University of Groningen

Are most published social psychological findings false?

Stroebe, Wolfgang

Published in:
Journal of Experimental Social Psychology

DOI:
[10.1016/j.jesp.2015.09.017](https://doi.org/10.1016/j.jesp.2015.09.017)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2016

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Stroebe, W. (2016). Are most published social psychological findings false? *Journal of Experimental Social Psychology*, 66, 134-144. <https://doi.org/10.1016/j.jesp.2015.09.017>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Are most published social psychological findings false?☆

Wolfgang Stroebe

Department of Social and Organizational Psychology, University of Groningen, Grote Kruisstraat 2/1, 9712TS Groningen, The Netherlands

ARTICLE INFO

Article history:

Received 24 October 2014

Revised 28 September 2015

Accepted 29 September 2015

Available online 24 March 2016

Keywords:

Bayesian analysis

False positives

Methodology

One-shot studies

Replicability

Mass replication

Ioannidis' claim

ABSTRACT

Based on Bayesian reasoning, Ioannidis (2005) made the bold claim that most published research findings are false. His claim has been widely cited. It also seems consistent with the findings of the Open Science Collaboration Project that a majority of psychological studies could not be replicated. In this article, I argue (1) that Ioannidis' claim has limited relevance for social psychology and (2) that mass replication does not allow general conclusions about the validity of social psychological research. Ioannidis' claim is valid only for one-shot studies without replication and with a low a priori probability that the tested hypothesis is true. Mass replication provides limited information about the validity of social psychological research, because failures to replicate do not prove that the original finding is invalid. More conclusive information is provided by meta-analytic tests of social psychological theories. Implications for research in social psychology are discussed.

© 2015 Elsevier Inc. All rights reserved.

Ioannidis (2005) postulated in a provocative article that most published research findings are false. This claim, which he based on Bayesian reasoning, had great impact on the scientific community. By September 2015 his article had received 1619 citations in the Web of Science. When psychologists (and particularly social psychologists) became concerned about the replicability of their research findings, the Ioannidis claim was frequently cited in support of the argument that we were in the midst of a replicability crisis. For example, in an article in the special section on replicability in *Perspectives on Psychological Science*, Pashler and Harris (2012) asked themselves whether the replicability crisis was overblown; answering their own question in the negative, they suggested that "Ioannidis's devastating surmise ("Most published results are false") could very easily be the case throughout our field" (p. 532). This is a serious accusation that is given additional weight by the apparent failure of mass attempts to replicate many findings of social psychological research.

I argue in the first part of this article that Ioannidis' claim has little relevance for social psychology. I then show in the second part that the findings of mass replications allow only limited conclusions about the validity of social psychological research. Instead, the question of validity should be addressed with meta-analyses, which use a superior approach to the synthesis of research findings, particularly when these

are guided by a strong theory. Finally, implications for social psychological research will be discussed.

1. The claim of Ioannidis

1.1. What is a false research finding?

Ioannidis (2005) defines research findings "as any relationship reaching formal statistical significance" (p. 0696). Such a research finding is false, if it cannot be replicated, because this suggests that the relationship claimed in a study does not actually exist. There can be two reasons for such false claims. First, and this is the reason exclusively considered by Ioannidis, a false claim could be due to statistical problems (e.g., lack of power). Using the terminology of test theory, Ioannidis' claim refers to the reliability of an experimental procedure in producing the same result on repeated occasions. Second, a claim could also be false because some third variable was associated with the assumed cause and this variable was actually responsible for the observed relationship (i.e., confounding). Thus, even though the observed relationship between two variables can be replicated, the causal interpretation of the finding might be false. Despite being reliable, the experimental procedure might have low validity.

In evaluating the implications of the Ioannidis' analysis for social psychology, one needs to take into account the scientific context in which his claim was made. John P. Ioannidis is a professor of medicine who specializes in epidemiology and his analysis is based on epidemiological examples, more specifically on genetic epidemiology. Epidemiology is the study of the determinants of health-related states or events

☆ The author is grateful to Georgios Abakoumkin, Gisela Böhm, Alice Eagly, Volker Gadenne, Miles Hewstone, Toon Kuppens, Bernard Nijstad, Tom Postmes, Fritz Strack and Martijn van Zomeren for helpful comments on earlier versions of the manuscript.

E-mail address: wolfgang.stroebe@gmail.com.

and their distribution in specified populations. In epidemiological research, finding an association between a risk factor and some disease state is usually the *final* answer to a research question. For example, once we know that a sedentary life-style is an important risk factor for heart disease (i.e., reliably increases the risk of heart disease), public health campaigns can be devised to change this life-style. Typically, such campaigns are designed for being maximally effective rather than for optimally testing of the theory on which they are based.

In contrast, social psychologists typically also want to know whether their findings are supportive of the theory being tested. In most of their research, finding a reliable association between two variables is only part of the answer. Even if the use of experimental methods succeeded in ruling out confounding variables as one threat to validity, it still leaves open the questions of construct validity. Construct validity refers to the validity of the assumption that independent and dependent variables reflect the theoretical constructs they are assumed to represent (e.g., Cook & Campbell, 1979). Thus, conclusions based on social psychological research are not only endangered by false positive findings, but also by measures and manipulations that do not adequately represent the theoretical constructs they are supposed to reflect.

1.2. Why most published research findings may be false

There is ample reason to assume that all scientific tests are flawed in so far as they sometimes confirm a hypothesis that is actually false (i.e., type I error or false positive). Sometimes a test might also fail to confirm a true hypothesis¹ (type II error or false negative). The probability that the test will reject the null hypothesis when that hypothesis is false reflects the power of that test. The “probability that a research finding is indeed true depends on the prior probability of it being true (before doing the study), the statistical power of the study, and the level of statistical significance” (Ioannidis, 2005, p. 0696). This post-study probability that a research finding is true is referred to as the Positive Predictive Value (PPV).

Following Diekmann's (2011) insightful discussion of the Ioannidis (2005) article, I will illustrate Ioannidis' argument using the example of a medical test, namely the mammogram for breast cancer. Thus, instead of asking about the probability of a hypothesis being “true” if it is supported by a positive test result, we ask about the probability of a woman having breast cancer, if she received a positive mammogram result. According to Bayesian analysis, the answer to this question does not only depend on the accuracy of the test, but also on the prior probability of the woman having cancer. This prior probability for a woman aged 40 is approximately one percent. If women have breast cancer, a mammogram will detect it in 80% of the cases. Roughly 10% of mammograms give erroneous positive results (i.e., detect breast cancer, if it is not there). Thus, if we have 10,000 women being tested, 100 will have breast cancer and of those 100, 80 will have a positive mammogram. However, of the 9900 women, who do not have breast cancer, 990 (i.e., 10%) will also have a positive mammogram. If we want to know the percentage of women with a positive mammogram, who actually have breast cancer, we have to divide the 80 with a correctly positive mammogram by all the women with a positive mammogram (i.e., 990 plus 80). Thus, only about 7.47% of women with a positive mammogram will actually have breast cancer and the error rate is 92.53%. This surprisingly high error rate is largely due to the fact that the percentage of women having breast cancer (i.e., the prior probability) is rather low. If this percentage had been higher (e.g., 10%), the percentage of women with a positive mammogram,

who actually had breast cancer, would have been considerably higher [$80 / (900 + 800) = 47\%$] and thus the error rate much lower (53%).

We now translate this into the language of science and assume that there is a population of 10,000 hypotheses, of which only one percent is correct (i.e., 100). If the power of a statistical test is 80%, the chance of the null hypothesis being incorrectly accepted is 20%. Thus, of the 100 true hypotheses, 80 will be detected by the test. Given that the chance of the null hypothesis being incorrectly rejected is 5%, of the remaining 9900 false hypotheses, 495 will be wrongly accepted as correct. Thus, the probability that a significant result actually identifies a true hypothesis (the PPV) is $80 / (495 + 80) = 0.139$. Conversely, the probability that a significant result is a false positive is $1 - 0.139$ or 0.861. Thus, according to this example, most published research findings would indeed be incorrect. But this outcome is largely determined by the low percentage of correct hypotheses.

1.3. How do we know the prior probability that a hypothesis is valid?

It should be apparent from this discussion that Ioannidis (2005) claim that most research findings are false is valid only if the prior probability that the tested hypothesis is valid is very low *and* if the empirical test of the hypothesis has low power. In social psychology, as in most sciences, the prior probability of any hypothesis being valid cannot be known in any precise way.

Ioannidis (2005) uses an example from genetic epidemiology to illustrate “science at low pre-study odds”. He argues that in a genome-wide association study of the genetic determinants of susceptibility to schizophrenia, one assesses which of the 100,000 gene polymorphisms will be associated with schizophrenia. Based on existing knowledge of the heritability of the disease it can be assumed that approximately 10 polymorphisms will actually be associated with the disease (Ioannidis, 2005). Thus the pre-study probability of any of the total population of polymorphism being associated with schizophrenia is extremely low, namely $10/100,000 = 10^{-4}$.

Genetic epidemiology is atypical – even for epidemiological research – because all the genes making up the human genome have been identified in the human genome project that was completed in 2003. As a result the number of potential risk factors (i.e., hypotheses) has become known in genome-wide association studies. In other areas of epidemiology, this is not the case. For example, if one studies the environmental impact on cancer, nearly anything could be a potential risk factor. It is therefore not surprising that this type of research has resulted in conflicting findings. Two decades ago, in an article entitled “Epidemiology faces its limits”, Taubes (1995) complained that over “the years such studies have come up with a mind-numbing array of potential disease-causing agents, from hair dye (lymphomas, myelomas, and leukemia) to coffee (pancreatic cancer and heart disease) to oral contraceptives and other hormone treatment (virtually every disorder known to woman)” (p. 164). And a study that randomly selected 50 ingredients from cookbooks found that 40 of these ingredients had research that reported on their cancer risk (Schoenfeld & Ioannidis, 2013).

With thousands of potential determinants and huge data sets, a data-driven (i.e., atheoretical) search is likely to identify significant associations that are unlikely to be replicated in alternative data sets. These are the kind of findings on which Ioannidis (2005) claim is based. In contrast, if the genetic epidemiology of schizophrenia were guided by a strong theory, one would expect that the 100,000 potential genetic determinants could be reduced to a smaller biologically plausible number. If hypotheses are based on such a strong theory, their prior probability of being true will be greatly improved (Diekmann, 2011). And the better the theory, the greater this prior probability, and, everything being equal, the greater the post-study probability that a research finding is true (i.e., the positive predictive value).

It is therefore surprising that in genetic epidemiology these data-driven genome-wide association studies are often considered a

¹ I would like to emphasize that whenever I refer to hypotheses as true, it is in the sense used by Ioannidis that a relationship between two *empirical* variables is replicable. Because theories cannot be proven to be true (Popper, 1959), it would not make sense to refer to a hypothesis about the relationship between two theoretical concepts (e.g., the effect of ego depletion on self-control) as true.

methodological advance over theory-guided studies in which “a few candidate risk factors were selected based on diverse considerations (often under the umbrella of “biological plausibility”) that suggested their possible importance” (Ioannidis, Tarone, & McLaughlin, 2011, p. 451). According to Ioannidis, these candidate gene studies resulted in substantial selective reporting of “positive findings” that – not surprisingly – were rapidly refuted by subsequent studies” (Ioannidis et al., 2011, p. 451). However, even though in the meantime genome-wide association studies seem to have become the standard paradigm for studying genetic associations, results have been disappointing and the replication rate has remained low (Ioannidis et al., 2011). The problem with genetic epidemiology appears not only to be caused by the great number of potential factors, but also by the apparent lack of theories that would allow the identification of the most likely determinants.

In social psychology, as in most other sciences, the total number of potentially valid hypotheses for addressing a specific research question is typically unknown and unknowable. Even if a hypothesis is based on a well-supported theory, there could always be alternative hypotheses explaining the finding.² Furthermore, even a well-tested theory can include some false assumptions resulting in research findings that are inconsistent with theoretical predictions (Slife & Williams, 1995). Thus, in contrast to genetic epidemiology, in social psychology there is no *objective* basis on which to determine the prior probability of a hypothesis being true before the study.

1.4. The role of theory

Because conducting research requires time, money and other scarce resources, scientists tend to focus on hypotheses that they expect to be valid and to be supported by their studies. In such a research program, every new study provides some information about the likelihood that hypotheses derived from that theory will be empirically supported. If the core assumptions of a theory are reasonably valid, and if in the course of conducting empirical tests researchers improve their ability to develop good operationalizations of the theoretical concepts involved, experimental support for their theoretical hypotheses will become increasingly more likely. Thus, even though we may not know the prior probability of the validity of a *specific* hypothesis, we can assume that hypotheses that are logically deduced from well-corroborated scientific theories have a greater likelihood of being valid than hypotheses based on ad hoc hunches (Schaller, 2016–in this issue). Since from a Bayesian perspective, the prior probability of a hypothesis being valid (before doing the study) is one of the determinants of a subsequent research finding being “true” (Ioannidis, 2005), a discipline that mainly focuses on theory-guided research has a greater likelihood of producing valid findings than a field that engages mainly in a theoretical research.

² A good example is the critique of the classic study by Aronson and Mills (1959) by Gerard and Mathewson (1966). Aronson and Mills tested the dissonance hypothesis that the severity of the initiation into a group increases the liking for a group, particularly if that group turns out to be pretty worthless. Under the guise of a screening test for their suitability for a discussion group, female college students in the severe initiation condition had to read texts from obscene literature and a list of dirty words out loud to the experimenter. When the group later turned out to be engaged in the discussion of the sex life of lower animals, the severely initiated woman liked the group more. Although these findings were supportive of dissonance theory, opponents of dissonance theory suggested a long list of plausible alternative explanations (Gerard & Mathewson, 1966). For example, reading the dirty words could have sexually aroused the women in the severe initiation condition to a greater extent than those in the mild initiation and they might therefore have been more keen to join a sexual discussion group. Or they might have been intrigued by the obscene material and hoped that it would be discussed in the group at some later time. Or the women in the severe initiation condition might have been relieved when they heard the innocuous discussion of the actual group and therefore liked it more. Or according to the “afterglow hypothesis”, participants in the severe initiation condition might have had a greater sense of accomplishment than the women in the mild initiation condition. Finally, according to the “contrast” hypothesis, the group discussion simply appeared more positive after the unpleasant severe than the more pleasant mild initiation.

However, some people fear that testing hypotheses that according to some theory have a high probability of being valid will result in what Kelley (1992) referred to as “Bubba” psychology, namely a body of knowledge that even our grandmother (i.e., our Bubba) knew to be true. In contrast, scientific innovation seems to require examining hypotheses that predict *nonobvious* findings and have a *low probability* of being valid. Popper (1959, 1963) certainly appeared to have favored the latter position, when he wrote: “It is easy to obtain confirmation, or verification, for every theory, if we look for confirmations. Confirmations should count only if they are the result of *risky predictions*; that is to say, if, unenlightened by the theory in question, we should have expected an event which was incompatible with the theory – an event which would have refuted the theory... Every genuine test of a theory is an attempt to falsify it, or to refute it” (p. 36).

This triviality concern is based on the confusion of a number of issues. First, the requirement to test hypotheses that are logically derived from a well-corroborated theory merely implies that those hypotheses have a greater chance of being empirically supported than hypotheses based on ad hoc observations. Whether these findings will be trivial will depend on the theory and not on the way the hypotheses were deduced. If the theory is a bold advance over present knowledge and makes numerous nonobvious predictions, then hypotheses derived from that theory add to existing knowledge. In contrast, if the theory is trivial (i.e., “Bubba” psychology) then hypotheses derived from that theory are also likely to be trivial.

Admittedly, some predictions derived even from a bold theory may be less risky than others. Therefore the requirement of testing “risky predictions” poses a psychological dilemma for researchers. They may not be motivated to falsify their own theories. They also may like to use their limited resources in a way that ultimately produces publishable results, if not for their own purpose, then for that of their graduate students. However, once researchers have developed a theory and published supporting research, they have the responsibility to test it in the strictest manner, even if they fear that such a test may fail to support the theoretical predictions. In the long run, such non-supportive results may help to improve and refine the theory in question.

Furthermore, the rewards of empirically supporting nonobvious predictions are much greater than those of demonstrating the obvious.³ It is doubtful that without Festinger and Carlsmith’s (1959) counterintuitive finding of greater attitude change under small (rather than large) rewards, dissonance theory (Festinger, 1957) would have become as famous as it did. Predicting an inverse relationship between incentive and attitude change was definitely “risky” in the sense that “unenlightened by the theory in question, we should have expected an event which was incompatible with the theory” (Popper, 1963, p. 36).⁴

³ The acclaim that can be earned with the publication of counter-intuitive findings is responsible for the vast number of sexy findings derived from ad hoc theories that are traditionally published in social psychology. An article published in 1967 by Kenneth Ring indicates that this is no recent phenomenon. In a section on “Fun and games in social psychology”, Ring (1967) complained that “clever experimentation on exotic topics with a zany manipulation seems to be the guaranteed formula for success... One sometimes gets the impression that an ever-growing coterie of social psychologists is playing ... a game of ‘can you top this?’ (p. 117). Whereas his complaint was probably directed at some of the dissonance research published in those days, I can think of several social psychological research areas it would apply today.

⁴ Whether one expects that testing a risky prediction will falsify one’s theory depends on the trust one has in the validity of that theory. A famous example for this is Einstein’s (alleged) answer to a journalist after Eddington’s test of the prediction of the general theory of relativity about the gravitational deflection of light. Although Newton’s theory also predicted some deflection, Einstein’s theory predicted twice the amount derived from Newton’s theory. In 1919, Eddington led expeditions to Brazil and to the island of Principe, to observe the solar eclipses of May 29, 1919 and thereby test the Einstein’s theory of general relativity. The purpose was to determine whether or not there is a deflection of light by the sun’s gravitational field and with which theory the extent of the deflection would be consistent. The findings supported Einstein rather than Newton. When asked what he would have done if the experiment had not supported his theory, he is supposed to have answered “Then I would feel sorry for God” (Trefil, 2015, p. 83).

Nevertheless, in conducting increasingly stricter tests of a theory, it is always possible that theory-inconsistent empirical findings will emerge. If other core assumptions of a theory have been empirically supported, this will initially lead to a modification of the theory to accommodate the inconsistent findings. However, extended non support should persuade researchers to relinquish their theory and attempt to develop a new theory that is not only consistent with the existing evidence but also allows them to generate novel predictions (Popper, 1959). Since this is the only way a science can develop, one-shot studies based on ad hoc theories contribute little to our knowledge. We can neither be confident that the finding would replicate nor that the explanation of the finding has any validity.

1.5. The role of methodology

The probability of published research findings being false (i.e., a false positive) does not only depend on the prior probability of our hypotheses being valid, but also on the quality of our research methods (e.g., Ioannidis, 2005). In conducting research we need to employ methods that maximize the probability of identifying valid and rejecting invalid hypotheses. In this section, I will argue that one reason why social psychological research findings may be more valid than those of epidemiology is the greater feasibility of using experiments.

“The randomized experiment is the gold standard to identify and test causal relationships” (Antonakis, Bendahan, Jacquart, & Lalive, 2014, p. 93). However, conducting experiments is often difficult or impossible in epidemiology. For example, in the two areas, which have interested me, namely the increase in mortality risk following the death of a partner (e.g., Stroebe, Stroebe, & Schut, 2007; Stroebe, Stroebe, Gergen, & Gergen, 1981) and the association between firearm availability and homicide rate (Stroebe, 2013), experiments would have been impossible for ethical or factual reasons. But even with issues such as the health effects of specific foods (e.g., Mediterranean diet), the practical problems of making large groups of people keep to a specific diet are simply insurmountable.

The major research tools in epidemiology are therefore observational population studies and case–control studies. In population studies, such as the classic Framingham study (Dawber, 1980), large population samples are observed over decades. Such studies allow one to assess the association between risk factors measured early in the study (e.g., smoking) to health outcomes many years later (e.g., heart disease, lung cancer). Case–control studies compare individuals who suffer from some kind of disease to a control group of healthy individuals with regard to the relative presence of some risk factor. The one area where experiments are possible and randomized controlled trials are standard is drug intervention studies.

1.6. The strength of social psychological research methods

Experiments have three advantages over observational studies: (1) Because the independent variable is manipulated, the experimenter can be confident that experimental and control conditions differ only with regard to the variable being studied and (2) that the experimental manipulation preceded any change in the dependent variable. (3) Finally, due to random assignment of participants to experimental conditions, the experimenter can be confident that (within statistical limits) any difference observed between conditions was due to the experimental manipulation (and not due to alternative third variables). Therefore experimental studies allow the elimination of numerous confounding factors that may have been responsible for differences observed in case–control studies.

For example, if one observes in a longitudinal population study that smokers are more likely to develop heart disease than non-smokers, one would face the problem that smokers are also more likely to engage in other health-impairing behaviors that could have contributed to the development of the illness (e.g., Poortinga, 2007). If one had been able

to conduct an experimental study, in which individuals had been randomly assigned to a smoker and a non-smoker condition, one could have been much more confident that apart from smoking status there were no other differences that could have accounted for the health difference. However, even though such confounds are an important source of error in epidemiology, they are not considered in Ioannidis (2005) analysis.⁵

1.6.1. The importance of replications

Although Ioannidis (2005) recognized the importance of replications in his paper, he did not discuss how the positive predictive value of the truth of a research finding (i.e., PPV) improved when multiple studies showed significant results. Moonesinghe, Khoury, and Janssens (2007) extended Ioannidis' analysis by deriving the effects of replications on the PPV for various prior probabilities and two different levels of statistical power. Table 1 reproduces their calculation of the PPV for at least r significant results out of ten independent studies for pre-study odds of a true relationship of $R = 0.1$ and $R = 0.5$ at a power of 20% and 80% and $p < .05$. As one can see, with a prior probability of .5 and a power of 80%, the probability of at least one significant research finding being correct (i.e., the PPV) is only 0.55. With two significant replications (i.e., after 3 significant studies out of ten independent studies) the PPV increases to 0.98. However, the probability of obtaining r significant results decreases with increasing r (not shown in the table). Thus, with a prior probability of a hypothesis being correct of $R = 0.5$ and a power of 80%, one has a 60% chance that at least one out of 10 studies is significant, but only a 34% chance of at least three significant out of 10 studies. With power of 20% (and $R = 0.5$) one even has only an 11% chance of getting at least three significant results.

1.6.2. Critical issues

One serious threat to the validity of research in social – as well as in other areas of psychology – is the perennial problem that studies are underpowered (Cohen, 1962; Fraley & Vazire, 2014; Sedlmeier & Gigerenzer, 1989). Statistical power determines the ability to identify valid hypotheses. It depends on both the size of an effect and the sample size. If – relative to the size of an observed effect – the sample size is too small, one runs the risk of drawing false conclusions from the results of a study. Repeated analyses of the power of studies in social psychology and personality journals conducted over more than half a century have indicated that this research is typically underpowered (Cohen, 1962; Fraley & Vazire, 2014; Sedlmeier & Gigerenzer, 1989). The most recent analysis has again found that the average sample size in social and personality research was 104 and that the power to find a medium size effect ($r = .20$; $d = .41$) was approximately 50% (Fraley & Vazire, 2014).

As the analysis of Moonesinghe et al. (2007) shows, the power problem can be alleviated to some extent with multiple replications. However, the problem with underpowered replications is that there is always a good chance that a new study will not reach an acceptable level of significance. The temptation then is to disregard that study and to conduct another one, in the hope that it will work out. This procedure is not only highly problematic (i.e., publication bias), but also unnecessary. Instead, one should combine the data from the different studies and compute a meta-analytical index to pool the individual effect sizes into a single estimate (Braver, Thoemmes, & Rosenthal, 2014). With the procedure they refer to as “continuously cumulating meta-analysis”, Braver et al. (2014) recommend that researchers perform meta-analytic calculations in a continuous fashion after

⁵ Even genetic epidemiology has a third variable problem. The aim at elucidating causal relationships can be frustrated by the fact that stretches of genome tend to be inherited as a unit (i.e., linkage equilibrium). Thus, the association of one genetic variant with a disease, however strong, may be non-causal, because the association is due to some neighboring genetic material in linkage disequilibrium with the one under study (Attia, Ioannidis, Thakkinian, et al., 2009).

Table 1

Post-study probability of a true relationship (PPV) when at least N studies (out of 10) have statistically significant results as a function of pre-study probability of a true relationship of $R = 0.1$ and 0.5 , power of 20% and 80% ($\alpha = .05$).

Adapted from Moonesinghe et al. (2007, Table 2, p. 0220).

Power = 20%	R = 0.1	R = 0.5	Power = 80%	R = 0.1	R = 0.5
N of significant results	PPV	PPV	N of significant results	PPV	PPV
1	0.1820	0.5266	1	0.1995	0.5548
2	0.4202	0.7837	2	0.5372	0.8530
3	0.7369	0.9333	3	0.8968	0.9775
4	0.9216	0.9833	4	0.9898	0.9999
5	0.9810	0.9961	5	0.9994	1.0
6	0.9957	0.9991	6	1.0	1.0
7	0.9991	0.9998	7	1.0	1.0
8	0.9998	1.0	8	1.0	1.0
9	1.0	1.0	9	1.0	1.0
10	1.0	1.0	10	1.0	1.0

completing each new replication attempt. Because it is based on more data than each individual study, this pooled estimate will be much more trustworthy. Therefore, if an effect is real, the pooled effect over two (or more) studies is highly likely to be significant, even if one of the replication attempts did not reach statistical significance. Even in the highly powered Schweinsberg et al. (2016), replications by individual research groups sometimes failed to reach statistical significance, whereas effects proved reliable when pooled in the meta-analysis. In contrast, if a pooled effect does not reach an acceptable level of significance, chances are that the effect is not real (i.e., a false positive) and that selectively only publishing the significant result is misleading.

Other concerns regarding the validity of social psychological research were raised in articles criticizing the way social (and other) psychologists conducted their studies and analyzed their data. Simmons, Nelson, and Simonsohn (2011) reported simulations that showed that a number of “questionable research practices” (e.g., stopping data collection on the basis of interim data analysis; dropping experimental conditions from your published reports) could result in an erroneous rejection of the null hypothesis (i.e., false-positive results).

Because the extent to which questionable research practices distort the knowledge gained from psychological research depends on the frequency of their use in research, the survey results of John, Loewenstein, and Prelec (2012) were particularly worrisome. These authors found that in their sample of psychologists 15.6% reported “stopping data collection earlier than planned because one found the result that one had been looking for” or 27.7% failed to report all of a study’s conditions and 38.2% reported “deciding whether to exclude data after looking at the impact of doing so on the results”. These results seemed to suggest that a substantial proportion of research findings were based on questionable research practices.

On closer look, however, their findings do not allow inferences about the prevalence of these practices in current research. To estimate the prevalence of these questionable research practices in the population of psychological researchers one would need to know the frequency with which researchers engage in these practices in a given period of time (e.g., last year). Instead, respondents were asked whether they had ever engaged in a given practice. Thus, when 15.5% of respondents reported that they stopped data collection earlier than planned because the results were already significant, we have no idea *when* they did this, whether they did it once or more often, and whether they were active contributors to the research literature.⁶

⁶ This last point is particularly important, because a very small proportion of scientists accounts for most of the output in psychology. As Simonson (2002) demonstrated, the top 10% of authors in terms of output account for 37% to 47% of all publications in psychology.

Although one can doubt the claims of John et al. (2012) about the frequency with which social scientists engage in questionable research practices, their survey suggests that social scientists do use such practices at least some of the time. Since their use substantially increases the likelihood of false positive results, these reports stimulated a methodological discussion that was good for the field. It has raised the awareness that these procedures were problematic and should not be used. It also stimulated procedural changes that make it easier to detect the use of some of the questionable research practices (e.g., dropping conditions, failing to report all dependent measures) and therefore discourages researchers from using them.

The most important change is that many journals (e.g., PLOS One) now require that all data of published research (as well as all research materials) have to be stored and made available to fellow researchers on request. This requirement, which should be adopted by all journals, should discourage those questionable research practices that are based on improper statistical analysis or partial reporting of findings. Even greater transparency is achieved by journals that publish studies that have been preregistered. For example a new journal of “*Comprehensive Results in Social Psychology*” has research plans reviewed, even before the research is being conducted. If the plan for research is accepted as being methodologically sound and theoretically important, authors are guaranteed publication of the manuscript irrespective of the outcome of data analysis. In this way, the journal will also publish null findings, if the research question investigated seemed important and the research design was found to be sound.

2. Why mass replication cannot provide the answer

At this point a reader may wonder about the use of speculating about the validity of social psychological research, when mass replication attempts such as the ManyLabs project (Klein, Ratliff, Vianello, et al., 2014a), the Schweinsberg et al. (2016) or the Open Science Collaboration (2015) seem to provide a definite answer. Unfortunately, this expectation may be overoptimistic. As the instigators of the Open Science Collaboration (2012) themselves caution in their mission statement, “a failure to replicate an effect does not conclusively indicate that the original effect was false. An effect may also fail to replicate because of insufficient power, problems with the design of the replication study, or limiting conditions whether known or unknown. (Open Science Collaboration, 2012, p. 658).”

2.1. The replication projects

At the time of writing this article, results of four mass replication projects were available: The ManyLabs project (Klein et al., 2014a), a set of preregistered replications, published in a special issue of *Social Psychology* edited by Nosek and Lakens (2014), the Open Science Collaboration (2015) and the Schweinsberg et al. (2016). In two of these projects, a group of researchers replicated a sample of studies, with each group replicating all studies. The effects were then synthesized into an overall effect for each study (Klein et al., 2014a; Schweinsberg et al., 2016). In the other two projects a number of research groups each replicated one study, selected from a sample of studies (Nosek & Lakens, 2014; Open Science Collaboration, 2015).

2.1.1. Samples

To provide information on the replicability of research in social psychology, the studies selected for replication would have to be a representative sample of research in social psychology. This is not the case for three of these projects. The selection criterion for the ManyLabs Replication Project was that studies could be administered quickly and featured simple designs. The ManyLabs team tested the replicability of 13 studies across 36 independent samples, totally 6344 participants. The selection criterion for studies in the Nosek and Lakens (2014) volume is not known. The studies were selected by potential contributors to

the volume, who then submitted their plans to the editors in a preregistration procedure. The volume contains highly powered replications of the 12 studies that were approved by the editors. In the Pipeline Project 25 research groups conducted high-powered replications of ten moral judgment effects, which the last author of the project and his collaborators had “in the pipeline” in 2014. This is the only project, where research groups were explicitly selected for their expertise and their access to appropriate subject populations.

The only project that attempted to select a representative sample of studies was the [Open Science Collaboration \(2015\)](#). The Open Science Collaboration sampled studies from three major psychology journals (*Journal of Personality and Social Psychology*, *Psychological Science*, and *Journal of Experimental Psychology: Learning, Memory, and Cognition*) for the year 2008. Whereas the first replication teams could choose from the first 20 articles from each journal published in the first 2008 issue, later teams had less freedom of choice. Finally, 100 studies published in these journals were subjected to highly powered replications. This sample of studies is likely to be representative of research published in three major psychology journals. It is less clear, whether all the replication teams were experienced researchers with expertise of the particular area of study.

2.1.2. Results

One difference between the findings of these four projects is that replicability of studies was higher in the two projects where several labs replicated the same set of studies ([Klein et al., 2014a](#); [Schweinsberg et al., 2016](#)) than in the projects, where each study was replicated only by one research group ([Nosek & Lakens, 2014](#); [Open Science Collaboration, 2015](#)). This could be due to the fact that a potential lack of experience of individual research teams has less impact on the outcome of replications that are conducted by multiple (rather than individual) research groups. In the ManyLabs Replication Project 11 out of 13 effects were replicated. Similarly, eight out of the 10 moral judgment studies were successfully replicated in the Pipeline Project. In contrast, nine of the 12 preregistered replication attempts in the special issue edited by [Nosek and Lakens \(2014\)](#) were unsuccessful.⁷ Finally, in the collaborative research project, only 36% of the effects could be replicated and social psychological studies were even less replicable than studies from cognitive psychology. Whereas between 48% (JEPLMC cognitive) and 53% (PSCI cognitive) studies were replicated in cognitive psychology, the replication percentage for social psychology was 23% (JPSP) and 29% (PSCI social). However, 68% of studies could be replicated, when the original and the replication effects were meta-analytically combined. This latter percentage is very similar to that of the two multi-lab studies.

This near doubling of replication success after combining original and replication effects is puzzling. Because these replications were already highly powered, the increase is unlikely to be due to the greater power of a meta-analytic synthesis. The two most likely explanations are quality problems with the replications or publication bias in the original studies or. An evaluation of the quality of the replications is beyond the scope of this review and should be left to the original authors of the replicated studies. However, the fact that all replications were exact rather than conceptual replications of the original studies is likely

to account to some extent for the lower replication rate of social psychological studies ([Stroebe & Strack, 2014](#)). There is no evidence either to support or to reject the second explanation.

2.1.3. Discussion

As [Stanley and Spence \(2014\)](#) demonstrated in a series of simulations, even under ideal conditions findings can differ widely merely as a result of measurement error. And strikingly, even though all replications in the ManyLabs Project were conducted via computer according to as standardized protocol, lab specific methods factors accounted for 40% of the total variability of the studies on average ([McShane & Böckenholt, 2014](#)). Because conditions are rarely ideal in the real world, replications often fail because they did not succeed in faithfully reproducing the conditions of the original study. For example, when the ManyLabs Project did not replicate the finding that priming participants with the American flag increases conservatism in US participants ([Carter, Ferguson, & Hassin, 2011](#)), [Ferguson, Carter, and Hassin \(2014\)](#) pointed at methodological differences between the two studies as a potential reason for this discrepancy. Whereas their study had been a stand-alone experiment, “participants in the ManyLabs study engaged in many experiments, of which some contained direct references to the US” (p. 301). This might have interfered with the effectiveness of the experimental manipulation (i.e., the flag priming). Or when [Johnson, Cheung, and Donellan \(2014\)](#), who contributed to the special issue of [Nosek and Lakens \(2014\)](#), failed to replicate [Schnall, Benton, and Harvey's \(2008\)](#) effect of cleanliness on moral judgment, [Schnall \(2014\)](#) argued that the failure could have been due to a ceiling effect in the replication study.

All four projects relied on exact replications, often using the material used in the original studies. However, as I argued earlier ([Stroebe & Strack, 2014](#)), even if an experimental manipulation exactly replicates the one used in the original study, it may not reflect the same theoretical variable. The failed replication of the flag study provides an example for such a discrepancy. As [Ferguson et al. \(2014\)](#) argued, “national flags, ex hypothesis, activate knowledge related to one's nation that is shaped by the prevailing political atmosphere, which is hardly inert. The original experiment was run in 2009 – shortly after the first African-American was sworn in as president of the US – whereas the ManyLabs research was run four years later, in the 5th year of his presidency. Although many effects should remain stable over time (e.g., numerical anchoring effects), stimuli that represent time-sensitive knowledge and events (e.g., political and national symbols) should be expected to change over time, making “direct” replication difficult” (p. 301). [Klein, Ratliff, Vianello, et al. \(2014b\)](#) agreed with these arguments. In their concluding comment, they stated: “Direct replication does not guarantee that the same effect was tested”. Although direct replications attempt to duplicate the conditions and procedure of the original experiment, “attempting to do so does not guarantee success. Every replication is different in innumerable ways from the original” ([Klein et al., 2014b, p. 307](#)).

2.1.4. Conclusions

Although the picture that emerges from these studies about the replicability of psychological research is less than clear, with 68% of studies being replicable, even the most optimistic estimate is not completely reassuring. It would therefore have been useful if the projects had identified predictors of replication failure. Here the information is disappointing. According to the [Open Science Collaboration \(2015\)](#), surprisingness of the original effect and the challenge of conducting the replication were related to replication failure, whereas the effect size of the original study was associated with replication success. Unless researchers decide to spend the remainder of their professional lives conducting uncomplicated studies that test unsurprising hypotheses, the only useful lesson they can draw from this project is to increase the sample size of their studies, an advice that might sound familiar to readers of [Cohen \(1962\)](#). The disappointing

⁷ For some reason, one study was replicated twice. Whereas [Gibson, Losee, and Vitiello \(2014\)](#) succeeded in replicating the stereotype susceptibility effect of [Shih, Pittinsky, and Ambady \(1999\)](#), a second replication by [Moon and Roeder \(2014\)](#) commissioned later failed to replicate the effect. In a response to these two replications, [Shih and Pittinsky \(2014\)](#) suggested various reasons for this discrepancy, which were then denied by Moon and Roeder in their response to the comments of [Shih and Pittinsky \(2014\)](#). For further reactions to this particular project read Simone Schnall's description of her experience. <http://www.spsblog.org/simone-schnall-on-her-experience-with-a-registered-replication-project/>. This all confirms the warning in the mission statement of the [Open Science Collaboration \(2012, p. 658\)](#) cited earlier that “a failure to replicate an effect does not conclusively indicate that the original effect was false”. Unfortunately, such finer points are typically not understood by the wider public.

replication rate of single studies does not mean, however, that our journals and textbooks are full of false findings. Good journals typically do not publish single studies but series of studies testing the same hypothesis. Similarly, good textbooks do not rely on single studies to draw conclusions about the validity of theories, but review research programs and meta-analytic research syntheses. And as we have seen in our discussion of the analysis of Moonesinghe et al. (2007), a few successful replications substantially increase the probability that the relationship that has been observed is a true relationship.

3. What we failed to learn from our previous crisis

It is interesting to note that an earlier crisis in social psychology that erupted at the end of the 1970's was also partly fueled by concerns over non-replicability of social psychological findings (e.g., Gergen, 1973; Orne, 1962; Rosenthal & Fode, 1963). But unlike the present crisis, the discussion went beyond methodological solutions and challenged the epistemological status of social psychology. In particular, an article by Gergen (1973) aroused a great deal of hostility in the field, even though his critique would be considered more acceptable today. I will briefly review one of his main points, because it suggests criteria for the appropriateness of exact versus conceptual replications.

3.1. Social psychology as history

In his highly cited article Gergen (1973) rejected the natural science approach to social psychology and the ambition to establish general laws of human behavior. He argued that “social psychology is primarily a historical inquiry” (p. 309). “The marked success of natural sciences in establishing general principles can importantly be attributed to the general stability of events in the world of nature” (p. 309). In contrast social psychology “deals with facts that are largely nonrepeatable and which fluctuate markedly over time” (p. 310). However, Gergen (1973) admitted that social phenomena “may vary considerably in the extent to which they are subject to historical change” and he suggested that we “must think, then, in terms of a continuum of historical durability, with phenomena highly susceptible to historical influence at one extreme and the more stable processes at the other” (p. 318).

Gergen's argument has important implications for decisions about the appropriateness of conceptual compared to exact replication. The more a phenomenon is susceptible to historical change, the more conceptual replication rather than exact replication becomes appropriate (Stroebe & Strack, 2014). But whereas Gergen argued that phenomena “closely tied to physiological givens” (p. 318) would be least susceptible to historical change, I would suggest that the more a social psychological phenomenon is exclusively dependent on basic psychological processes, the more it is likely to replicate with exact replications. In contrast, social psychological phenomena that depend partly on social learning or cultural norms will only be reproducible by exact replication, if the context of learning or the cultural norms has remained unchanged (and the participants are drawn from the same or a comparable population).

Let me illustrate this with another example from the ManyLabs project (Klein et al., 2014a), where the largest effect size resulted from replications of the anchoring effect (Jacowitz & Kahneman, 1995). With this task, participants had to estimate the size (e.g., of the population of Chicago) or a geographical distance (e.g., from San Francisco to New York) after receiving a number that is clearly too large or too small. They first had to decide whether the estimate given is too large or too small. After that, they have to estimate the actual size or distance. The anchoring effect predicts that an initial number influences (anchors) later estimates.

According to the selective accessibility model developed by Strack and Mussweiler (1997), participants solve the comparative task by testing the possibility that the target possesses the anchor value and by trying to construct a mental model that includes information that is

maximally consistent with the anchor value. This process is dependent on a memory search. The fact that different participants might be more or less knowledgeable about the task will influence their estimates, but, unless they happen to know the answer, it will have no influence on the anchor effect. In terms of Gergen's “continuum of historical durability” this is an example of a process at the durable end. The other extreme is represented by findings such as the “flag study” of Carter et al. (2011). As Ferguson et al. (2014) argued, this is basically a historical effect, namely that at a certain time in American history, American citizens became more conservative, when primed with the American flag.

But many findings in the social psychological literature may reflect general laws, even though the experimental manipulations used are susceptible to historical change. Those are the studies, where conceptual instead of exact replication is appropriate. A classic example is the study by Aronson and Mills (1959) on the effects of severity of initiation on liking for a group (see Footnote 2). Although their severity manipulation would be unlikely to be effective these days, this would not invalidate their theoretical hypothesis. If their manipulation was replaced by some other task that was considered effortful or embarrassing by present day students, their finding might be replicated. Another example would be the manipulation of argument quality in tests of dual process models of persuasion, where argument strength is typically defined empirically for each specific subject population (e.g., Petty & Cacioppo, 1986).

Historical change can also affect the dependent measure, even in experiments that test basic psychological processes. To give one example, Strack, Martin, and Stepper (1988) tested the hypothesis that people's facial activity influences their affective responses. They had their student participants hold a pen in their mouth in ways that either inhibited or facilitated muscles typically associated with smiling. Their dependent measure consisted of ratings of the funniness of a set of cartoons that had been rated as moderately funny in a pretest. In support of the facial feedback hypothesis, Strack et al. (1988) found that the cartoons were rated least funny, when the muscle activity associated with smiling was inhibited. Whereas their manipulation of facial activity should not be affected by historical change, it is unlikely that the cartoons, which German students found moderately funny in the late 80ies, would still be considered funny today. However, their findings should be replicable using a new set of cartoons that was considered moderately funny by students assessed today (see also, Fabrigar & Wegener, 2016—in this issue).

I would therefore disagree with Gergen (1973) conclusion that “the continued attempt to build general laws of social behavior seems misdirected, and the associated belief that knowledge of social interaction can be accumulated in a manner similar to the natural sciences appears unjustified” (p. 316). Since the theoretical assumption tested by Aronson and Mills implied a specific effect of a severe initiation, the fact that half a century ago, severity of an initiation would have to be manipulated differently from today does not invalidate the general conclusion. Similarly, the fact that different cartoons are considered funny today than three decades ago does not invalidate the general finding of the study of Strack et al. (1988).

3.2. The development of meta-analysis to synthesize research findings

These concerns about the apparent lack of cumulative knowledge in social psychology were the reasons for the development of meta-analytic methods that allow the statistical integration of the results of independent studies (e.g., Borenstein, Hedges, Higgins, & Rothstein, 2009; Johnson & Eagly, 2014). The well-known advantage of meta-analyses over the traditional manner of summarizing studies through box scores is that even if results seem to be inconsistent with some replications producing non-significant findings, the overall effect might still be statistically significant. A further advantage is that meta-analyses are conducted on existing data and that they can be used to evaluate both the solidity of social psychological effects as well as the empirical support for

social psychological theories (Chan & Arvey, 2012). Thus, meta-analyses offer a more economical and methodologically advanced alternative to mass replication (see also, Fabrigar & Wegener, 2016—in this issue).

One serious disadvantage of meta-analysis is that conclusions can be affected by publication bias that is the selective reporting of significant studies. However several methods have been developed that help to detect publication bias and to estimate effects of potential bias on the outcome of a meta-analysis (Johnson & Eagly, 2015). The most direct strategy for the detection of publication bias is the comparison of effects in published research with those in research that has been informally or not published (e.g., dissertations, conference presentations, unpublished manuscripts, unpublished data sets). Another early method was Rosenthal's (1991) fail-safe index that estimates the number of missing samples with no effect that would be necessary to erase the significance of the effect that has been observed.

In the meantime, other strategies have been added that typically use funnel plots to detect publication bias (e.g., Duval & Tweedie, 2000; Egger, Smith, Schneider, & Minder, 1997). Funnel plots are plots of the effect sizes of trials against sample size. Since the precision in estimating a treatment effect is likely to increase with increasing sample size, results from smaller studies will scatter more widely than results from larger studies. However, if there is a publication bias in the sense that non significant findings are not reported, “a funnel plot should reveal few entries in the smaller effect size portion of the graph for smaller sample sizes” (Johnson & Eagly, 2014). Rather than relying on visual inspection, one can use quantitative methods to estimate the extent to which publication bias is present in a distribution of effect sizes (e.g., Egger et al., 1997). The “trim-and-fill technique developed by Duval and Tweedie (2000) even allows to estimate the extent to which filling in the (presumably) missing values changes the outcome.

It is intriguing, that meta-analyses present a picture of the replicability of social psychological research that differs considerably from that suggested by mass replication. There are dozens of meta-analyses summarizing social psychological research being published each year demonstrating the validity of important findings in social psychology. To give an example, I shall briefly review meta-analyses of the effect of intergroup contact on prejudice. In his classic volume on the “Nature of Prejudice”, Allport (1954) suggested the contact theory that interactions between members of different groups would improve attitudes towards the other group, at least under certain conditions, namely equal status, common goals, no intergroup competition and authority sanction. In the meantime, a meta-analysis of 713 independent samples from 515 studies demonstrated that intergroup contact does indeed reduce prejudice and that the effect is small-to-medium size (Pettigrew & Tropp, 2006). The authors also showed that their findings are unlikely to have resulted from either participant selection or publication bias. According to Rosenthal's (1991) fail-safe index, it would require 1200 missing samples that average no effect to erase the significance of the effect of intergroup contact on prejudice. A comparison of published and unpublished sources also found no significant difference in the size of the observed contact effects. The relationship between effect size and sample size was practically zero. Finally, a scatter diagram using the two variables resulted in a near symmetrical funnel. Although Allport's original conditions for optimal contact facilitated the effect, they were not necessary conditions for the effect to occur (Pettigrew, Tropp, Wagner, & Christ, 2011). This meta-analysis further demonstrated that contact is not only beneficial for contact with ethnic outgroups, but also for contact with other discriminated outgroups such as homosexuals, the disabled and the mentally ill. Finally, a meta-analysis testing mediators of the contact effect found that it reduced prejudice by enhancing knowledge about the outgroup, by reducing anxiety about intergroup contact and by increasing empathy and perspective taking (Pettigrew & Tropp, 2008).

Because much of this research is based on correlational findings, it is important that a further meta-analysis demonstrated the effectiveness of contact-based interventions on the reduction of prejudice (Lemmer

& Wagner, 2015). They included programs realizing direct or indirect contact and measuring its effect either shortly after the end of the intervention ($N = 11,371$) or one month later ($N = 1650$). The results showed that these programs did not only improve attitudes towards individuals but also towards the outgroup as a whole.

Meta-analyses can also show that there is no reliable support for some effect for which empirical support has been conflicting. For example, Nieuwenstein, Wierenga, Morey, et al. (2015) recently published a meta-analysis of studies testing the hypothesis derived from the unconscious thought theory of Dijksterhuis and Nordgren (2006) which states that complex decisions are best made without careful and elaborate deliberation. Since the evidence on this phenomenon was contradictory, Nieuwenstein et al. (2015) first conducted a large-scale replication study. When this failed, they decided to do a meta-analysis based on 61 studies conducted by different researchers. This meta-analysis resulted in a small but significant effect, which became non-significant, after correction for publication bias. From their perspective “previous reports of significant results were confined to studies that were relatively unreliable due to the use of small samples of participants” (Nieuwenstein et al., 2015, p. 9). The fact that they felt the need to conduct a meta-analysis even after a large-scale replication failed to show an effect, supports my assertion that results of a meta-analysis provide more valid information than a large scale replication about the existence of an effect or the validity of a theory. If they had decided to first conduct a meta-analysis, they might have considered it unnecessary to conduct a time intensive replication. This would have been a more effective strategy in terms of resources including time.

4. Implications for research in social psychology

“A good theory is a filter to separate likely hypotheses from unlikely ones” (Diekmann, 2011). Because theory-guided research is a means to increase the a priori probability of a hypothesis being valid, it reduces the proportion of published research findings that are false. We should therefore aim at conducting theory-guided research rather than demonstrating “sexy” counterintuitive phenomena that are derived from everyday observations and lack theoretical foundation (e.g., Schaller, 2016—in this issue). However, the fact that a novel hypothesis has been derived from a well-corroborated theory should not influence our methodological decision about how to test that hypothesis. As discussed earlier, the prior probability of any novel hypothesis is unknown and unknowable. Even with a sequence of theory-testing studies conducted in the context of a research project and based on a successfully tested theory, we cannot be confident that the pre-study probability of any novel hypothesis derived from that theory is above chance. Thus, novel findings should not be published unless they have been replicated at least once, but preferably several times. This raises the question whether such replications need to be exact or whether conceptual replication would suffice as well.

Moonesinghe et al. (2007) argued that any true replication should be an exact replication, “a precise process where the exact same finding is reexamined in the same way”. However, conceptual replications are often more informative than exact replications, at least in studies that are testing theoretical predictions (Stroebe & Strack, 2014). Because conceptual replications operationalize independent and/or dependent variables in a different way, successful conceptual replications increase our trust in the predictive validity of our theory. They have the disadvantage that if they replicate a finding, they do not rule out the possibility that the original study was a false positive. It is therefore important to replicate a novel experimental finding at least once with an exact replication study of sufficient power. Once an effect has been replicated successfully in an exact replication study, the question whether a study was a good test of the underlying theory becomes more important than the question whether the original manipulation can reliably produce a given effect. And this first question is best examined with a conceptual replication.

Such replications are not always conducted. For example, the classic experiment by Festinger and Carlsmith (1959) tested a novel hypothesis. Their finding was not only counterintuitive, it was also inconsistent with the then dominant learning theory orientation. Moreover, because only one of their four dependent variables showed a statistically significant effect, it was really doubtful whether a replication of the study would have resulted in the same findings. Therefore, an editor would have been well-advised to require a replication before accepting the paper for publication. In defense of the editor, who accepted the manuscript, one has to remember that Festinger (1957) had earlier published his monograph “A theory of cognitive dissonance” that presented an elaborate theory and contained multiple successful empirical tests of that theory. One could therefore argue that the prior probability of the Festinger and Carlsmith hypothesis being correct was not all that low. And yet, this was a novel prediction and a replication of the finding should have been requested to demonstrate that the experimental manipulation was reliable in producing this effect. This argument is even more valid for studies based on ad hoc (mini-) theorizing that was specifically developed for a given set of studies.

Even today many findings are published that are either based on novel derivations from existing theories or even untested ad hoc theories with no attempt made to establish the reliability of the experimental procedures that are being used. For example, Vess (2012) recently tested a novel prediction that was based on an integration of research on attachment anxiety and studies linking physical temperature to perceptions of intimacy (e.g., Ijzerman & Semin, 2010). Because anxiously attached individuals seek proximity under distress (Mikulincer, Orbach, & Iavnieli, 1998), Vess (2012) expected “that attachment anxiety would positively predict preference for warmth following a threat” (p. 472). To test this prediction, he manipulated attachment concerns by having participants either reflect on a past romantic breakup or on an ordinary event (Experiment 1). When participants were then asked to rate the desirability of warm-temperature refreshment, those with high (but not low) levels of anxious attachment and high (but not low) induced attachment concerns reported a preference for warm foods.

Instead of replicating these findings in a second study, to show that this experimental manipulation is reliable in producing this effect, Vess (2012) tested another prediction derived from attachment theory in Experiment 2, namely that compared to semantically activating the concept of coldness activating the concept of warmth would increase the relationship satisfaction of anxiously attached people. Although both studies are loosely derived from attachment theory, it is important to note that the second study does not constitute a conceptual replication of the first study. Both these hypotheses are plausible, but as LeBel and Campbell (2013) argued, before one tries to explain a phenomenon one should make sure that the phenomenon is reproducible. Given that neither effect has been replicated even once, the fact that LeBel and Campbell (2013) failed in two exact replications of Vess (2012) Experiment 1 to reproduce his effects is disturbing.

To present one more example, Vohs, Redden, and Rahinel (2013) recently published three studies testing the novel and interesting prediction that disorderly environments lead people towards breaking with convention, whereas orderly environments trigger conventional behavior. Experiment 1 showed that participants in an orderly room chose healthier snacks and donated more money to charity than did individuals who were sitting in a disorderly room. In Experiment 2, participants sitting in an disorderly room were more creative at the task of finding new uses for a ping-pong ball. Finally in Experiment 3, participants in an orderly room preferred a snack option labeled as “classic”, whereas those in a disorderly room were more likely to choose an option labeled as “new”. Since each of these experiments assessed a different theoretical concept as dependent measure, they do not constitute conceptual replications. I am sympathetic to the authors’ theoretical reasoning here, particularly with regard to the relationship between disorder and creativity. However, in their efforts to demonstrate the

wide applicability of their theoretical hypothesis, they did not adhere to the methodological requirement to first establish the reliability of their experimental procedures. They conducted three unrelated tests of their theory without assessing beforehand the reliability of any of the experimental procedures.

When replications are difficult, as is often the case with non-experimental studies, other strategies should be used to establish the reliability or validity of a finding. Although one could again mention many examples, I will limit my discussion to one where the original researchers themselves realized that they had fallen prey to publishing a false-positive finding. They coauthored a paper with Simonsohn in which they showed that their original finding constituted a false positive (Silberzahn, Simonsohn, & Uhlmann, 2014). In the original article (also published in *Psychological Science*), Silberzahn and Uhlmann (2013) presented data that appeared to show that Germans with noble-sounding surnames such as Kaiser (emperor), König (king) or Fürst (prince) were more likely to hold managerial positions than Germans with last names that either refer to common occupations (e.g., Koch/cook; Bauer/farmer) or to no occupation at all. They used hierarchical linear modeling to control for name frequency.

When Silberzahn et al. (2014) later reanalyzed the data comparing each noble surname with a set of similarly frequent non-noble names, the effect disappeared. There is no established psychological theory from which one could derive this finding. However, Silberzahn and Uhlmann (2013) developed the ad hoc hypothesis that somebody with a noble name (e.g., König/King) will be addressed as “king” at the workplace “potentially conditioning a psychological association between Herr König and high status. This association may in turn influence the extent to which he is viewed well-suited for managerial roles” (p. 2438). This is a hypothesis that could have been tested experimentally and an editor would have been well advised to request such additional evidence before accepting the article.

But the case of the notorious “professor priming” studies (Dijksterhuis & van Knippenberg, 1998) shows that publications of novel and counterintuitive findings can be challenged, even if researchers report both exact and conceptual replications of their findings in their primary publication. When Dijksterhuis and van Knippenberg (1998) published the counterintuitive effects of primes on intellectual performance, they reported a series of four related studies that all demonstrated that effect. In the first study they primed their participants either with the stereotype of professor or of secretaries or did not prime them. They found that participants primed with the professor stereotype outperformed those in the secretaries-prime or no prime condition on multiple-choice questions from trivial pursuit. Whereas they did not pretest their assumption that intelligence and knowledgeability are really part of the stereotype of professors held at the University of Nijmegen, they pretested that soccer hooligans (used as prime in Experiment 3) were considered stupid. Consistent with their hypotheses, they found that participants primed with the stereotype of soccer hooligans performed worse than participants in a “no prime” condition (i.e., a conceptual replication). In Experiment 4 they directly tested the assumption underlying their experiments that priming participants with the stereotype of professor or soccer hooligan increases the cognitive accessibility of the trait “intelligent” or “stupid” and it is these traits that influence performance. They primed participants either with “professor” or “soccer hooligan” (or “intelligent” or “stupid”) and found a significant effect of direction of prime. Participants primed with “professor” or “intelligent” outperformed participants primed with “soccer hooligan” or “stupid”. Since they also replicated the professor priming in their second experiment, their paper contains three replications of the professor priming effect and two replications of the hooligan priming effect.

Their report would have been even more conclusive, if they had assessed whether the professor prime increased the cognitive accessibility of the cognitive representation of the trait “intelligent” (e.g., with a lexical decision task) and demonstrated that these changes

in cognitive accessibility were responsible for the changes in performance on the trivial pursuit. The problem with this approach is that it carries the risk that having “intelligent” as a target word in the lexical decision task might act as a prime and thus alleviate the differences in the cognitive accessibility of this trait between experimental and control group. But even without this additional study, the evidence they present in their paper is very convincing.

It is therefore ironic that Shanks et al. (2013) conducted nine exact replication studies that failed to yield statistically significant effects on performance. The publication by Shanks et al. (2013) is also a good example that publishing exact replications of a finding that has already been replicated several times does not result in a great deal of new information. If such a study replicates the original finding, nothing new has been learnt. If it fails to replicate, the reason for the failure are unclear. Therefore, it would have been wise after the second failed replication to explore whether the manipulations and dependent measures used in the original study were reflecting the same constructs when used in a different country with a different subject population.

There is now a multi-lab replication attempt planned to find out whether the professor priming effect is a “true” finding.⁸ Whatever the outcome of this exercise, it seems to me that the Dijksterhuis and van Knippenberg (1998) study is a good example of how a novel theoretical prediction should be tested. One should first establish the reliability of the experimental procedure in producing the predicted effect; then one should conceptually replicate the effect and finally (or at the same time) test the validity of the theoretical assumptions with regard to mediating processes.

5. Concluding remarks

This article critically reviewed the evidence that has been advanced to support the claim that social psychology is in a replicability crisis and that most of our published findings may be false. The claim has been based on three types of evidence, namely Ioannidis (2005) Bayesian analysis, the chronically low statistical power of social psychological studies and the findings of mass replication studies. As I argued, all of this evidence can be questioned. I do not want to reiterate my critique here, but limit myself to one crucial shortcoming that is common to all of the evidence, namely their focus on single unrelated studies.

As Moonesinghe et al. (2007) demonstrated in extending Ioannidis' analysis, the probability of research findings being valid increases substantially with every successful replication. The analysis by Moonesinghe et al. (2007) is also relevant for the power problem. Reports that the average power of social psychological research is low imply that a considerable proportion of studies conducted in social psychology may have resulted in false positive or false negative findings (Fraley & Vazire, 2014). Again this estimate applies only to single studies. If studies are part of a research program with multiple studies replicating and extending the original findings, low power of individual studies becomes less problematic, especially if the effects of multiple studies are pooled meta-analytically.

The mass attempts at replicating social psychological findings also focus on single unrelated studies. Admittedly, these replications all have high power. But in contrast to meta-analyses, which statistically synthesize the effects of a variety of studies typically testing the same set of theoretical hypotheses, mass replication focuses on the effects of theoretically unrelated studies. This difference in approach is probably one of the major reasons for the puzzling discrepancy in the findings of mass replication attempts and those of meta-analyses.

In his classic volume on “Conjectures and Refutations” Popper (1963) suggested that science develops by replacing existing theories

with theories that do not only explain all the previous findings (theory-consistent as well as inconsistent), but also generate novel predictions that cannot be derived from the “old” theory. Theories consist of a network of related hypotheses, each of which will usually have been supported by dozens of individual studies. Therefore theories are not refuted by a single inconsistent finding but by studies that support an alternative theory that has greater empirical content. In this Popperian system, the statement that the findings of only a certain percentage of studies can be replicated in a given discipline would not be considered very helpful in advancing our knowledge. Rather than trying to find out whether single effects are replicable, it would appear to be more fruitful to assess meta-analytically whether a given theoretical prediction is empirically supported.

References

- Allport, G. W. (1954). *The nature of prejudice*. Reading, MA: Addison-Wesley.
- Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2014). Causality and endogeneity: Problems and solutions. In D.B. Vay (Ed.), *The Oxford handbook of leadership and organizations* (pp. 93–117). New York: Oxford University Press.
- Aronson, E., & Mills, J. (1959). The effect of severity of initiation on liking for a group. *Journal of Abnormal and Social Psychology*, 59, 177–181.
- Attia, J., Ioannidis, J.P., Thakkinian, A., et al. (2009). How to use an article about genetic association A: Background concepts. *Journal of the American Medical Association*, 301, 74–81.
- Borenstein, M., Hedges, L.V., Higgins, J.P., & Rothstein, H.R. (2009). *Introduction to meta-analysis*. New York: John Wiley.
- Braver, S.L., Thoenes, F.J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis. *Psychological Science*, 9, 333–342.
- Carter, T.J., Ferguson, M.J., & Hassin, R.R. (2011). A single exposure to the American flag shifts support toward Republicanism up to 8 months later. *Psychological Science*, 22, 1011–1018.
- Chan, M.E., & Arvey, R.D. (2012). Meta-analysis and the development of knowledge. *Perspectives on Psychological Science*, 7, 79–92.
- Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, 65, 145–153.
- Cook, T.D., & Campbell, D.T. (1979). *Quasi-experimentation: Design & analysis issues in field settings*. Chicago: Rand McNally.
- Dawber, T.R. (1980). *The Framingham study*. Cambridge, MA: Harvard University Press.
- Diekmann, A. (2011). Are most published research findings false? *Jahrbücher für Nationalökonomie und Statistik*, 231, 628–635.
- Dijksterhuis, A., & Nordgren, L.F. (2006). A theory of unconscious thought. *Perspectives on Psychological Science*, 1, 95–106.
- Dijksterhuis, A., & van Knippenberg, A. (1998). The relation between perception and behavior or how to win a game of trivial pursuit. *Journal of Personality and Social Psychology*, 74, 865–877.
- Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56, 455–463.
- Egger, M., Smith, D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple graphical test. *British Medical Journal*, 315, 629–634.
- Fabrigar, L.L., & Wegener, D.T. (2016). Conceptualizing and evaluating the replication of research results. *Journal of Experimental Social Psychology*, 66, 68–80 (in this issue).
- Ferguson, M. J., Carter, T. J., & Hassin, R. R. (2014). Commentary on the attempt to replicate the effect of the American flag on increased Republican attitudes. *Social Psychology*, 45, 299–311.
- Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford, CA: Stanford University Press.
- Festinger, L., & Carlsmith, J.M. (1959). Cognitive consequences of forced compliance. *Journal of Abnormal and Social Psychology*, 58, 203–210.
- Fraley, R.C., & Vazire, S. (2014). The N-Pact factor: Evaluating the quality of empirical journal with respect to sample size and statistical power. *Plos One*, 9, e109019.
- Gerard, H.B., & Mathewson, G.C. (1966). The effects of severity of initiation on liking for a group: A replication. *Journal of Experimental Social Psychology*, 2, 278–287.
- Gergen, K.J. (1973). Social psychology as history. *Journal of Personality and Social Psychology*, 26, 309–320.
- Gibson, C.E., Losee, J., & Vitiello, C. (2014). A replication attempt of stereotype susceptibility (Shih, Pittinsky & Ambady, 1999): Identity salience and shifts in quantitative performance. *Social Psychology*, 45, 194–198.
- Ijzerman, H., & Semin, G.R. (2010). Temperature perceptions as a ground for social proximity. *Journal of Experimental Social Psychology*, 46, 867–873.
- Ioannidis, J.P. (2005). Why most published research findings are false. *PLoS Medicine*, 2, e124.
- Ioannidis, J.P.A., Tarone, R., & McLaughlin, J.K. (2011). The false-positive to false-negative ratio in epidemiological studies. *Epidemiology*, 22, 450–456.
- Jacowitz, K. E., & Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin*, 11, 1161–1166.
- 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. <http://dx.doi.org/10.1177/0956796111430953>.
- Johnson, B.T., & Eagly, A.H. (2014). Meta-analysis of research in social and personality psychology. In C.M. Judd, & H.T. Reis (Eds.), *Handbook of research methods in social and personality psychology* (pp. 675–707) (2nd ed.). New York: Cambridge University Press.

⁸ The reason for this lingering doubt seems to be the discrepancy in estimates of the prior probability for priming between the skeptics and the believers. If, like some of the skeptics, you believe that priming without awareness is impossible, supportive evidence is unlikely to convince you of its feasibility.

- Johnson, B. T., & Eagly, A. (2015). Meta-analysis of research in social and personality psychology. In H. T. Reis, & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 677–709) (2nd ed.). New York: Cambridge University Press.
- Johnson, D.J., Cheung, F., & Donellan, M.B. (2014). Does cleanliness influence moral judgment? A direct replication of Schnall, Benton, and Harvey (2008). *Social Psychology*, 45, 209–215.
- Kelley, H. (1992). Common-sense psychology and scientific psychology. *Annual Review of Psychology*, 43, 1–23.
- Klein, R.A., Ratliff, K.A., Vianello, M., et al. (2014a). Investigating variation in replicability: A “ManyLabs” replication project. *Social Psychology*, 45, 142–152.
- Klein, R.A., Ratliff, K.A., Vianello, M., et al. (2014b). Theory building through replication: Response to commentaries on the “ManyLabs” replication project. *Social Psychology*, 45, 307–310.
- LeBel, E.P., & Campbell, L. (2013). Heightened sensitivity to temperature cues in individuals with high anxious attachment: Real or elusive phenomenon. *Psychological Science*, 24, 2128–2130.
- Lemmer, G., & Wagner, U. (2015). Can we really reduce ethnic prejudice outside the lab? A meta-analysis of direct and indirect contact interventions. *European Journal of Social Psychology*, 45, 152–168. <http://dx.doi.org/10.1002/ejsp.2079>.
- McShane, B. B., & Böckenholt, U. (2014). You cannot step into the same river twice: When power analyses are optimistic. *Perspectives on Psychological Science*, 9, 612–625.
- Mikulincer, M., Orbach, I., & Iavnieli, D. (1998). Adult attachment style and affect regulation: Strategic variations in subjective self–other similarity. *Journal of Personality and Social Psychology*, 75, 265–280.
- Moon, A., & Roeder, S.S. (2014). A secondary replication attempt of stereotype susceptibility. *Social Psychology*, 45, 199–201.
- Moonesinghe, R., Khoury, M.J., & Janssens, C.J.W. (2007). Most published research findings are false—But a little replication goes a long way. *PLoS Medicine*, 4, 0218–0221. <http://dx.doi.org/10.1371/journal.pmed.0040028>.
- Nieuwenstein, M.R., Wierenga, T., Morey, R., et al. (2015). On making the right choice: A meta-analysis and large-scale replication attempt of the unconscious thought advantage. *Judgment and Decision Making*, 10, 1–17.
- Nosek, B.A., & Lakens, D. (2014). Special issue on registered research. *Social Psychology*, 45(3).
- Open Science Collaboration (2012). An open, large-scale collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, 7, 657–660. <http://dx.doi.org/10.1177/1745691612462588>.
- Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349, 943–951.
- Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, 17, 776–783.
- Pashler, H., & Harris, C.R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*. <http://dx.doi.org/10.1371/journal.pmed.0020124>.
- Pettigrew, T. E., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology*, 90, 751–783.
- Pettigrew, T. F., & Tropp, L. R. (2008). How does intergroup contact reduce prejudice? Meta analytic test of three mediators. *European Journal of Social Psychology*, 38, 922–935.
- Pettigrew, T.F., Tropp, L.R., Wagner, U., & Christ, O. (2011). Recent advances in intergroup contact theory. *International Journal of Intercultural Relations*, 35, 271–280. <http://dx.doi.org/10.1016/j.ijintrel.2011.03.001>.
- Petty, R., & Cacioppo, J. (1986). *Communication and persuasion*. New York: Springer.
- Poortinga, W. (2007). The prevalence and clustering of four major lifestyle risk factors in an English adult population. *Preventive Medicine*, 44, 124–128.
- Popper, K.R. (1959). *The logic of scientific discovery*. London: Hutchinson.
- Popper, K.R. (1963). *Conjectures and refutations: the growth of scientific knowledge*. London: Routledge & Kegan Paul.
- Ring, K. (1967). Experimental social psychology: Sober questions about some frivolous values. *Journal of Experimental Social Psychology*, 3, 113–123.
- Rosenthal, R. (1991). *Meta-analytic procedures for social research*. Newbury Park, CA: Sage.
- Rosenthal, R., & Fode, K. L. (1963). Psychology of the scientist: Three experiments on experimenter bias. *Psychological Reports*, 12, 491–511.
- Schaller, M. (2016e). The empirical benefits of conceptual rigor: Systematic articulation of conceptual hypotheses can reduce the risk of non-replicable results (and facilitate novel discoveries too). *Journal of Experimental Social Psychology*, 66, 107–115 (in this issue).
- Schnall, S. (2014). Commentary and rejoinder of Johnson, Cheung, and Donellan (2014a). Clean data, statistical artifacts wash out replication effort. *Social Psychology*, 45, 315–320.
- Schnall, S., Benton, J., & Harvey, S. (2008). With a clean conscience: Cleanliness reduces the severity of moral judgment. *Psychological Science*, 19, 1219–1222.
- Schoenfeld, J.D., & Ioannidis, J.P. (2013). Is everything we eat associated with cancer? A systematic cookbook review. *American Journal of Clinical Nutrition*, 97, 177–184.
- Schweinsberg, M., Madan, N., Vianello, M., Sommer, S. A., Jordan, J., Tierney, W., ... Uhlmann, E. L. (2016). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology*, 66, 55–67.
- Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on power studies? *Psychological Bulletin*, 105, 309–316.
- Shanks, D. R., Newell, B. R., Lee, E. H., et al. (2013). Priming intelligent behavior: An elusive phenomenon. *PLoS One*, 8, e56515.
- Shih, M., & Pittinsky, T.L. (2014). Commentary and rejoinder on Gibson, Losee, and Vitiello (2014) and Moon and Roeder (2014). *Social Psychology*, 45, 335–338.
- Shih, M., Pittinsky, T.L., & Ambady, N. (1999). Stereotype susceptibility: Identity salience and shifts in quantitative performance. *Psychological Science*, 10, 80–83.
- Silberzahn, R., & Uhlmann, E.L. (2013). It pays to be Herr Kaiser: German with Nobel sounding surnames more often work as managers than as employees. *Psychological Science*, 24, 2437–2444.
- Silberzahn, R., Simonsohn, U., & Uhlmann, E.L. (2014). Matched-names analysis reveals no evidence of name-meaning effects: A collaborative commentary on Silberzahn and Uhlmann (2013). *Psychological Science*, 25, 1504–1505.
- 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.
- Simonton, D.K. (2002). *Great psychologists and their times*. Washington, DC: American Psychological Association.
- Slife, B.D., & Williams, R.N. (1995). *What is behind the research?* Thousand Oaks, CA: Sage.
- Stanley, D.J., & Spence, J.R. (2014). Expectations for replications: Are you realistic? *Perspectives on Psychological Science*, 9, 305–318.
- Strack, F., & Mussweiler, T. (1997). Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *Journal of Personality and Social Psychology*, 73, 437–446.
- Strack, F., Martin, L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology*, 54, 768–777.
- Stroebe, M.S., Stroebe, W., & Schut, H. (2007). Health consequences of bereavement: A review. *The Lancet*, 370, 1960–1973.
- Stroebe, W. (2013). Firearm possession and violent death: A critical review. *Aggression and Violent Behavior*, 18, 209–221.
- Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, 9, 59–71.
- Stroebe, W., Stroebe, M.S., Gergen, K., & Gergen, M. (1981). The broken heart: Reality or myth? *Omega*, 12, 87–106.
- Taubes, G. (1995). Epidemiology faces its limit. *Science*, 269, 164–169.
- Trefil, J. (2015). *The Routledge guidebook to Einstein's relativity*. New York: Routledge.
- Vess, M. (2012). Warm thoughts: Attachment anxiety and sensitivity to temperature cues. *Psychological Science*, 23, 472–474.
- Vohs, K.D., Redden, J.P., & Rahinel, R. (2013). Physical order produces healthy choices, generosity, and conventionality, whereas disorder produces creativity. *Psychological Science*, 24, 1860–1867.