.41）の方が低かった（$F(1, 144)=4.55$, $p=.04$, $\eta_p^2=.03$）。この結果は，Schnall, Benton, et al.（2008）ならびに Johnson et al.（2014）では見られなかった効果である。その他の項目では有意な効果が認められず（Fs<2.20），効果サイズも小さかった（η_p^2s<.02）。この点は先行研究と一致していた。

道徳判断　Schnall, Benton, et al.（2008）では道徳判断課題の回答を平均し指標化していたが，α係数は報告されていなかった。本研究ではα=.62となり，十分な信頼性が保証されているとは言い難かった。この全体指標と各項目の値を従属変数として参加者間1元配置の分散分析を行った（Table 5）。その結果，財布課題において有意な効果が認められた（$F(1, 144)=4.11$, $p=.04$, $\eta_p^2=.03$）。統制条件（M=6.55）よりも清浄条件（M=5.73）で不道徳でないと判断していた。全体指標とその他の項目では有意な効果は認められなかった（Fs<1.90, η_p^2s<.02）。

考察　多くの項目，特に道徳判断の全体指標において先行研究の結果は再現されなかった。情動状態，道徳判断の各1項目において清浄プライミングの影響が見られたが，Schnall, Benton, et al.（2008），Johnson et al.（2014）のいずれにおいても報告されていない現象であった。道徳判断の全体指標での結果が Schnall, Benton, et al.（2008）の核心であることを考えると，清浄プライミングが道

Table 5　事例 2：条件別に見た道徳判断の平均値と標準偏差（全体データ，*N*=146）

	全体（*N*=146）	統制（*n*=71）	清浄（*n*=75）	*F* 値	η_p^2
飼い犬	4.55（2.78）	4.83（2.81）	4.29（2.75）	1.37	.01
トロッコ	3.10（2.20）	3.31（2.29）	2.91（2.11）	1.23	.01
財布	6.13（2.46）	6.55（2.17）	5.73（2.65）	4.11*	.03
飛行機	6.11（2.63）	6.13（2.66）	6.09（2.63）	0.01	.00
履歴書	5.64（2.36）	5.72（2.23）	5.57（2.49）	0.14	.00
ネコ	2.83（2.70）	2.94（2.66）	2.72（2.75）	0.25	.00
全体	4.56（1.43）	4.72（1.41）	4.40（1.44）	1.85	.01

注　カッコ内は標準偏差。*：$p<.05$

Table 6　事例 2：条件別に見た情動状態の平均値と標準偏差（一橋データ，*n*=67）

	全体（*n*=67）	統制（*n*=33）	清浄（*n*=34）	*F* 値	η_p^2
くつろいだ	7.52（5.51）	9.48（6.08）	5.62（4.16）	9.27**	.13
怒った	3.27（4.60）	2.45（3.90）	4.06（5.12）	2.07	.03
しあわせな	4.52（5.04）	5.70（5.55）	3.38（4.27）	3.68†	.05
かなしい	4.09（5.67）	3.52（5.21）	4.65（6.11）	0.66	.01
恐れた	4.01（4.71）	2.88（3.76）	5.12（5.30）	3.96†	.06
沈んだ	6.48（6.04）	6.42（6.33）	6.53（5.84）	0.01	.00
吐き気のする	3.87（5.71）	2.91（4.70）	4.79（6.48）	1.85	.03
動揺した	6.33（6.20）	4.64（6.23）	7.97（5.80）	5.15*	.07
混乱した	5.18（5.51）	4.27（5.23）	6.06（5.70）	1.78	.03

注　カッコ内は標準偏差。**：$p<.01$，*：$p<.05$，†：$p<.10$

徳判断に及ぼす影響は再現されなかったと結論できる。

3.4　付加分析あるいは *p*-hacking の実践

本研究で最初に取得し，もっとも大きなデータでもある一橋大学データ（*n*=67，男性 47 名，女性 19 名，未回答 1 名；平均年齢 19.66 歳，*SD*=1.38）に着目した。

情動状態　各項目に対し参加者間 1 元配置の分散分析を行った（Table 6）。その結果，くつろいだ（$F(1, 65)=9.27$, $p=.003$, $\eta_p^2=.13$），動揺した（$F(1, 65)=5.15$, $p=.03$, $\eta_p^2=.07$）で有意な効果が，しあわせな（$F(1, 65)=3.68$, $p=.06$, $\eta_p^2=.05$），恐れた（$F(1, 65)=3.96$, $p=.05$, $\eta_p^2=.06$）で有意に近い効果が認められた。統制条件と比べて清浄条件の方が快感情を経験せず，不快感情を経験していた。Schnall, Benton, et al.（2008）や Johnson et al.（2014）では認められていない効果であった。

Table 7　事例 2：情動状態と道徳判断との相関係数（一橋データ，n=67）

	全体	飼い犬	トロッコ	財布	飛行機	履歴書	ネコ
くつろいだ	−.09	−.15	−.04	.04	−.21†	.25*	−.13
怒った	.25*	.33**	.04	−.11	.11	.02	.36**
しあわせな	−.10	−.08	−.03	−.08	−.17	.20	−.09
かなしい	.11	.26*	.10	−.24*	−.04	.01	.26*
恐れた	.20	.22†	.18	−.09	.14	.01	.16
沈んだ	.11	.14	.01	−.06	.11	.00	.11
吐き気のする	.09	.29*	−.09	−.15	−.04	.01	.23†
動揺した	.07	.14	.03	−.10	−.01	−.12	.25*
混乱した	.16	.23†	−.08	−.11	−.09	.17	.37**

注　**：p<.01，*：p<.05，†：p<.10

道徳判断課題の特徴　道徳判断課題 6 問のα係数は .49 となり，信頼性が保証されているとは言えなかった。課題項目間の相関係数を算出したところ，飼い犬課題とネコ課題との間にのみ正相関が認められた（r=.61, p<.001）。その他は無相関であった（rs<.25）。このことは，課題を合成することが妥当でない可能性を示す。

また，Schnall, Haidt, et al.（2008）では嫌悪感が道徳判断の核心にあると考えていた。そこで，実験時の情動状態と道徳判断との関連を検討するため，両者間の相関係数を算出した（Table 7）。その結果，一部の情動状態と道徳的判断との間に相関が認められた。特に，怒り，悲しみ，吐き気，混乱は飼い犬課題とネコ課題との間に正相関を示していた。このうち吐き気が嫌悪感に直接関連する項目であり，元の論文が仮定する関連が飼い犬課題とネコ課題の 2 課題でのみ見られたことになる。この飼い犬課題とネコ課題は，Schnall, Haidt et al.（2008）において嫌悪感を含むと仮定された 3 課題のうち 2 つであった。

道徳判断　道徳判断に対してプライムを条件とする参加者間 1 元配置の分散分析を行った。その結果，全体指標，項目のいずれにおいても有意な効果は認められなかった（Fs<1.87, η_p^2s<.04）。

しかし，本研究における清浄プライミング操作は，情動状態に影響していた。そして，情動状態は一部の道徳判断に影響を及ぼしていた。これらのことから，清浄プライミング操作が情動経験を経由して，仮説で想定していない間接的な影響を及ぼしている可能性がある。そのため，これまでの付加分析では仮説を支持する結果が得られなかった可能性がある。また，道徳判断の指標は信頼性が低かった。課題間の相関関係からは，飼い犬課題とネコ課題との間にのみ正相関が見られた。これらは Schnall, Haidt, et al.（2008）において嫌悪感を含む課題であると想定されたものである。そこで，本研究の道徳判断の指標としては，これら 2 課題の結果を平均したものが妥当であると考えた。

Table 8　事例 2：情動を統制した共分散分析の結果（一橋データ，n=67）

変数	F 値	η_p^2
くつろいだ	0.77	.01
怒った	2.41	.04
しあわせな	0.35	.01
かなしい	1.14	.02
恐れた	0.25	.00
沈んだ	2.20	.04
吐き気のする	0.05	.00
動揺した	0.34	.01
混乱した	1.40	.02
清浄プライミング	4.39*	.07

注　*：p<.05

以上をうけて，飼い犬課題とネコ課題の平均（M=3.04, SD=2.28）に対して，本研究で測定した 9 つの情動状態を共変量として投入し，清浄プライミング操作を要因とする共分散分析を実施した。その結果，9 つの情動状態はいずれも有意な効果は示さなかったが（Table 8），清浄プライミング操作は有意な効果を示した（F(1, 56)=4.39, p=.04, η_p^2=.07）。共変量を投入した調整平均をみ

ると，統制条件が3.67なのに対し清浄条件で2.42となり，清浄条件の方が低くなっていた。このことは，情動経験の影響を取り除いた場合，統制条件よりも清浄条件で嫌悪感を含みうる道徳場面での判断が緩くなる事を示している。この結果は，元の研究の仮説を支持するものであった。

4. 事例における *p*-hacking の指摘

事例1，事例2のいずれにおいても元の研究と同様の分析を行った結果は，元の研究と同方向の有意な効果が認められず，元の研究を再現しなかった。これらのデータは十分な検出力を備えたものであり，いずれの事例においても直接的追試に失敗したと結論づけられるものである。その一方で，付加的分析を行うことで，いずれの事例においても元の研究の仮説を支持する結果を得た。これらは *p*-hacking を行った結果である（平石・池田，2015；池田・平石，2016；Nosek et al., 2012；Simmons et al., 2011）。

4.1 データ選択による分析機会の増大

いずれの事例においてもデータ取得した大学を選択することでデータの削除を行った。いずれの事例ももっとも大きいデータとして，さらに事例2では最初に取得したとしてデータ選択をしたが，いずれに関しても理論的根拠はない。実のところ付加的分析においては，データを選択することで検出力が低まることを期待した。大学ごとに分析を分けることで，検出力の低い分析を複数機会行うことができ，いずれかの分析で仮説に対し肯定的な（もちろん偽陽性の）結果を得ようとしたのである（Ioannidis, 2005；Nosek et al., 2012）。

4.2 従属変数の選択と事後的な変数生成

次に，いずれの事例においても仮説に反する結果をもたらす変数を削除し，事後的に従属変数を変更，生成した。事例1においては複数の従属変数を個別に検討し，仮説に反する結果を示した従属変数を除外した。こうすることで仮説を支持する結果のみを残した。事後生成した従属変数で同様の分析をデータ全体（N=396）で実施した場合，仮説を支持する交互作用効果は得られない（F<1, η_p^2=.00）。その意味で，事例1の付加分析は，データを選択した状況で従属変数を取捨選択したあげく偽陽性な結果を得たと結論できる。

その後，後付けで，仮説を支持した従属変数の背後にある共通要因を探し出した。事例1における主食系，間食系という違いが理論的意味を持つかは著者らには不明である。しかし，探索的分析の結果見いだした何らかの効果にあたかも意味があるように考察することで，論文独自の新たな貢献可能性として出版されることもあるかもしれない[8)]。

事例2では元の研究（Schnall, Benton, et al., 2008）で信頼性の報告がなされていないことを逆手にとり，道徳判断課題間の相関分析を行った。そこで互いに関連のあった2測定だけを残した。これら2測定は否定的情動との関連があったので，この事実を，従属変数を取捨選択する正当化の理由として用いた。さらに，この2測定はSchnall, Haidt, et al.（2008）で嫌悪情動を含むとされたことまで説明した。

しかしながら，Schnall, Haidt, et al.（2008）で嫌悪情動を含むとされた課題には事例2で使用した飛行機課題も含まれるが，この変数を除外する理由は説明されていない。そもそもSchnall, Haidt, et al.（2008）は，道徳判断課題に嫌悪情動が含まれるか否かは嫌悪情動が道徳判断に影響することには関わらないと指摘している。これらのことから，事例2での従属変数の選択を理論的に正当化することは実は困難である。また，Schnall, Benton, et al.（2008）では従属変数に関するα係数は報告されていないため，信頼性を高めることが元の研究の結果を再現することにつながるのか自明ではない。

4.3 実験条件の削除

事例1では最終的に近接条件を分析から除外し，有意な交互作用効果を得ている。このような実験条件の削除はSimmons et al.（2011）も指摘している技法であり，偽陽性をもたらす可能性を高める。実際，同様の分析をデータ全体（N=396）

8) ここでは，論文の考察部分での後付け解釈が，その論文の独自貢献として認められる場面を想定している。その意味では従来の探索的分析の貢献と同じである。この後付け解釈があたかも最初から想定されていたように研究を再構成した場合，5.1で議論するHARKingになる。

で実施した場合，仮説を支持する交互作用効果は得られない（$F<1, \eta_p^2=.00$）。また，事例 1 において期待される効果を心的距離の増大にのみ求める理論的裏付けはない。実験条件を除外するにあたり中庸条件を除外することも可能であったはずだが，この場合，仮説に関わる交互作用効果は有意にならない（$F(1, 106)=3.11, p=.08, \eta_p^2=.03$）。事例 1 における実験条件の削除は明らかに *p*-hacking として行われている。

4.4　不要な共変量の投入

事例 2 では情動状態に関わる 9 変数をすべて共変量として投入し共分散分析を行った。情動状態の中には確かに道徳判断と相関を示したものも含まれるが，無相関であったものもあり，これらすべてを共変量として投入する積極的理由はない。また，共分散分析の結果においていずれの共変量も有意でなく（Table 8），過剰統制を行った可能性が高い。このような共変量の投入は，Simmons et al.（2011）も指摘する通り，偽陽性の可能性を高める。実際，同様の共分散分析をデータ全体（$N=146$）で実施した場合，仮説を支持する効果は認められない（$F(1, 135)=1.65, ns, \eta_p^2=.01$）。

5. 偽陽性を回避し再現性を高める

5.1　都合の悪い報告の抑制

事例 1，事例 2 で示した *p*-hacking の技法は指摘されずとも認識可能かもしれない。これらは理論的にも統計技法的にも明らかに不適切であり，容易に疑義を挟むことができるようにも思われる。その一方，*p*-hacking の普及度に鑑みると（John et al., 2012），多くの研究論文において *p*-hacking が見過ごされている可能性がある。このような看過の背景には，都合の悪い報告を抑制するという問題ある研究実践がある。

事例 1，事例 2 では実施した *p*-hacking をすべてその結果とともに記載をした。そのため，多重検定を実施していること，有意とならなかった分析結果が多く含まれることが明らかであった。これに対し，有意にならなかった結果を敢えて報告しないことで，多重検定を行った事実の隠蔽が可能になるかもしれない（平石・池田，2015；Simmons et al., 2011）。意図的に隠蔽しないとしても，紙幅の都合を考え，結果の記述を単純化，省略することもあるだろう。その結果，*p*-hacking を行った事実が隠蔽される可能性が高まる。また，場合によっては，最初から予測できていたかのように論文を作成することも可能かもしれない。このような事後的仮説生成（Hypothesizing After the Results are Known）は HARKing と呼ばれる。HARKing を行う事で，読みやすい論文作成が可能になる一方，*p*-hacking を行った事実が隠蔽される可能性が生じる（Kerr, 1998）。

5.2　研究の事前登録（pre-registration）

p-hacking を可視化して偽陽性の結果報告を減少させるためには，都合の悪い報告の抑制と HARKing に基づく論文作成をできなくする実践が必要である。その方法の一つとして研究の事前登録がある。研究の事前登録とは，研究目的，実験参加者数，実験手続き，分析方法を事前に Open Science Framework（https://osf.io/）等へ公的に登録することを指す。

研究の事前登録には様々なメリットがある。研究計画を事後に変更できなくすることで *p*-hacking の実践とその隠蔽の可能性が低まるのである。たとえば，研究目的の事前登録は，HARKing を不可能にする。実験参加者数の登録は，検定力の小さい研究を複数実施することや選択的なデータ収集に対して抑制的に働く可能性がある。いくつかの心理学雑誌がサンプルサイズの事前決定とその根拠の明確化を推奨するに至っているが（Association for Psychological Science, 2015；Varize, 2016），事前登録と併用することで *p*-hacking 抑制効果が期待できる。

さらに，手続きおよび分析方法の事前登録は，*p*-hacking それ自体を難しくする。事前登録にない分析は論文を批判的に検討することを促すので，*p*-hacking の検出が容易になる。少なくとも，事前登録された分析以外の分析はすべて探索的分析とみなすことが可能になる。また，本研究の事例 1 および事例 2 で行ったような付加分析を事前に予告することは，それを正当化する理由がない限り *p*-hacking と見なされる可能性が高くなる。

これは探索的分析の事前登録を否定するものではない。理論上あるいは方法論上の理由を明示して探索的分析を事前登録することは正当な行為で

あろう。たとえば，予備研究において探索的に複数の変数を用意して事前登録し，これらの変数を統制変数として取捨選択しながら本実験の統制状況を検討することもあるだろう。また，事後的な探索的分析であっても，事後に行ったことを明確にしてその結果をすべて公表することは許容されるだろう。ただし，事前，事後のいずれであったにせよ，探索的検討で得られた結果に関しては，新たに研究を立案し追試すべきである。探索的分析と事後解釈は，探索的であり事後であることが読者に明確に理解されるべきであり，仮説を支持すると言及したり，事前から予測されていたものと再構成したりしてはならない。事前登録を通じて仮説検証たる分析と探索的分析の違いを研究者が明確に知ることが，実験科学としての議論を保証することにつながるのである。

5.3 否定的な結果の公表

研究を事前登録することで結果の偽陽性の可能性が低まるとすれば，それは仮説を支持しない陰性の研究結果が増えるということでもある。帰無仮説検定を用いている現状では，仮説を支持する結果に比べて，帰無仮説を採択する陰性の結果から直接的に主張できることは少ない。その結果，陰性の結果よりも，偽陽性も含めた陽性の結果の方が出版されやすくなる出版バイアスが生じやすくなる。

実は，研究知見の再現性は特定データ単独の結果からは保証されない。仮説を支持する陽性の結果であっても第一種の誤りの可能性があるし，何の結果も得られない場合でも第二種の誤りの可能性がある[9]。現状では，複数の直接的追試によって得られた効果サイズをメタ分析的に検討することが，再現性の判断にはもっとも有効であろう（e.g., Klein et al., 2014）。その一方で，出版バイアスによって陽性の結果のみが報告されてしまうと，メタ分析の元となる研究に偏りが生じてしまう。

その意味では，仮説を支持しない陰性の結果も含めて，どのような結果であれ出版されることが望ましい。前節 5.2 で事前登録の重要性を主張したが，結果の公表に関しても事前登録が重要な役割を果たすだろう。研究を事前登録することで，その研究の実施が公開される。結果がもし公表されなければ，何らかの不都合が存在することを示すことになろう。場合によっては外部第三者から結果を請求することも可能になる。現状では陰性の結果の出版は難しいが，PsychFileDrawer（http://psychfiledrawer.org/）などの登録機関を利用して公表することは可能である。このように実施された研究結果が隠匿されることなく出版されることで，効果サイズのメタ分析の精度が上がり，研究知見の再現性を示すことができることだろう。

理想論かもしれないが，再現性検証に資するデータを蓄積するには，事前審査（pre-review）制度の導入が有効であろう。事前審査制度とは，研究計画の段階で査読を行い，どのような結果であれ雑誌に掲載するという制度である。事前審査制度を取り入れた雑誌である *Comprehensive Results in Social Psychology* 誌が 2016 年に刊行予定であり，今後の普及が期待される。

6. 最 後 に

その有効性に関する議論の余地はあるとはいえ，社会心理学，特に社会的認知領域は実験を通じた実験科学の方法論に基づき理論構築をしてきた。このようなアプローチをとる限り，研究知見の再現性は一定以上保証される必要がある。ここ数年議論されてきた社会心理学における再現性問題は，学界における自省を促すものであり，新しい知見と同じぐらいに確実な知見が重要であることを再認識させた。本論文で示した *p*-hacking をはじめとする問題ある研究実践を改め，再現性を検証しながら研究知見を積み重ねることで，社会心理学および社会的認知領域においてその理論に関わる根本的な議論が再活発化することを期待したい。

9) その意味で，本論文の事例だけでは各研究の再現性に結論をだすことはできない。しかしながら，本研究の目的は *p*-hacking の危険性を指摘することにあった。

文　　献

Association for Psychological Science (2015). Replication in psychological science. *Psychological Science*, *26*, 1827–1832.

Bargh, J. A. (2012). Priming effects replicate just fine, thanks: In response to a ScienceNews article on priming effects in social psychology. *Psychology Today*. Retrieved from

https://www.psychologytoday.com/blog/the-natural-unconscious/201205/priming-effects-replicate-just-fine-thanks

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, *71*, 230–244.

Bargh J. A., Gollwitzer, P. M., Lee-Chai, A., Barndollar, K., & Trötschel, R. (2001). The automated will: Nonconscious activation and pursuit of behavioral goals. *Journal of Personality and Social Psychology*, *81*, 1014–1027.

Barsalou, L. W., Niedenthal, P. M., Barbey, A. K., & Ruppert, J. A. (2003). Social embodiment. *Psychology of Learning and Motivation*, *43*, 43–92.

Clark, H. H. (1973). Space, time, semantics, and the child. In T. E. Moore (Ed.), *Cognitive development and the acquisition of language* (pp. 27–63). Academic Press.

Doyens, S., Klein, O., Pichon, C-L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE*, *7*, e29081.

Fauconnier, G., & Turner, M. (2002). *The way we think: Conceptual blending and the mind's hidden complexities*. Basic Books.

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.

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.

Harris, C. R., Coburn, N., Rohrer, D., & Pashler, H. (2013). Two failures to replicate high-performance-goal priming effects. *PLoS ONE*, *8*, e72467.

平石 界・池田功毅（2015）心理学的な心理学研究：Questionable Reasearch Practice　心理学ワールド, *68*, 5–8.

池田功毅・平石 界（2016）心理学における再現可能危機：問題の構造と解決策　心理学評論，*59*, 3–14.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, e124.

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.

Johnson, D. J., Cheung, F., & Donnellan, M. B. (2014). Does cleanliness influence moral judgments? A direct replication of Schnall, Benton, & Harvey (2008). *Social Psychology*, *45*, 209–215.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*, 196–217.

Klein, R. A. et al. (2014). Investigating variation in replicability: A "Many Labs" replication project. *Social Psychology*, *45*, 142–152.

Lakens, D. (2014). Grounding social embodiment. In D. C. Molden (Ed.)., *Understanding priming effects in social psychology* (pp. 175–190). Guilford Press.

Lakoff, G. (2012). Explaining embodied cognition results. *Topics in Cognitive Science*, *4*, 773–785.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.

Landau, M. J., Meier, B. P., & Keefer, L. A. (2010). A metaphor-enriched social cognition. *Psychological Bulletin*, *136*, 1045–1067.

LeBel, E. P., & Wilbur, C. J. (2014). Big secrets do not necessarily cause hills to appear steeper. *Psychonomic Bulletin & Review*, *21*, 696–700.

Lynott, D., Corker, K. S., Wortman, J., Connell, L., Donnellan, M. B., Lucas, R. E., & O'Brien, K. S. (2014). Replication of "Experiencing physical warmth promotes interpersonal warmth" by Williams & Bargh (2008). *Social Psychology*, *45*, 216–222.

Mandler, J. M. (1992). How to build a baby: II. Conceptual primitives. *Psychological Review, 99*, 587–604.

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.

大久保街亜（2016）帰無仮説検定と再現可能性　心理学評論，*59*, 57–67.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349*, 943.

Pashler, H., Coburn, N., & Harris, C. R. (2012). Priming of social distance? Failure to replicate effects on social and food judgments. *PLoS ONE*, *7*, e42510.

Rozin, P., Haidt, J., & McCauley, C. (2000). Disgust. In M. Lewis & J. M. Haviland (Eds.), *Handbook of Emotions* (2nd ed. pp. 637–653). New York: Guilford.

Schnall, S. (2014). Clean data: Statistical artifacts wash out replication efforts: Commentary and Rejoinder on Johnson, Cheung, and Donnellan (2014a). *Social Psychology*, *45*, 315–317.

Schnall, S., Benton, J., & Harvey, S. (2008). With a clean conscience: Cleanliness reduces the severity of moral judgments. *Psychological Science*, *19*, 1219–1222.

Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (2008). Disgust as embodied moral judgment. *Personality and Social Psychology Bulletin*, *34*, 1096–1109.

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.

Slepian, M. L., Masicampo, E. J., Toosi, N. R., & Ambady, N. (2012). The physical burdens of secrecy. *Journal of Experimental Psychology: General*, *141*, 619–624.

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, *37*, 1–2.

Trope, Y., & Liberman, N. (2003). Temporal construal. *Psychological Review*, *110*, 403–421.

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review*, *117*, 440–463.

Varize, S. (2016). Editorial. *Social Psychological and Personality Science*, *7*, 3–7.

Williams, L. E., & Bargh, J. A. (2008a). Experiencing physical warmth promotes interpersonal warmth. *Science*, *322*, 606–607.

Williams, L. E., & Bargh, J. A. (2008b). Keeping one's distance: The influence of spatial distance cues on affect and evaluation. *Psychological Science*, *19*, 302–308.

— 2016. 3. 10　受稿，2016. 3. 15　受理 —