Japanese Psychological Review
2016, Vol. 59, No. 1, 98–107

心理学のデータと再現可能性

渡　邊　芳　之
帯広畜産大学

Psychological data and their reproducibility

Yoshiyuki WATANABE
Obihiro University of Agriculture and Veterinary Medicine

Herein, I discuss some methodological aspects on the reproducibility of psychological data. Reduced reproducibility of psychological data can occur as a consequence of research misconduct, inadequate statistical methods, effects of uncontrolled latent variables, and low occurrence probability of the studied phenomena itself. As such, the validity of data and their analyses may be subject to these sources of irreproducibility, and the expected reproducibility level depends on theoretical and methodological characteristics of the studied phenomena and related variables. Until recently, psychologists' historical reluctance for replication studies, derived from demonstrational and anecdotal usage of psychological data, has prevented psychologists from considering the issues related to reproducibility. Some possible alternatives for psychologists addressing reproducibility issues are also discussed.

Key words: reproducibility, latent variables, evidence, demonstration, anecdotes
キーワード：再現可能性（再現性），潜在変数，エビデンス，デモンストレーション，逸話

「心理学の再現可能性」という問題にはとてもたくさんのことが関わってくるが，この論文ではまず低い再現性がいったい何を反映しているのかということと，心理学でなぜこれまで再現可能性が大きな問題にならなかったのかということについて考えてみたい。そのうえで，心理学がこれからそれぞれの問題にどのように対処していくべきかについて検討する。なお，ここではまだ行われていない研究や追試における可能性も含めたマクロな問題として再現性を考えるときにだけ「再現可能性」という言葉を用い，具体的な研究と追試において生じる問題については「再現性」という言葉を用いることにする。

1.「再現性のなさ」のいくつかのすがた

再現性を「ある実験や調査とそのデータの分析によって得られた結果が，それと同じ実験や調査が反復されたときに再現されること」というごく狭い意味に考えても，再現性が低くなる原因には大きく分けて，当初の実験や調査の手続きやデータ，その分析方法に何らかの間違いや「正しくなさ」があるために再現性が低くなる場合と，当初の実験や調査の手続きやデータや分析には間違いがないにも関わらず再現性が低くなる場合の2つがあると思われる。そして，その2つのなかにもまたいくつかのパターンがある。そのうち，まず当初の実験に間違いがある場合から考える。なお「実験や調査」「反復される実験」と繰り返すのは煩雑なので，これ以降は実験や調査などのデータ収集手続きを代表して「実験」とよび，それが反復されることを「追試」とよぶ。

1.1　実験が間違っている場合

当初の実験の手続きやデータ，分析が間違っているために再現性が低くなる場合の最悪の例は，データが捏造されたり変造されたりしていることである。捏造や変造とまでは言えなくても，実験仮説が支持されやすいような方向にデータを取り増すなどいわゆる「問題のある研究実践」（QRPs: Questionable Research Practices，藤島・樋口，2016；池田・平石，2016）も実験結果を誤ったものとし，

その実験が捏造や変造などのない「正しい」方法で追試され分析されたときに得られる結果と一致しないようにするので，再現性を低める。

研究者にそのような意図がなくても，実験要因を操作したり測定したりする手続きに信頼性がないときや，データの統計的分析に問題があるとき（大久保，2016）には，その実験結果と「正しい」（あるいはまた別の方向に間違った）追試結果とは一致しなくなるために，やはり再現性が低くなる。

こうした場合，当初の実験結果として報告される内容は実験の状況や文脈で実際に起きたことを正確に反映していない点で妥当でないために，追試で得られる妥当な結果（あるいは別の方向に妥当でない結果）とは一致せず，再現性が低くなる。こうした場合には，データの捏造や変造，問題のある研究実践を防ぐとともに，実験方法や測定の信頼性を高めたり，統計分析の方法を改善したりすることによって再現性が高まることが期待される。再現性を高めるためのそうした取り組みについては，この特集の他の論文でそれぞれくわしく検討されている（池田・平石，2016；大久保，2016）。

1.2　実験が間違っていない場合

しかし，再現性が低くなるのは必ずしも当初の実験結果が妥当でない場合だけではない。当初の実験には特に不正や問題点はなく，実験要因の操作や測定にも信頼性があって，データの分析も問題なく行われているにも関わらず，追試でそれが再現されないということがありうる。

「心理学の再現可能性」が議論されるときには，どちらかというと先に述べたような「当初の実験に不正や間違いがあって結果が妥当でない場合」とその解決法のほうが注目される傾向がある（たとえば三浦，2015）。しかし，実際にはそうした問題がなくても実験結果が再現されないことはあり，そうした場合のほうが心理学の再現可能性にとってより大きな問題であるようにも思われる。こうした場合に，当初の実験は実験状況で起きたことを妥当に報告しているにも関わらず，追試でそれが再現されないということが起きるのはなぜだろうか。

そうした場合のシンプルな例が，実験の手続き（論文でいう「方法」のセクション）が十分にくわしく記述されていなくて，追試が当初の実験の状況や文脈を正しく再現できていないために，結果が再現されない場合である。もちろんこれは心理学に限ったことではなく，生命科学・医科学で発表される論文の多くにおいて，その研究方法を正確に再現できるような情報が欠けているという報告がある（Iqbal et al., 2016）。

こうした場合には，それぞれの研究者が実験結果を報告する際に実験の方法や手続き，要因の操作や測定の方法，統計分析の方法などをできる限りくわしく記述して，追試において実験条件や実験状況を正確に再現できるような情報を提供することが重要であり，それによって再現性が向上することが期待できる。そのためには，論文の著者がそのように努力するだけでなく，学術誌の側にも詳細な記述を可能にし促進するような工夫が必要だろう（池田・平石，2016）。さきにこの問題をシンプルと形容したのは，比較的容易に実行可能な解決方法が示されるからである。

1.3　潜在変数がある場合

それでは，当初の実験やデータ分析が正しく行われ，その方法が詳細に記述されて，当初の実験状況が追試で正確に再現されれば再現性は担保されるかというと，必ずしもそうではない。潜在変数が実験結果に影響を与えている場合と，研究の対象となっている現象自体の生起確率が低い場合には，依然として再現性が低くなりうる。まず潜在変数がある場合から考える。

当初の実験結果が，研究者も認識していない潜在変数の影響を受けていると，その変数は追試で統制されない。その結果として，当初の実験結果が追試で再現されるかどうかは，追試の実験状況で潜在変数が確率的にとる値に依存するようになる。

ある実験結果がそれを最初に行った研究室では繰り返し再現されるのに，他の研究室が追試を試みると再現されないという「研究室効果」とでもいうべき現象はよく見られる。こうした場合の多くでも，オリジナルの実験状況の中に実験結果に強く影響する潜在変数があるが，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ことで，追試の再現性が低まる（澤・栗原，2016）。

こうした潜在変数には古くから「実験者効果」とよばれたような，実験者と実験参加者（被験者）との間で社会的に構成される状況要因も含まれる（西阪，2001）。

潜在変数の影響は，そうした潜在変数を発見して実験要因に加えたり，系統的誤差として統制したりすることができれば低減でき，再現性が高まると期待できる。先に述べた，実験方法の記述が不十分なために再現性が低くなる場合でも，実験者が気づいていない変数は論文に記述されないし，いっぽうで実験状況を詳細に記述して再現可能にすることは，潜在変数もそれに付随して再現される可能性を高めるから，実験結果の再現性を高めるだろう。

1.4　現象自体の生起確率が低い場合

そうした潜在変数の多くが発見されて，操作されたり統制されたりするようになっても，依然として再現性が向上しない場合には，研究の対象になっている現象自体がもともと生起確率の低いもので，同じ条件が繰り返し与えられても低い確率でしか再現しない現象であることが考えられる。

木から落ちる枯葉がどのようなルートを経てどこに落ちるかは，風や温度などの条件が同じでもそのつど異なり，同じルートで同じ場所に落ちる確率は非常に低い（中谷，1958）。世界にはそうした現象は他にも存在するだろうし，人間行動にもそうした性質のものは多いと思われる。そうした現象を取り扱う場合には，これまで述べたような再現性低下の原因をすべて解決したとしても，その実験結果の再現性はその現象がもともと持っている生起確率に依存するようになるだろう。

同じように再現性が低くても，潜在変数がある場合と現象自体の生起確率が低い場合とでは，前者は影響の大きい潜在変数を統制できていないという点で当初の実験が妥当でないといえるが，後者では実験結果は現象自体の生起確率を正しく捉えているのだから妥当といえる。

しかし現実には，観察された再現性のなさが潜在変数の存在によるのか，現象自体の生起確率が低いのかを区別することは困難である。潜在変数が発見されれば実験結果の変動が潜在変数の影響であることはわかるが，潜在変数が発見されない段階では，その再現性のなさが潜在変数の影響なのか，現象自体の生起確率の影響なのかは原理的に区別できないからである。先の落ち葉の例でも，枯葉の落下ルートや落下場所が未知の変数によって決定されていて，それを発見し統制できれば同じ落ちかたを再現できる可能性はつねに残っている。

もしその事象に影響するすべての潜在変数が発見されれば，潜在変数の影響と現象自体の確率性を切り分けることができるが，とくに非実験系の心理学が取り扱うような現象で，その現象に影響するすべての変数が特定され，操作や統制が可能になるということは考えにくい。多くの現象は潜在変数の影響と現象自体の生起確率とが混交した力によって変動し，再現性が低くなる。同時に，実験結果が「潜在変数を特定できていない」という点で妥当でないのか，現象を正しく捉えているので妥当なのかも判断できないままになる。

ただし，この区別ができないのは研究者だけでなく，そうした現象を実際に体験する一般の人々にとっても同様である。現象に影響する変数で研究者が発見できないようなものは一般の人にも不可視であることがほとんどであり，素朴な認知は潜在変数の影響と現象自体の生起確率を区別することができない。したがって一般の人からはこうした現象の多くは，実際には何らかの潜在変数によって強く影響されていても「ある条件のもとでときどき起きること」と捉えられ，それが社会的な現実になっていると考えられる。その意味では，潜在変数と現象自体の生起確率の混交によって生じる再現性の低さは「現実」を妥当に反映しているとも言える。

1.5　再現性と妥当性

このように，実験結果の再現性が低くなるのは，必ずしも実験やデータ，分析方法に問題があるときだけではなく，再現性の低さがなんらかの現実を正しく反映しており，妥当である場合もある。しかし，ある実験の結果が追試で再現されない時に，それが実験が妥当でないからなのか，実験自体は妥当だが再現されないのかを，実験のデータ自体から判断することはできない。また，再現性がある程度認められるとして，いったいどのくらいの再現性が必要なのか，たとえば39%という再現性（Open Science Collaboration, 2015）

は十分なのか不十分なのかということも，やはりデータの外側にあるなんらかの基準と照らし合わせなければ判断できない。それでは，こうした再現性と実験やその結果の妥当性との関係はどのようなことから判断できるだろうか。

第一に，他の知見やデータとの関係がある。心理学の再現可能性が話題になり，心理学全般の再現性が39%「しか」ないという報告（Open Science Collaboration, 2015）を見た時に，私がすぐに連想したのはパーソナリティの一貫性問題である（渡邊，2010）。パーソナリティの一貫性問題では，性格検査などのパーソナリティ・アセスメントが，それによって測定される構成概念と関係するような実際の行動を30%くらいしか予測できない，ということが指摘され（Mischel, 1968），パーソナリティ・アセスメントの妥当性や有用性に疑問が呈されたために，大きな論争になった[1)]。50年近くを経た結論からいえば，パーソナリティ・アセスメントの行動予測力が30%程度であるというMischelの主張にはその後も効果的な反証はなく，それを改善するための具体的な努力や技術開発も特に行われていない。

しかし，50年の間にパーソナリティ・アセスメントが否定されたり，使われなくなったりすることはなかった。パーソナリティ心理学にさまざまな生理学的指標や遺伝的指標が導入され（国里・山口・鈴木，2007；De Young et al., 2010），統計的な分析手法や測定技法が飛躍的に進歩した中でも（森尾，2013；清水，2013），質問紙法性格尺度に代表されるさまざまなパーソナリティ・アセスメントは依然としてパーソナリティを把握するための中心的な技法として用いられ続けている。このようにパーソナリティ・アセスメントの妥当性や有効性が疑われない大きな原因の一つは，アセスメントの結果が他の多くの指標と相関したり，他の指標の将来の値を予測したりしていることと考えられる。

パーソナリティ・アセスメントの結果はさまざまな生理学的指標や遺伝的指標，経済状態や学歴，健康や幸福感など，実に多くの外的な指標と相関を示すし，アセスメント結果からそうした指標の将来の値を予測することもできる（Ozer & Benet-Martínez, 2006）。このことは，パーソナリティ・アセスメントがそれらの外的基準と連動するような「なにか」を確かに測定しているという基準関連妥当性を担保している。だからパーソナリティ・アセスメントは「行動を30%しか予測できなくても」使われ続けるし，そのことが結果的に一貫性論争を止揚してしまったのが現在だと言える。

同じことは再現可能性の問題についても言えるだろう。比較的低い再現性しか示されないような実験結果であっても，その結果が（概念的に関係づけられるような）他のさまざまな現象や事象と関係していることが示されれば，それが「なにか」を妥当に捉えている可能性は高まるし，同時に低い再現性が持つ意味（間違いなのか，現象の生起確率が低いのか）についての考察も発展すると思われる。実験結果の妥当性は再現性と外的基準との関係とを組み合わせて検討されるべきだろう。

第二に，理論的考察の重要性がある。そもそも，心理学の実験結果の再現性はどのくらいあればよいのか。100%でなければいけないということはないとして，80%ならいいのか，50%でいいのか，それどころか39%でも特に問題はないのか，こうしたことも実験やそのデータの内部では判断できないことである。

たとえば求められる再現性は，説明や予測の対象となる現象の重大性・深刻性や，現象への対処可能性によっても大きく異なる。「戸締りしなかった窓から泥棒が入る」という現象は確かにあるが，それがどのくらい再現するか，家の窓に鍵を締めなかったらどのくらいの確率で泥棒が入るかといったら，その再現性はそれほど高くなく，せいぜい数%しかないと思われる。しかし家に泥棒が入ることは非常に深刻な現象なのでわれわれは戸締りをするし，戸締りすることで（もともと低い）泥棒が入る確率をさらに低下させることができる。飛行機事故の原因解明などでも同様で，同じ条件が揃えば10%の確率で墜落するならば，その条件は改善されるべきものになる。求

1)　心理学研究の測定やデータ，妥当性が信頼できない，という議論ではしばしば「再現性が30%しかない」「予測力はわずか30%である」というように30%前後の数値が「足りないもの，少ないもの」として提示される。私はこの30%という数字を「心理学係数」と名づけることを提唱する。

められる再現性は現象の性質によって異なり，それはデータの外側で，理論的に分析されるべきことである。

以上のことから，心理学の再現可能性問題については，まず実験の手続きやデータ，データの分析方法，方法の記述のしかたなどの問題から生じる再現性の低さを具体的な手続きによって低減した上で，研究対象とする現象の性質や，それに影響すると思われるさまざまな要因を踏まえて，必要な再現性の大きさや性質についての理論的考察を進めることが重要といえる。

2. デモンストレーションとしての心理学データ

ここまで，心理学における再現可能性を低めるさまざまな要因とその含意について簡単に検討した。さて，こうした要因が心理学の実験や調査の結果の再現性を低めるようになったのは最近のことではなく，ずっと以前からそうだったはずだが，それが今まであまり問題にならず，最近になって急に問題にされるようになったのはなぜだろう。そのひとつの原因は，心理学がこれまで追試にあまり熱心でなかったことである。

自然科学の多くの分野では新しい重要な知見が報告されると，関係する研究を行っている世界中の研究室で追試が行われて，繰り返し再現された知見だけが事実として残っていくことが多い。少し前には常温核融合が（渡辺，2001），最近ではSTAP細胞（De Los Angeles et al., 2015）が，そうした追試に耐えられずに否定されるという運命をたどっている。いっぽう，心理学で重要な知見が報告されても，そのように世界中で追試されるということはこれまであまりなかった。別の研究の中で先行研究の知見が再検討され，検証されることはあるが，そのときには研究対象も，データの収集方法も分析方法も先行研究とは異なっていることが多く，追試としてのその結果の解釈にはさまざまな問題が生じる（三井，1981）。卒業研究など教育の文脈で行われる追試は比較的多いが，その結果は公式には報告されない。

追試されないならば，その再現性が問題になることもないので，これまで心理学の再現可能性はほとんど問題にされなかった。皮肉なことに，日本の心理学においてもっとも多く追試されたのは，心理学がみずから主張したのではない「血液型性格関連説」である（佐藤・渡邊，1992；最近では縄田，2014）。しかしこの場合には心理学者の側に「血液型と性格との関連を否定したい」という強い動機づけがアプリオリにあり，じっさい追試（で関連が再現されないこと）によってそれが否定されるという結果になっているので，これも本来の意味での追試とは呼びにくい[2)]。

2.1　デモンストレーションとしてのデータ

これまで心理学が追試に熱心でなかったこと，言いかえれば心理学では追試によって確認されなくても実験や調査の結果がそのまま受け入れられやすく，疑われなかったことにはさまざまな原因が関係していると思われる。そのひとつの大きな理由は，心理学におけるデータが，エビデンスとしてよりもデモンストレーションとして機能していたことにあると考えられる。ここでいうエビデンスとは，それによって仮説や理論が検証，反証されるようなデータのことをいう。いっぽうデモンストレーションとは，仮説や理論の検証や反証のためではなく，その仮説や理論の正しさをアピールするために提示されるデータのことをいう。

デモンストレーションの代表的なイメージは，学校の理科の授業で行われる実験である。理科の授業では，実験されている知識は既に確立されたもので，実験を行う目的は知識の検証や反証ではなく，それが取り扱っているような現象について生徒に実際に体験させ，理科の知識についての理解を深めることが目的となっている。デモンストレーションとしてのデータの意味は，その理論や仮説が正しいかどうかを検証・反証することではなく「こんなことが本当に起きるんだよ」ということを多くの人にアピールし，理解させることにある。

2)　Bem（2011）の「予知研究」とその追試（Ritchie, Wiseman, & French, 2012）をめぐる一連の出来事も，研究が主張する知見が心理学者の常識や他の心理学的知識と整合しないことで追試の対象になった，という点でわが国における血液型性格関連説の追試と類似した性質を持っている。また，そうした試みが結果として心理学の研究方法そのものへの疑問や批判につながった（佐藤・渡邊，1992）という点でも類似していると言える。

2.2　社会心理学とデモンストレーション

「心理学の実験」としてもっともよく言及されるもののひとつにS. Milgramの服従実験，いわゆる「アイヒマン実験」がある（Milgram, 1963；Blass, 2004）。この実験は，ナチスドイツによるユダヤ人虐殺が当時のドイツ人の「狂気」や「異常」の結果ではなく，「普通の人々」の命令への服従によって粛々と行われた，ということを実験的に示したものとして歴史に残っている。

さて，Milgramは実験以前に「人は命令への服従によって残虐になりうる」という仮説を持っていて，それを実験で確かめたわけだが，もしこの実験が失敗に終わり，誰も「XXX」（危険なレベルを超えた最高電圧）まで電圧を上げ続けたりしなかったならば，Milgramはその仮説を捨てたり，再検討したりしただろうか。あるいは「人が命令への服従によって残虐になることはない」と結論しただろうか。おそらくそうはならなかっただろう。実験の失敗はあくまで実験方法や実験状況の問題で，Milgramは「人が命令への服従によって残虐になりうる」ことを示すことのできる別の実験方法や実験状況を引き続き探し続けただろうし，実験の失敗がそのまま理論や仮説への反証とはならなかったと思われる。同じ意味で，実験が追試で再現されなくても，Milgram自身が成功させた実験のデモンストレーションとしての意味が大きく揺らぐことはなかっただろう[3)]。

Milgramの時代の社会心理学者には「人間は状況によって大きく影響される」「どんな人でも状況によって善良にも残虐にもなり得る」という状況主義のイデオロギーが強く共有されていて，それはデータによって確認されるまでもない理論的前提であった。もちろんそうしたイデオロギーが共有されることには第二次世界大戦の悲惨な体験があり，またMilgram自身も含めて当時の社会心理学にはユダヤ系の研究者が多かったことも影響していたと思われる（Blass, 2004）。彼らにとっての実験やデータの意義はそうした理論的前提を検証したり反証したりすることではなく，そうした理論から予想される現象が実際に起きる，ということをデモンストレーションすることだっただろう。

こうしたことはMilgramの実験に限らず，線の長さの判断において生じる同調行動を明らかにしたS. Aschの線分組み合わせ実験（Asch, 1955），監獄に似た状況に置かれた人々が容易に看守や囚人の役割に陥ったP. Zimbardoのスタンフォード監獄実験（1971年，Zimbardo, 2008）など，1950～1970年代のいわゆる社会心理学黄金時代に行われた社会心理学の実験の多くに言えることである。そこでは実験やそのデータは研究者の理論や仮説が正しいかどうかを検証するために用いられているのではなくて，研究者の理論から予想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して，社会に対して何かの示唆や警告を与えることが重視されていた。

「人は命令への服従によって残虐になりうる」といった仮説は，たとえば「シロバナスイートクローバーが産生する成育阻害物質クマリンの産生量は個体へのストレスによって増加する」という仮説のようにデータによるエビデンスが伴わなければ正しいとも間違っているとも想像がつかないような仮説とは違って，常識や関連する様々な知識，人類が積み重ねてきた過去の経験などからデータ以前に「そうなのではないか」「そうともいえるのではないか」と誰でも想像できるものである。

歴史的に，とくに非実験系の心理学で実験や調査によって検証される仮説のほとんどはこのようにもともと「条件や状況によってはそうだろう」「そういう場合もあるだろう」と想像のつくもので，もしそれが実験で検証されなくても，その仮説自体が完全に否定されるわけではなかったし，実験の失敗が研究者にその仮説を捨てさせることもなかった。いっぽうで実験が成功すれば，研究者はその事実を「そういうことが本当にあるのだ」「この仮説が正しいのだ」ということを自分以外の人に納得させ，受容させるために用いることができた。

いずれにしても，心理学の実験の成否は研究者

3)　もっとも，Milgramの実験自体については追試によって結果がおおむね再現されたという報告がある（Burger, 2009）。また，MilgramやZimbardoの実験があまり追試されなかったことには倫理的な問題も影響していると思われるが，倫理的な理由で追試できないということもデータのエビデンスとしての価値を低めると思われる。追試できないことがとくに知見の価値を低めなかったことも，これらの実験データがデモンストレーションであったことを示している。

の持っている仮説そのものを検証したり反証したりするエビデンスとしての役割はあまり果たしておらず，むしろ仮説やそれの基礎となる理論，思想，イデオロギーなどの正しさを（おもに研究者以外の）人々に納得させるためのデモンストレーション」として機能していた。デモンストレーションされるのは個別の心理学理論だけでなく，遺伝論や環境論のような巨大な「パラダイム」であるときもあったし，人間の可塑性や可能性についてのイデオロギーであったり，ナチスや全体主義への批判のような政治的主張であったり，教育の方法や枠組みに関する主張であったりした。そして，デモンストレーションとしての説得力はそれが「科学的な実験」という枠組みで行われることでますます強まった。

2.3　デモンストレーションと逸話

同じような目的のデモンストレーションとしては，古くから逸話や事例が用いられてきた。たとえばアヴェロンの野生児や「アマラとカマラ」の事例は人間の発達に幼少期の経験や環境が大きな影響を与えるということのデモンストレーションとして活用された。最近になってこれらの事例が事実でない，あるいはかなり事実と違って報告されている，ということが明らかになってきたが（鈴木，2008），そのことによって「人間の発達に幼少期の経験や環境が大きな影響を与える」という仮説が否定されるかというと，そうはならないだろう。逸話や事例はエビデンスではなくデモンストレーションなので，事例が否定されることがそのまま仮説の否定にはつながらない。忠犬ハチ公の逸話（一ノ瀬・正木，2015）は「犬は飼い主への強い忠誠心をもつ」ということのデモンストレーションだが，もしハチ公が亡き飼い主を待っていたのではなく屋台の客が餌をくれるから渋谷駅に毎日通っていたのだとわかっても，「犬は飼い主への強い忠誠心をもつ」という考え自体は否定されないのと同じである。

どのような逸話が採用され，事実とされるかは理論に依存し，逸話は理論を検証したり反証したりしない。したがって逸話が疑われたり，否定されたりすることは理論の説得力の低下の結果として起きる。アマラとカマラやアヴェロンの野生児の逸話が疑われ，事実関係が検証されるようになったのは，人間発達の環境論の影響力が徐々に低下して遺伝や生得的な要因の力が見直されるようになったという理論やイデオロギー，時代精神の変化の結果であって，逸話が疑われるようになったから環境論の説得力が低下したわけではない。その意味でも逸話はエビデンスではなくデモンストレーションなのである。

いっぽうで，逸話や事例というのはその目撃談や記録の信頼性が常に疑われるものである。もし仮説のデモンストレーションが科学的に統制され，客観的に記録される実験や調査という枠組みの中で実現すれば，その説得力は著しく高まる。そういう現象が少なくとも一回は間違いなく起きた，ということが科学的な実験の枠組みで保障されるからである。それを最大限に活用したのが，とくに第二次大戦後から1970年代までの社会心理学であった。先に挙げたMilgramやZimbardoの実験はその代表的なものである。そこでは，それまでおもに実験心理学の分野で蓄積されてきた心理学実験の方法論が活用されて，「科学的に統制され記録された逸話」が次々と提供されていったが，それらが追試によって検証されることは多くなかった。それは研究者の怠慢や背徳ではなくて，もともとデモンストレーションとしての実験は追試される必要がなかったし，もし追試で再現されなくてもそれが理論や仮説の反証になるとはあまり考えられなかったのである。この点で，現在の心理学の再現可能性問題が，おもに社会心理学で最も大きな問題になっているのは当然とも言える。

2.4　デモンストレーションからエビデンスへ

こうした傾向に変化が生じて，心理学の実験・調査のデータの位置づけがデモンストレーションからエビデンスへと顕著に変化してきたのは1980年代以降と思われる。諸領域で「大きな仮説」や「大きな差」がひと通り研究され尽くして，研究者がより小さな仮説，より小さな差を研究対象にするようになると，実験・調査のデータはデータがなくても常識的にありそうな大きな仮説のデモンストレーションから，データがなければ判断がつかないような小さな仮説のエビデンスへとその性質を変えていったし，データの意味も心理学の外側への説得やアピールから，心理学内

部での納得の基準，研究者個人の判断の基準へと変化していった（尾見・川野，1996）。

こうした変化には，さまざまな生理学的指標や脳神経科学的な指標，生命科学に強く影響を受けたモデルなどが心理学に導入されて，データとエビデンスに関する自然科学や医学保健学的な考え方が心理学の中に広がってきたことも関係しているだろう。EBM（Evidence Based Medicine）の思想は，もともと実証に裏付けられない医学理論や臨床経験によって医療が行われることへのアンチテーゼとしての性質が強く，それはデータの逸話的な用法を強く断罪するものでもあった（伊藤，2002）。ただ，その結果が実験や調査によるエビデンスの確立にではなく，むしろ実験や調査によるデータの再現性のなさが問題にされることにつながったのには，心理学者の考え方やデータ収集，データ分析の方法が，追試を必要としないデモンストレーションの時代からあまり大きく変化しなかったことと，前に考えた潜在変数や現象自体の生起確率の問題が関わっている。

研究が扱う現象や仮説の枠組みが大きなものから条件特定的な小さいものに変化すると，一般的には対象とする現象の生起に影響する変数の数は少なくなるだろうから，発見されない潜在変数の影響も小さくなって，結果の再現性が高くなることが期待される。しかし，非実験系の心理学が扱うような現象では，もともと関係する潜在変数の数が多く，影響が大きいために，仮説の枠組みが小さくなっても潜在変数の影響が再現性を高めるほどには減少しないかもしれない。その結果，仮説の枠組みが小さくなっても扱う現象の確率的な変動があまり小さくならず，再現性も高まらないということがあるだろう。

このように，心理学が追試をあまり行わず，これまで再現可能性の問題にあまり注意を払わなかったことには，心理学の実験や調査のデータが歴史的に，事実を検証・反証するエビデンスとしてよりも理論的立場やイデオロギーの説得力を高めるためのデモンストレーション，科学的に統制され記録される逸話としての性質を強く持っていたことが関係している。それが変化して，心理学のデータも自然科学と同様にエビデンスとしての性格を期待されるようになったことと，心理学のデータが依然として持っている特性との不一致が，現在の再現可能性問題を生んでいると考えることができる。

3. おわりに

ここまで，心理学の再現可能性が低くなる理由，そしてこれまで心理学が追試に熱心でなく，再現可能性の問題にあまり興味を持たなかった理由について考えてきた。それでは，これから心理学者はこうした問題にどのように対処していくべきだろうか。

まず，実際には再現性のある現象についてのデータから再現性を奪うような，不正や捏造，変造を追放するとともに，方法論上，統計解析上の問題を解決して，実験のデータを「正しいもの」にする必要があることは前にも述べた。その上で，自分たちが研究対象にしている現象について，どのくらいの再現性が求められ，かつ可能であるのかについての理論的・方法論的な検討をさらに進めて，問題ごとに必要な再現可能性をあきらかにして，その実現を目指すことが求められるだろう。

その上で，再現可能性を「目的」にしないことが重要である。データや研究成果の再現性はしばしば「科学であること」の条件とされているために，再現性は高ければ高いほど望ましく，それによって心理学も科学の仲間入りができるというような考えが起きやすい。しかし，先にも述べたようにわれわれ心理学者が扱う現象には非常に多くの変数が関わっていて，そのわずかな変動によっても心理現象は大きく変動するし，そうした変数のうちかなりの部分は今後も潜在変数のままで残る。それだけでなく，対象とする現象自体の生起確率が低い可能性も常に残る。

必要な再現性の大きさは研究対象や研究の目的，方法によって異なるはずだし，低い再現性が検出されることのほうが現象を妥当に反映している可能性もある。データの再現可能性はあくまでもわれわれが研究対象とする現象を正しくとらえるために参考にできるさまざまな指標の一つに過ぎない。再現可能性のために他の重要なことを犠牲にする必要はなく，自分たちの研究や対象や研究目的に応じて，現象を正しくとらえるための他の基準とうまくバランスをとりながら，適切な再

現性のレベルを決めて実現すればよい[4)]。

心理学に必要なのは「科学」の特定の基準に自らを適合させることではなく，心理学がその対象と目的にしたがってどのような科学になるのか，心理学の科学性はどのように実現されるのかを自ら考えることである。行動主義の時代に心理学が「科学」になるために研究できる対象，使える方法を極端に狭めてしまったことは忘れてはいけないことで，再現可能性もそれが目的化されれば，再現可能性を高めやすい研究対象や研究方法に研究の可能性が限定されて，第二の行動主義時代を招いてしまうことは容易に想像できる。再現可能性の向上はあくまでも心理学の研究をよりよく行うための手段に過ぎず，心理学の目的ではないのである。

文　　献

Asch, S. (1955) Opinions and social pressure. *Scientific American*, *193*, 31–35.

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.

Burger, J. M. (2009) Replicating Milgram — Would people still obey today? *American Psychologist*, *64*, 1–11.

Blass, T. (2004) *The man who shocked the world: The life and legacy of Stanley Milgram*. Basic Books. 野島久雄・藍澤美紀（訳）（2008）服従実験とは何だったのかスタンレー・ミルグラムの生涯と遺産　誠信書房.

De Los Angeles, A., Ferrari, F., Fujiwara, Y., Mathieu, R., Lee, S., Lee, S., Tu, H., Ross, S., Chou, S., Nguyen, M., Wu, Z., Theunissen, T. W., Powell. B. E., Imsoonthornruksa, S., Chen, J., Borkent, M., Krupalnik, V., Lujan, E., Wernig, M., Hanna, J. H., Hochedlinger, K., Pei, D., Jaenisch, R., Deng, H., Orkin, S. H., Park, P. J., & Daley, G. Q. (2015) Failure to replicate the STAP cell phenomenon. *Nature*, *525*, E6–E9.

De Young, C. G, Hirsh, J. B., Shane, M. S., Papademetris, X., Rajeevan, N., & Gray, R. (2010) Testing predictions from personality neuroscience: Brain structure and the big five. *Psychological Science*, *21*, 820–828

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.

一ノ瀬正樹，正木春彦（編）（2015）東大ハチ公物語―上野博士とハチ，そして人と犬のつながり　東京大学出版会.

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.

Iqbal, S. A., Wallach, J. D., Khoury, M. J., Schully, S. D., & Ioannidis, J. P. A. (2016) Reproducible research practices and transparency across the biomedical literature. *PLoS Biol*, *14*: e1002333.

伊藤幸郎（2002）EBMの思想的基盤　生命倫理，*12*, 25–31.

国里愛彦・山口陽弘・鈴木伸一（2007）パーソナリティ研究と神経科学をつなぐ気質研究について　群馬大学教育学部紀要　人文・社会科学編，*56*, 359–377.

Milgram, S. (1963) Behavioral study of obedience. *Journal of Abnormal and Social Psychology*, *67*, 371–378.

Mischel, W. (1968) *Personality and assessment*. New York: Wiley. 詫摩武俊（監訳）（1992）パーソナリティの理論―状況主義的アプローチ　誠信書房.

三井宏隆（1981）社会心理学における追試の問題　実験社会心理学研究，*20*, 171–178.

三浦麻子（2015）心理学研究の「常識」が変わる？～心理学界における再現可能性問題への取り組み　心理学ワールド，*68*, 9–12.

森尾博昭（2013）潜在的な個人差の測定　二宮克美・浮谷秀一・堀毛一也・安藤寿康・藤田主一・小塩真司・渡邊芳之（編）　パーソナリティ心理学ハンドブック（pp. 669–675）福村出版.

中谷宇吉郎（1958）科学の方法（岩波新書青版313）岩波書店.

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.

西阪 仰（2001）心と行為～エスノメソドロジーの視点　第5章　心理学の社会学　岩波書店，pp. 173–218.

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.

Ozer, D. J., & Benet-Martínez,V. (2006) Personality and the prediction of consequential outcomes. *Annual Review of Psychology*, *57*, 401–421.

尾見康博・川野健治（1996）納得の基準　―心理学者がしていること―　東京都立大学人文学部　人文学報，*269*, 31–45.

Open Science Collaboration (2015) Estimating the reproducibility of psychological science. *Science*, *349*, aac4716.

小塩真司（2016）心理尺度構成における再検査信頼性係数の評価―「心理学研究」に掲載された文献のメタ分析から―　心理学評論，*59*, 68–83.

Ritchie, S. J., Wiseman, R., & French, C. C. (2012) Failing

4)　実際に求められる「適切な再現性のレベル」はこの論文で述べたように研究者が持つ研究目的や理論的背景によって決まるだけでなく，心理学者間の社会的合意にも大きく影響されるだろう（小塩，2016）。これから各領域でそうした合意形成が進むことが期待される。

the future: Three unsuccessful attempts to replicate Bem's 'Retroactive Facilitation of Recall' Effect. *PLoS ONE*, *7*: e33423.
澤 幸祐・栗原 彬（2016）動物心理学における再現可能性の問題　心理学評論，*59*, 46–56.
清水和秋（2013）構造方程式モデリング　二宮克美・浮谷秀一・堀毛一也・安藤寿康・藤田主一・小塩真司・渡邊芳之（編）　パーソナリティ心理学ハンドブック（pp. 669–675）福村出版.
佐藤達哉・渡邊芳之（1992）現代の血液型性格判断ブームとその心理学的研究　心理学評論，*35*, 235–269.
鈴木光太郎（2008）オオカミ少女はいなかった～心理学の神話をめぐる冒険　新曜社.
渡辺 正（2001）常温核融合騒ぎの顛末（疑似科学を通して考える）　化学と教育，*49*, 688–691.
渡邊芳之（2010）性格とはなんだったのか―心理学と日常概念　新曜社.
Zimbardo, P. G. (2008) *The Lucifer effect: Understanding how good people turn evil*. New York, Random House. 鬼澤 忍・中山 宥（訳）（2015）ルシファー・エフェクト　ふつうの人が悪魔に変わるとき　海と月社.

― 2016. 3. 10　受稿，2016. 3. 27　受理 ―