What Can We Learn from Many Labs Replications?

Stroebe, Wolfgang

Published in:
Basic and Applied Social Psychology

DOI:
10.1080/01973533.2019.1577736

Publication date:
2019

Basic and Applied Social Psychology

ISSN: 0197-3533 (Print) 1532-4834 (Online) Journal homepage: https://www.tandfonline.com/loi/hbas20

What Can We Learn from Many Labs Replications?

Wolfgang Stroebe

To cite this article: Wolfgang Stroebe (2019) What Can We Learn from Many Labs Replications?, Basic and Applied Social Psychology, 41:2, 91-103, DOI: 10.1080/01973533.2019.1577736

To link to this article: https://doi.org/10.1080/01973533.2019.1577736

BASIC AND APPLIED SOCIAL PSYCHOLOGY
2019, VOL. 41, NO. 2, 91–103
https://doi.org/10.1080/01973533.2019.1577736

Routledge
Taylor & Francis Group

What Can We Learn from Many Labs Replications?

Wolfgang Stroebe

University of Groningen

ABSTRACT

Several hundred research groups attempted replications of published effects in so-called Many Labs studies involving thousands of research participants. Given this enormous investment, it seems timely to assess what has been learned and what can be learned from this type of project. My evaluation addresses four questions: First, do these replication studies inform us about the replicability of social psychological research? Second, can replications detect fraud? Third, does the failure to replicate a finding indicate that the original result was wrong? Finally, do these replications help to support or disprove any social psychological theories? Although evidence of replication failures resulted in important methodological changes, the 2015 Open Science Collaboration findings sufficed to make the point. To assess the state of social psychology, we have to evaluate theories rather than randomly selected research findings.

What can we learn from many labs replications?

For the last few years, social psychology has been in a crisis of confidence, brought about by a loss of trust in the reproducibility of research findings. The first event that gave rise to these doubts was the fraud case involving Dutch social psychologist Diederik Stapel. He was one of the young stars of European social psychology until it was discovered in 2011 that more than 50 of his publications had been based on invented data (Levelt et al., 2012; Stroebe, Postmes, & Spears, 2012). This widely publicized case, which harmed the image of social psychology in the eye of the public, was soon followed by findings of an apparently high prevalence of questionable research practices (e.g., John, Loewenstein, & Prelec, 2012; but see also Fiedler & Schwarz, 2016). At the same time, there were reports of failures to replicate high-profile priming studies; probably most publicized was Doyen, Klein, Pichon, and Cleeremans's (2012) nonreplication of the iconic study of Bargh, Chen, and Burrows (1996). Bargh and colleagues had reported that priming individuals with elderly primes slowed down their walking speed. Even though these events had already received more publicity than one would expect from an internal dispute within a scientific discipline, the press coverage became frantic after an open letter by Nobel Laureate Kahnemann, in which he warned young social psychologists not to engage in work on priming (Yong, 2012). These doubts motivated numerous research groups to engage in collaborative, large-scale, and highly powered (above .80) replication projects. These projects attempted exact and preregistered replications of the original studies with the aim to provide empirical estimates of the extent to which (social-) psychological research findings could be replicated.

The first and largest of these projects was the Open Science Collaboration (OSC; 2015), a collaborative effort that attempted exact replications of 100 studies from cognitive and social psychology. Ninety-seven of the selected studies had reported significant effects. These studies had been selected from three major psychology journals. A different research team conducted each replication attempt. Only 36% of the attempted replications were successful. Since then, results of numerous additional replication projects have been published. In a special issue of *Social Psychology* edited by Nosek and Lakens (2014), 14 preregistered replication attempts were reported, out of which eleven resulted in failure.

In the same special issue, the first of several Many Labs Studies was published (Klein et al., 2014). In Many Labs studies, several research groups attempt

CONTACT Wolfgang Stroebe wolfgang.stroebe@gmail.com Department of Social and Organizational Research, University of Groningen, Grote Kruisstraat 2/1, 9712 TS, Groningen, the Netherlands.

the replication of either one particular study or a whole set of studies. In the Many Labs 1 project, 36 research groups tried to replicate the same 13 effects with a total of 6,344 participants (Klein et al., 2014). As in the other Many Labs studies of this type, suitability for online presentation and brevity of study were major criteria for the study selection. Ten of the 13 replication attempts were successful. In the Many Labs 2 study, which is only now being published, Klein and colleagues (2018) conducted replications of 28 contemporary published findings. The total set of studies was divided into halves that could be run in 30 min. Each half was administered to half of 125 samples of more than 15,000 participants from 36 countries and territories. Fifteen (54%) of the studies showed significant effects in the same direction as the original finding. Because a major aim of both these studies was to assess the variation across samples and settings, Klein and colleagues compared the intraclass correlations (ICC) of effects across samples with that of samples across effects. As one would expect, the ICC of effects across samples was quite large in both studies, indicating that between 75% and 78% of the variance was accounted by the different effects, leaving approximately 20% of the variance to lab or sample specific aspects such as potential effects of cultural variation.[1] That the ICC of samples across effects is approximately zero in both Many Labs studies suggests that "samples would elicit larger magnitudes for some effects and smaller magnitudes for others" (Klein et al., 2014, p. 149). This assumption is supported by the finding that tests of heterogeneity were significant in 50% (Many Labs 1) and 39% (Many Labs 2) of the studies.

The Many Labs 3 study was organized by Ebersole et al. (2016) to replicate 10 known effects in 20 participant pools. Of these replication attempts, only three were successful. A second aim of that study was to detect effects of the time of the semester, when the study was being run. These time effects were weak. A modified Many Labs setup has been used by Schweinsberg et al. (2016), in a study that became known as the Pipeline Project, because 25 research groups conducted replications of up to 10 moral judgment studies, which one of the coauthors of the Schweinsberg et al. project had conducted but not yet published (i.e., in the pipeline).[2] Each of the groups attempted a direct replication of between three and 10 of these studies. Six of the 10 effects replicated robustly across all laboratories.

Camerer et al. (2018) set out to replicate 21 social science experiments that had been published in *Nature* and *Science* between 2010 and 2015. The replication attempts were conducted by five teams from the United States, Sweden, Singapore, and Austria. Again deviating from the Many Labs design, each team was responsible for only three to five replications. These researchers found significant effects in the same direction as the original study for 12 replications (57.1%). After additional data collection to increase statistical power, two more studies replicated to bring the total to 14 (67%) successful replications.

In addition to these projects, there were also seven preregistered replications published in *Perspectives on Psychological Science*, of which six were replication failures (Bouwmeester et al., 2017; Cheung et al., 2016; Eerland et al., 2016; Hagger et al., 2016; O'Donnell et al., 2018; Wagenmakers et al., 2016) and only one was successful (Alogna et al., 2014). These replication attempts used a different type of Many Labs framework with all the labs attempting to replicate one specific effect.

Given the enormous investment of time and resources in these replication attempts, the time may be ripe to assess what we have learned from this research and what can potentially be gained from these types of projects (see also Strack, 2017; Strack & Stroebe, 2018; Stroebe, 2016a). In this evaluation, I address four questions: First, to what extent do these replication studies inform us about the replicability of social psychological research? Second, can replications detect fraud and have these projects helped us to uncover fraud cases? Third, does the failure to replicate a study finding indicate that the original finding was wrong? Finally, do these replication attempts help to support or disprove any social psychological theories?

Can replication studies assess the replicability of social psychological research?

To assess the replicability of social psychological research, a replication project would have to attempt replications of a representative sample of social psychological research. Obviously, this is an impossible task, because it would have to include unpublished research and unpublished research—even if it is recent—is notoriously difficult to trace. But even if one would limit such a sample to published research, the numbers would be enormous, given that the first social psychological studies have been published in the decade before 1900 (Stroebe, 2012). Although it would be feasible to create a representative sample of research published during the last decade in journals specifically devoted to social psychology, the numbers

of studies to be replicated would still be exceedingly large.

It is therefore not surprising that none of the Many Labs studies made any attempt at sampling. The only project that made such an attempt was the OSC (2015). All effects were selected from one of three journals: *Psychological Science*, *Journal of Personality and Social Psychology*, and *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Replication teams could select from a pool of the first 20 articles from each journal, starting with the first article published in the first 2008 issue. Because most articles publish multiple studies, the last experiment in an article was selected for replication.

One can doubt whether articles from three prime journals of the discipline are representative for psychological research. The OSC claim that "the resulting open data set provides an initial estimate of the reproducibility of psychology" (pp. 4716-1) as well as their title "Estimating the Reproducibility of Psychological Science" considerably overstates their case. Even a more modest title such as "Estimating the Reproducibility of Psychological Research Published in 2008 in Three Prime Journals" might be somewhat of an overstatement. As Fiedler and Prager (2018) criticized, the instruction to select the last study in each article may have captured studies testing problematic and tangential hypotheses that were forced on researchers by reviewers.

In contrast to the attempt of the OSC (2015) to achieve some measure of representativeness, the Many Labs projects clearly stated that their collection of studies was in no way representative of social psychological research. For example, the studies in the Many Labs 1 and 2 projects were selected so that they could be easily run in a computer laboratory, and in one session. Schweinsberg et al. (2016) in their pipeline project attempted replications of studies on moral judgment that had just been conducted by one of the coauthors, certainly a very idiosyncratic selection criterion. Their replication rates are therefore at best representative for social psychological studies that are suitable for online presentation and are quick to run.

With regard to the preregistered replications published in *Perspectives on Psychological Science*, the selection criteria are often that a study has received many citations. However, this is not a particularly valid selection criterion. For example, the Pen-Study of Strack, Martin, and Stepper (1988) has been cited 703 times (Web of Science, October 1, 2018). Yet it is only of historical interest. Even a failure of replicating the pen-effects would not reduce our trust in the facial feedback theory it originally tested. Similarly questionable are the reasons for the selection of the so-called Professor Study of Dijksterhuis and van Knippenberg (1998). This study claimed to show that priming student participants with the word *professor* increased their performance on a Trivial Pursuit task. Although this is an interesting effect, there has never been a satisfactory theoretical explanation. Different from other studies that suggest that category priming can affect behavior such as walking speed (e.g., Bargh et al., 1996), the professor priming influences performance on a skill task (i.e., trivial pursuit). Furthermore, the effect has frequently been replicated, most recently in 2014 by Dijksterhuis, van Knippenberg, and Holland (mentioned in Dijksterhuis, 2018). They conducted an exact replication at the same university as the original study and found the effect for male but not for female students.

The second and perhaps even more important question is whether the failure to replicate a study finding means that that particular study finding was wrong. As I discuss this point extensively later, I only briefly touch on two problems here. First, and this is a problem specific to social psychology, operationalizations in social psychology often derive their meaning from the historical, social or cultural context (Crandall & Sherman, 2016; Stroebe & Strack, 2014). Exact replications of a study, as they were attempted in all replication projects, may fail to reflect the same theoretical variables that were manipulated or measured in the original study. As a result, the failure of an exact replication of a given study does not necessarily imply that the theoretical hypothesis originally tested in that study was refuted. Another potential problem, which is also discussed later, is the possibility that even in attempts to conduct an exact replication of a study, experimenters might introduce unintended deviations from the original procedure, which will result in a failure to replicate the original findings.

What can we conclude from these replications projects? Although they do not permit inferences about the replicability of (social-) psychological research in general, they do suggest that our methodology could profit from improvements. Social psychologists have learned to take the complaints about the insufficient statistical power of their research (e.g., Cohen, 1962; Fraley & Vazire, 2014; Sedlmeier & Gigerenzer, 1989) more seriously and have considerably increased the number of participants on which they base their studies. Most social psychology journals now require power analyses and editors discourage studies that

have less than .80 power.[3] As another positive consequence, researchers are now typically required to submit all experimental materials and all relevant data to either the journal publishing their studies or some publicly available server. This will considerably reduce questionable research practices such as failing to publish relevant but inconsistent findings. It will also facilitate the detection of fraud. Easy access to the original data of studies will also greatly facilitate meta-analyses. Thus these replication projects have resulted in changes in the way we do our research. However, the publication of the OSC (2015) would probably have sufficed to achieve this result. Social psychologists had learned their lesson and further multiple-lab studies were not really required.

Can replication studies detect fraud?

It is a widely accepted myth in psychological science that replications are useful in detecting fraud. For example, Crocker and Cooper (2011, p. 1182) stated in an editorial in a special section in *Science* on data replication and reproducibility that "scientists generally trust that fabrication will be uncovered when other scientists cannot replicate (and therefore validate) findings." Along similar lines, Roediger (2012), a former Association for Psychological Science president, commented on the Stapel case that "if others had tried to replicate his work soon after its publication, his misdeeds might have been uncovered much more quickly" (para. 6). A study of the history of science would have indicated that fraud is rarely detected through failed replications (e.g., Broad & Wade, 1982; Stroebe et al., 2012).

There are several reasons why replications do a poor job as fraud detectors. There are typically so many alternative explanations for replication failures that fraud is not even considered. Science is based on trust, and before the recent fraud cases, it would probably have not even occurred to scientists to suspect a colleague of fraud. Based on information found on the Internet, an analysis of 40 fraud cases observed that in the overwhelming number of cases, colleagues or students who became suspicious detected the fraud (Stroebe et al., 2012). In only two, and two rather unusual, cases was fraud discovered by replication failure. Even the Stapel fraud was revealed by his research students, who had become suspicious of his unusual success in empirically supporting the most daring hypotheses (Stroebe et al., 2012). With the new rule that data for published research have to be made available, it can be expected that fraud cases will increasingly be detected because of suspicious data patterns.

Another reason for replications being poor fraud detectors is that clever fraudsters, who stick closely to predictions that are plausible in the light of existing literature, have a very good chance that their research will be successfully replicated by their colleagues.[4] For example, DeCoster and Claypool (2004) published a meta-analysis of priming effects on impression formation supporting a general model of information bias. The literature was very coherent and supportive of their model. The only unexpected finding was that effect sizes of studies conducted in Europe were substantially greater than those of American studies. They attributed this to cultural differences. However, when I checked the authorship of the European studies, it turned out that the majority had been conducted by Stapel, and many of these studies later turned out to be fraudulent (Levelt et al., 2012). Thus, in inventing data, Stapel managed to get the priming effects right but overestimated the size of these effects.

Does replication failure falsify the original finding?

The German Research Association (Deutsche Forschungsgemeinschaft, 2017) published a position paper in which they made the obvious—but sometimes overlooked—point that the finding that a scientific research result is replicable or not replicable is itself a scientific finding and, as such, not final. Like all scientific findings, it is subject to scientific skepticism and further examination. More specifically, they stated that "replication failure is no general proof of falsification" (my translation). From reading some of the Many Labs replication attempts, one gets the impression that authors believe that by having a given effect replicated by more than a dozen laboratories from several continents and more than 1,000 research participants, they can somehow identify the "true effect" of a manipulation and thus arrive at conclusions that are final.[5]

This is an illusion. There are both statistical and methodological problems making it difficult to draw clear-cut conclusions from replication studies. The *statistical problem* is that there "is no single standard for evaluating replication success" (OSC, 4716-2). Because much has been written on this (e.g., Fiedler & Prager, 2018; Trafimow, 2018), I limit my discussion to the effect of regression shrinkage, because it invalidates an apparently straightforward and intuitively appealing criterion for evaluating

replications: Replications are considered successful if they show a statistically significant effect in the same direction as the original study. As Trafimow (2018) has recently pointed out, this is a poor decision criterion. Like any statistic, *p* values have a sampling distribution. Thus, if a researcher would exactly replicate the same experiment multiple times, he or she would get a distribution of *p* values. Without publication bias, regression toward the mean would sometimes lift the *p* value of a replication of a nonsignificant effect above the threshold of significance and sometimes lower the *p* value of the replication of an originally significant effect below the significance level. However, given the fact that only significant effects are being published, it is likely that some of the researchers of the original study were lucky and hit a *p* value that was at the upper tail of the distribution. As a consequence, regression effects are likely to have lowered the *p* value of the replication. "Therefore, one way to view the Open Science Collaboration finding is that it provides empirical confirmation that psychology results are not immune to statistical regression" (Trafimow, 2018, p. 3).

Regression toward the mean also invalidates the second indicator of replicability used by the OSC, namely, their evaluation of the effect size of the replication against the original effect sizes. The OSC found that of the 99 studies for which an effect size could be computed for both the original and the replication study, 82 showed a stronger effect size in the original study. Again, this difference can be attributed to regression toward the mean. Strong original effects are likely to shrink when replicated (Fiedler & Prager, 2018). However, because—unlike with *p* values—there is no (direct) pressure to publish only studies with large effect sizes, regression toward the mean can also, in cases of studies with relatively small effects, result in replications to have larger effects than the original study (Fiedler & Prager, 2018).

The *methodological problem* is that it is practically impossible to conduct exact replications of the original study. The aim of exact replications is to re-create the manipulations and measures of the original experiment as closely as possible. In trying to achieve this, replicators are confronted with two problems. The first problem is that descriptions of experimental procedures in journal publications are typically not very complete. For example, in attitude change experiments, the messages used are rarely fully described, leaving it to the replicator to create a communication in line with the often scanty description given in the original publication. Similarly, the instructions likely are only paraphrased and the dependent measures are not described in full, with only a few examples given in illustration. A partial remedy here is to get in touch with the original authors in the hopes that they did keep a complete record. With experiments done in a time before computers, such attempts are likely to be unsuccessful.

The second problem, which is a major problem for social psychological research, is that manipulations and measures often derive their meaning from the historical, social, and cultural context at a given time (Crandall & Sherman, 2016; Gergen, 1973; Stroebe & Strack, 2014). As a result, in a replication attempt of a theory-testing study conducted many years after the original study, the manipulation used in the original study may no longer represent the theoretical construct it reflected at that time. Similarly, the dependent measure might no longer assess the theoretical variable it measured originally. For this reason, many researchers have argued that we should focus on conceptual replications of theory-testing research and use manipulations and measures that are more appropriate in terms of the historical, social, and cultural context in which the experiment is replicated (e.g., Crandall & Sherman, 2016; Stroebe & Strack, 2014).

A good illustration that historical or social change can make a manipulation inappropriate is the classic study by Aronson and Mills (1959) on the effects of the severity of initiation on the liking of a group. This study showed that female students, who had to undergo a severe initiation to join a group (actually a very dull discussion), liked the group more than women, who had undergone a mild initiation. The theoretical hypothesis, derived from dissonance theory, was that undergoing a severe initiation to join what turned out to be a dull group would create more dissonance than having gone through only a mild initiation. The severity manipulation involved having to either read aloud dirty words from a text or read neutral words. I am not aware of any attempt at exactly replicating this manipulation, but I would not expect it to be successful. Whereas reading aloud dirty words in front of a male experimenter might have been an embarrassing experience for young women in 1959, this is probably no longer the case today. Thus, an exact replication of this study would be unlikely to replicate the original findings even though a conceptual replication might do so.[6]

An early conceptual replication of this study was conducted by Gerard and Mathewson (1966) to rule out some alternative explanations of the original finding. I report it here because it nicely illustrates the

principle of conceptual replications. Instead of trying to vary the theoretical variable "severity of initiation" by having female students read dirty or neutral words, the initiation consisted of experiencing either weak or strong electric shocks. Thus this conceptual replication used a totally different empirical manipulation to manipulate the theoretical variable severity of initiation. Consistent with the "suffering leads to liking hypothesis," students who had to suffer the strong shocks liked the discussion groups more than did students, who experience only mild shocks. As with all conceptual replications, it can be argued that giving electric shocks might not constitute the same type of initiation as having female students read out dirty words. It should be pointed out, however, that any manipulation in any theory-testing experiment relies on auxiliary hypotheses that can be criticized for failure to truly reflect the theoretical variable it is supposed to vary (Trafimow, 2012). This is one the reasons for the rejection of "naïve falsificationism," that is, the assumption that a single inconsistent finding can falsify a theory (Popper, 1959).

That changes in the social context might have affected findings of the OSC (2015) has been suggested by van Bavel, Mende-Siedlecki, Brady, and Reinero (2016). These researchers had the OSC studies rated in terms of contextual sensitivity, that is, the extent to which the research topic in each study was deemed by raters to be contextually sensitive. Success of replication attempt correlated negatively with contextual sensitivity ($r = -.23$), an association that remained significant even after controlling for effect size and statistical power. This could be an indication that some of the replication failures might have been due to the impact of historical change and a failure of manipulations or measures to reflect the original theoretical constructs. Although Inbar (2016) criticized the interpretation of van Bavel et al. pointing out that the association between contextual sensitivity and replicability becomes nonsignificant if one controls for discipline (i.e., cognitive vs. social psychology), this could be the result of a restriction in range: Contextual sensitivity is a problem specific to social psychology (Gergen, 1973).

In another comment on the OSC (2015), Gilbert, King, Pettigrew, and Wilson (2016) criticized that some replications used participant populations that differed from those used in the original study. This can be a problem even if an apparently identical American student population is used, because students today are likely to differ in many ways from those who were studied 20 years ago. It becomes even more problematic, when a study originally conducted with American students is replicated in another country or with nonstudents. Gilbert et al. mentioned that an original study that measured American attitudes toward African Americans (Payne, Burkley, & Stokes, 2008) was replicated with Italian students, who were unlikely to share the same stereotypes. Similarly, a study by Risen and Gilovich (2008) that asked college students to imagine being called by a professor was replicated with individuals who had never been to college. Gilbert et al. further criticized that some replication attempts used procedures that differed in substantial ways from the protocol of the original study and thus failed to exactly replicate that study. The most striking example was the replication of a study that asked Israelis to imagine the consequences of military service (Shnabel & Nadler, 2008) and asked Americans to imagine the consequences of a honeymoon. Such deviations may be necessary when a conceptual replication is considered more appropriate than an exact replication. But there is no reason why an exact replication could not have been conducted with Israeli participants.

Although the critique of Gilbert et al. (2016) is persuasive, it suffers from the weakness that they cannot demonstrate that the original findings would have been replicated had the original procedures been followed with the proper subject populations. In the meantime, there are studies that have done so, albeit not with OSC experiments. One example is the replication by Luttrell, Petty, and Xu (2017) of a failed replication attempt by the Many Labs 3 project (Ebersole et al., 2016) of a study by Cacioppo, Petty, and Morris (1983). Cacioppo et al. tested a prediction of the elaboration likelihood model (Petty & Cacioppo, 1986) about the effects of differences in Need for Cognition (NFC) on persuasion. People with high need for cognition enjoy engaging in cognitive activity and are assumed to scrutinize message contents to a greater extent than people with low need for cognition. According to the ELM, they should therefore be more influenced by the quality of the arguments presented in a communication. This hypothesis was experimentally supported by Cacioppo et al. in a study in which they manipulated argument quality and measured NFC. In contrast, an attempt at exactly replicating this study in the Many Labs 3 project (Ebersole et al. 2016) failed to find the expected interaction between NFC and argument quality on persuasion.

In their comments on the Many Labs 3 replication failure, Petty and Cacioppo (2016) pointed at several

discrepancies between the original study and the replication. Because the original communications were no longer available, Ebersole et al. (2016) had to re-create those messages. According to Petty and Cacioppo (2016), the new communications were unusually short. More important, different from the original study, participants in the Ebersole et al. (2016) study were also told that the change in examination procedure announced in the message would shortly be introduced at their own university. This instruction is known to increase processing motivation (e.g., Petty, Cacioppo, & Goldman, 1981) and thus likely to eliminate the motivational differences due to measured NFC. When Luttrell et al. (2016) replicated the study, comparing the suboptimal condition (short message, high processing motivation) of Ebersole et al. (2016) with an optimal condition (longer message, low processing motivation) supposedly used in the original study, they found a third-order interaction (Condition × Argument Quality × NFC). They replicated the NFC × Argument Quality interaction in the optimal condition (i.e., longer message, low processing motivation) but not in the suboptimal condition used by Ebersole et al. (2016).

But the story does not end on this happy note. Ebersole et al. (2017) attempted to replicate the Luttrell et al. (2016) study in nine laboratories with 1,219 participants using the Luttrell et al. materials. They failed to replicate the third-order interaction of Luttrell et al. (2016). However, they found a significant second order interaction (NFC x Argument Quality) in the optimal condition, but not in the suboptimal condition. Although they replicated the original effect of Cacioppo et al. (1983) in the optimal condition, they could not reliably establish that the effects of the two experimental manipulations differed. But since the original question was whether the Cacioppo et al. (1983) findings could be replicated, the fact that Ebersole et al. (2017) have now replicated it, might deter future replicators from pursuing this issue any further.

Another example of a failed replication—the attempt by Wagenmakers et al. (2016) to replicate the so-called Pen-Study of Strack, Martin, and Stepper (1988)—illustrates that exact replications can become inexact when experimenters try to improve on the original procedure. The Pen-Study of Strack et al. was a test of the facial feedback hypothesis. This hypothesis, originally proposed by Darwin (1872), suggests that certain facial expressions and postures exert an influence on emotional responses. Research on this theory had originally been hampered by the difficulty of manipulating facial expressions unobtrusively. The innovative contribution of Strack and colleagues (1989) was to suggest a method by which frowns and smiles could be induced without giving specific instructions about facial muscles. Holding a pen with one's teeth forces facial muscles into a smiling position, whereas holding it with one's lips inhibits smiling. The dependent measure in this study was ratings of the funniness of moderately funny cartoons. In support of the facial feedback hypothesis, participants reported more intense humor when cartoons were presented under conditions that facilitated smiling rather than inhibiting it. Although important at the time as the first strict test of the facial feedback hypothesis, the study is only of historical importance today because the facial feedback hypothesis is well supported by a multitude of psychological and neurological data: The manipulation of facial expressions has been shown to affect emotional responses as well as prefrontal activation in the amygdala (Price & Harmon-Jones, 2015). A recent study demonstrated even that the two ways of holding the pen were associated with different fMRI fluctuation in areas related to the initiation of positive emotions (Chang et al, 2014; see also Hennenlotter et al., 2009).

Given that multiple successful conceptual replications of the Pen-Study had already been published,[7] it came as a surprise that the 17 laboratories from eight countries failed to find an effect of the pen manipulation on ratings of the funniness of cartoons (Wagenmakers et al., 2016). However, on closer reading of the publication, the reason for this failure became apparent (Strack, 2016; Stroebe, 2016b). In their pursuit of "true effects," Wagenmakers and colleagues had decided to videotape participants during the experiment to make sure that they followed instructions. This would have been no problem had they used hidden cameras. But participants were told that they would be videotaped and had the camera in full sight. As Strack (2016) pointed out in his reflections on the Wagenmakers et al. (2016) study, this might have been responsible for the failure of the pen manipulation having an effect. Independently, I wrote in a blog in the Digital Newspaper of Utrecht University,

> I could image that they might have felt a bit strange sitting in a laboratory with a pen in their mouths rating cartoons. Most participants will probably have been concerned about looking rather silly. It is easily imaginable that these concerns will have drowned the subtle cues emanating from their facial muscles. (Stroebe, 2016b, para. 11)

It is surprising that none of the 17 researchers involved in the replication attempt objected to this change in the original procedure.

In the meantime, there is evidence to suggest that the cameras were indeed responsible for the replication failure. Noah, Schul, and Mayo (2018) recently published a further replication of the study, in which they implemented the pen manipulation under two conditions, one with a video camera and one without. Whereas the condition without a camera replicated the original effect of Strack et al. (1988), the insertion of a camera eliminated the effect and thus replicated the null-finding of Wagenmakers et al. (2016).

Even more recently, a further successful replication of the original effect has been published (Marsh, Rhoads, & Ryan, 2018). These authors used a classroom demonstration to replicate the original Pen-Study with 446 male and female students. Students were divided into two groups on the basis of their seating position in the classroom. Students on the right side of the room were instructed to hold the pen in their teeth, whereas students on the left side were instructed to hold it in their lips. While holding the pens in their mouths, students were shown cartoons via overhead projector and were asked to rate how funny these cartoons were. They were then asked to switch the pen to the other position and were shown a new set of cartoons they had to evaluate. Thus, the pen effect was tested in both a within-subjects and a between-subjects design. The mean difference between the cartoon ratings under the two conditions was highly significant.

The problematic replication studies discussed earlier illustrate the point made by the Deutsche Forschungsgemeinschaft (2017) that replications—even if they are as massive as the OSC or the Many Labs studies—cannot provide definitive answers. However, it is not my intention to imply that social psychological study findings cannot be effectively falsified. In fact, the Many Labs framework could be well suited for testing whether published research findings can be replicated. If several laboratories, following the same protocol, fail to replicate an earlier finding, it is probable that this finding is false (Earp & Trafimow, 2015). However, the likelihood that a finding is false depends not only on the power of the replication attempt but also on the quality of the experimental protocol. If the experimental procedure followed in the replication attempt is suboptimal, either because it does not represent an exact replication (if an exact replication is intended) or because a conceptual replication would have been more appropriate, then even the results of powerful Many Labs studies have to be doubted.

Following the same protocol can also become a weakness, at least with manipulations or measures that are culturally sensitive, if the research groups involved in that replication attempt come from different cultural contexts. It is probably out of misguided ideas about external validity (see Stroebe, Gadenne, & Nijstad, 2018) that most of these Many Labs projects have involved research groups from a variety of countries. For example, O'Donnell et al. (2018), who attempted to replicate the effect reported by Dijkterhuis and van Knippenberg (1998) that priming students with the word *professor* increased their performance on a Trivial Pursuit task, proudly reported that "the participating labs represent five continents and 19 countries" (p. 272). The problem here is not the fact that theories are tested in several countries but that an experimental manipulation or measure that has been developed in one country may not reflect the intended theoretical constructs when used in a different cultural context. The fact that the knowledge items—used as dependent measure by all research groups in the O'Donnell et al. replication—had been standardized with an undergraduate student sample of the University of California at Berkeley constitutes a methodological problem. As Dijksterhuis (2018) pointed out in his comments on the replication, "this obviously made for a less sensitive dependent variable than would have been possible with a tailor-made set of questions pilot-tested by the individual labs" (p. 296). Some other replication attempts suffer from the same problem. For example, Wagenmakers et al. (2016) used cartoons as dependent measure that had been pretested with Dutch students from the University of Amsterdam. It is hardly likely that students in other countries, who participated in this study, had the same sense of humor as these Dutch students.

Even if a Many Labs replication would unequivocally suggest that a particular effect that had been produced in the original experiment could not be reproduced, this finding would per se not be particularly informative. In theory-testing research, we are interested in manipulations and measures only to the extent that they inform us about the validity of the theory being tested. For example, the phenomenon that holding a pen between the lips rather than with one's teeth affects a person's evaluation of the funniness of cartoons is important only because it supports the facial feedback hypothesis. It has no other purpose. Because the facial feedback hypothesis has in the

meantime been supported by numerous much stronger studies (i.e., studies that used better manipulations of facial muscles and better measures of affective responses), nonreplication of the Pen Study would hardly reduce our trust in the validity of the facial feedback hypothesis. Thus, evidence that a particular effect does not exist is informative only because of its relevance for the theory that predicted that effect. And this relevance might decrease with the availability of better methods of testing that theory.

Can failure to replicate a single finding falsify a theory?

A theory consists of abstract and unobservable constructs and assumptions about the relationship between these constructs. To test the theory empirically, these unobservable theoretical constructs have to be operationalized, that is, translated into observable terms with empirical hypotheses. The assumptions that link unobservable theoretical constructs to empirical manipulations or measures are "auxiliary hypotheses" that can themselves be true or false (Gadenne, 1984; Trafimow, 2012). For example, the unobservable theoretical construct of NFC has been translated into an empirically measurable construct through the development of the NFC questionnaire (Cacioppo et al., 1983). The theoretical construct of argument quality has been operationalized through sets of arguments that have been evaluated as high or low quality in a pretest. Finally, the unobservable theoretical concept of persuasion has been translated into an empirically measurable construct through an attitude scale (Trafimow, 2012).

Unfortunately, there is always the possibility that researchers' auxiliary hypotheses are invalid (Gadenne 1984; Popper, 1959; Trafimow, 2012). For example, the NFC scale might have low reliability or validity, the piloting of argument quality might have been done with high school rather than university students, and the measure of persuasion might have been unreliable. Most important, however, due to cultural, social, or historical change, the empirical realization of a theoretical construct in the original study may have no longer reflected this theoretical construct in a replication attempt conducted decades later (Stroebe & Strack, 2014).

This is one of the reasons why there can be no crucial experiments that can determine the fate of a theory. According to Popper's (1959) methodological falsificationist position, theories are only falsified by multiple negative results and/or an alternative theory of greater empirical content.[8] Theories have to be evaluated with research *programs* that consist of whole sequences of studies that assess the validity of hypotheses derived from a given theory. The purpose of every single study in that program is to test whether a given finding is consistent or inconsistent with the theory tested. A positive outcome of a theory-testing experiment will increase our trust in that theory, whereas a negative outcome will decrease it. Thus, each outcome is a small building block that contributes to our knowledge of the phenomenon under study (see Earp & Trafimow, 2015, for a formalization of this process). Although we can never hope to prove that a theory is valid, the aim of a research program involving such a sequence of theory-testing studies is to leave us either with a great deal of trust in the validity of that theory or with the conviction that the theory is untenable, at least in its present form.

This raises the question about the purpose of these multiple attempts at mass replication of *unsystematically* selected studies. Even if a Many Labs replication would *unequivocally* suggest that a particular effect that had been produced in the original experiment could not be reproduced, this finding would merely reduce our trust in that theory (assuming that the theory was otherwise well supported) but not result in its falsification. But the knowledge is in the theories and not the effects being studied. The finding that the way one holds a pen in one's mouth influences ratings of the funniness of cartoons is per se a totally useless bit of knowledge.[9] Thus, if one wanted to evaluate the state of social psychology as a science, one had to assess the validity of social psychological theories and not of social psychological findings.

Conclusions

To assess the state of social psychological science, one has to evaluate the validity of social psychological theories, because the body of knowledge of a science is in its theories and not its effects. Exceptions are applied sciences where applications become important. For example, if medical theories make the prediction that a particular drug would be helpful in treating a disease, there will be a long sequence of drug tests to establish the effectiveness of that drug. Here multiple drug trials are essential to make sure that a drug really works. But these studies are typically not theory-testing but assessing whether a particular intervention has the expected effect. Social psychology has also multiple applications such as health, organizational, or consumer psychology, where repeated assessments of

the effectiveness of applications are important. However, in evaluating social psychological science—the aim of practically all of these replication attempts—the concern should be with the validity of our theories rather than that of unsystematically selected research findings.

According to Popper (1959), theories cannot be proven to be true. Although we can increase our trust in the validity of a theory by conducting strict tests of that theory, there will never be complete certainty that a theory is valid. Theories can also not be falsified by single experiments. They have to be evaluated by research programs that assess the validity of multiple hypotheses that can typically be derived from a theory. Therefore, the attempt to replicate single *unsystematically* selected psychological effects does not allow general statements about the validity of the theories that were originally tested with these studies. Given the additional problem that the outcomes of replication attempts are themselves scientific findings and open to skepticism, their contribution to social psychological knowledge is extremely limited. One can therefore question whether the knowledge they add justifies the immense investment in time and resources that have been expended by the various multilab studies. Although it is possible to evaluate theories by replicating studies, one would have to *systematically* select sets of studies that provide the central evidence for a given theory. An alternative procedure would be to assess the overall predictive validity of that theory through meta-analysis. However, the optimal strategy, and one that would result in the greatest gain in knowledge, would be to develop and test a better theory that has greater empirical content.

Notes

1. Because all studies were conducted at the same time, one would not expect effects due to changes in historical context. Changes in historical context should result in replication failures. The fact that the participants in these studies had access to (or familiarity with) computers should have reduced the likelihood of cultural variations.
2. I must admit that the purpose of this exercise was not totally clear to me. Making this a prerequisite of any empirical publication would be a rather cumbersome requirement.
3. For example, the author instructions of the *Journal of Experimental Social Psychology* state, "Each original empirical study with existing data should report, for its key hypothesis tests, a *sensitivity power analysis*. … This should assume an alpha significance criterion (normally .05, two-tailed), and a standard power criterion (normally 80%), and report the minimum effect size" (https://www.elsevier.com/journals/journal-of-experimental-social-psychology/0022-1031/guide-for-authors).
4. If Stapel had kept to this recipe and not become overconfident in his later research, his fraud might never have been detected.
5. Pashler and DeRuiter (2017) suggested that "all reviews of summaries produced by psychological scientists … need to explicitly and conservatively label the degree of support. The highest credibility category—call it 'Class 1'—must be reserved for findings that have been confirmed in one or more preregistered replications, where publication bias, HARking (hypothesizing after the results are known), and p-hacking can all be confidently excluded."
6. To support this argument empirically, one would have to replicate the original embarrassment manipulation and demonstrate with a manipulation check that it does not result in significant differences in embarrassment in the replication study. As Erdfelder and Ulrich (2018) argued correctly, "An exact replication can be questioned only, if .. it can be demonstrated that the experimental material at the time and place of the replication study fails to have the same psychological effect as at the time of the original study" (pp. 6–7, my translation). Although Aronson and Mills (1959) did not report a manipulation check, they assumed that their manipulation had a differential effect on embarrassment.
7. A list of the studies can be retrieved from https://www.dropbox.com/s/5ttmh4swuhwgs17/Literature.xlsx?dl=0.
8. A theory B is said to have greater empirical content than a theory A if it can (a) explain all the findings supportive of A, (b) explain the findings that are nonsupportive of A and (3) make some new predictions that cannot be derived from A.
9. This is often used by journalists to ridicule social psychology. For example, if one describes the Pen-Study without mentioning that is was conducted to test an important theoretical hypothesis, it appears a useless—even inane—exercise.

Acknowledgments

The article is based on one of the Adair Distinguished International Visitors lectures presented to the Department of Psychology of the University of Manitoba, Canada in October 2018. I thank Alice Eagly, Volker Gadenne and Bernard Nijstad for helpful comments on aspects of this article.

References

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahník, Š., Birch, S., Birt, A. R., … Zwaan, R. A. (2014). Registered replication report: Schooler and Engstler-Schooler (1990). *Perspectives on Psychological Science*, *9*(5), 556–578. doi: 10.1177/1745691614545653

Aronson, E., & Mills, J. (1959). The effect of severity of initiation on liking for a group. *Journal of Abnormal and Social Psychology*, *19*, 177–181. doi:10.1037/h0047195

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behaviour: Direct effects of trait construct and stereotype activation. *Journal of Personality and Social Psychology*, *71*(2), 230–244. doi:10.1037//0022-3514.71.2.230

Bouwmeester, S., Verkoeijen, P. P. J. L., Aczel, B., Barbosa, F., Bègue, L., Brañas-Garza, P., … Wollbrant, C. E. (2017). Registered replication report: Rand, Greene and Nowak (2012). *Perspectives on Psychological Science*, *12*(3), 527–542. doi:10.1177/1745691617693624

Broad, W. J., & Wade, N. (1982). *Betrayers of the truth*. New York, NY: Simon & Schuster.

Cacioppo, J. E., Petty, R. E., & Morris, K. J. (1983). Effects of need for cognition on message evaluation, recall, and persuasion. *Journal of Personality and Social Psychology*, *45*(4), 805–818. doi:10.1037//0022-3514.45.4.805

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., … Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. doi:10.1038/s41562-018-0399-z

Chang, J., Zhang, M., Hitchman, G., Qiu, J., & Liu, Y. (2014). When you smile, you become happy: Evidence from resting state task-based fMRI. *Biological Psychology*, *103*, 100–106. doi:10.1016/j.biopsycho.2014.08.003

Cheung, I., Campbell, L., LeBel, E. P., Ackerman, R. A., Aykutoğlu, B., Bahník, Š., … Yong, J. C. (2016). Registered replication report: Study 1 from Finkel, Rusbult, Kumashiro & Hannon (2002). *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *11*(5), 750–764. doi:10.1177/1745691616664694

Crandall, C. S., & Sherman, J. (2016). On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology*, *66*, 91–99. doi:10.1073/pnas.1521897113

Crocker, J., & Cooper, L. (2011). Editorial: Addressing scientific fraud. *Science*, *334*(6060), 1182. Retrieved from http://www.sciencemag.org.proxy.library.uu.nl/content/334/6060/1182.full doi:10.1126/science.1216775

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*, 145–153.

Darwin, C. (1872). *The expression of emotions in man and animals*. London, UK: John Murray.

DeCoster, J., & Claypool, H. M. (2004). A meta-analysis of priming effects on impression formation supporting a general model of informational bias. *Personality and Social Psychology Review*, *8*(1), 2–27. doi:10.1207/S15327957PSPR0801_1

Deutsche Forschungsgemeinschaft (2017). *Stellungsnahme zur Replizierbarkeit von Forschungsergebnissen (Position paper on the replicability of research findings)*. Retrieved from http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/2017/170425_stellungnahme_ replizi erbarkeit_forschungsergebnisse_de.pdf

Dijksterhuis, A. (2018). Reflection on the professor-priming replication report. *Perspectives on Psychological Science*, *13*, 295–296. doi:10.1177/1745691618755705

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*(4), 865–877.

Dijksterhuis, A., van Knippenberg, A., & Holland, R. (2014). Evaluating behavior priming research: Three observations and a recommendation. *Social Cognition*, *32*(Supplement), 196–208. doi:10.1521/soco.2014.32.supp.196

Doyen, S., Klein, O., Pichon, C., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS One*, *7*(1), e29081. doi:10.1371/journal.pone.0029081

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*, 621. doi:10.3389/fpsyg.2015.00621

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., … Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82. doi:10.1016/j.jesp.2015.10.012

Ebersole, C. R., Alaei, R., Atherton, O. E., Bernstein, M. J., Brown, M., Chartier, C. R., … Nosek, B. A. (2017). Observe, hypothesize, test, repeat: Luttrell, Petty and Xu (2017) demonstrating good science. *Journal of Experimental Social Psychology*, *69*, 184–186. doi:10.1016/j.jesp.2016.12.005

Eerland, A., Sherrill, A. M., Magliano, J. P., Zwaan, R. A., Arnal, J. D., Aucoin, P., … Prenoveau, J. M. (2016). Registered replication report: Hart & Albarracín (2011)). *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *11*(1), 158–171. doi:10.1177/1745691615605826

Erdfelder, E., & Ulrich, R. (2018). Zur Methodologie von Replikationsstudien. (The methodology of replication studies). *Psychologische Rundschau*, *69*(1), 3–21. doi:10.1026/0033-3042/a000387

Fiedler, K., & Prager, J. (2018). The regression trap and other pitfalls of replication science illustrated by the report of the Open Science Collaboration. *Basic and Applied Social Psychology*, *40*(3), 115–125. doi:10.1080/01973533.2017.1421953

Fiedler, K., & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science*, *7*(1), 45–52. doi:10.1177/1948550615612150

Fraley, R. C., & Vazire, S. (2014). The N-Pact factor: evaluating the quality of empirical journals with respect to sample size and statistical power. *PLos One*, *9*(10), e109019doi:10.1371/journal.pone.0109019

Gadenne, V. (1984). *Theorie und Erfahrung in der psychologischen Forschung (Theory and experience in psychological research)*. Tübingen, Germany: Mohr & Siebeck.

Gerard, H. B., & Mathewson, G. C. (1966). Effects of the severity of initiation for a group – replication. *Journal of Experimental Social Psychology*, *2*(3), 278–287. doi:10.1016/0022-1031(66)90084-9

Gergen, K. J. (1973). Social psychology as history. *Journal of Personality and Social Psychology*, *26*(2), 309–320. doi:10.1037/h0034436

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comments on "Estimating the reproducibility of psychological science". *Science*, *351*(6277), 1037. doi:10.1126/science.aad7243

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., … Zwienenberg, M. (2016). A multilab pregegistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, *11*(4), 546–573. doi:10.1177/1745691616652873

Hennenlotter, A., Dresel, C., Castrop, F., Ceballos-Baumann, A. O., Wohlschläger, A. M., & Haslinger, B. (2009). The link between facial feedback and neural activity within central circuitries of emotion – new insights from botulin toxin induced denervation of frown muscles. *Cerebral Cortex*, *19*(3), 537–542. doi:10.1093/cercor/bhn104

Inbar, Y. (2016). Association between contextual dependence and replicability in psychology may be spurious. *Proceedings of the National Academy of Sciences*, *113*(34), E4933–E4934. doi:10.1073/pnas.1608676113

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. doi:10.1177/0956797611430953

Klein, R. A. et al. (2018). Many Labs 2: Investigating variation in replicability across samples and setting. *Advances in Methods and Practices in Psychological Science*. *1*(4), 443–490. doi:10.31234/0sf.io

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., … Nosek, B. A. (2014). Investigating variation in replicability. A "Many Labs" replication project. *Social Psychology*, *45*(3), 142–152. doi:10.1027/1864-9335/a000178

Levelt, N. & Drenth Committees (2012). Flawed science: The fraudulent research practices of social psychologist Diederik Stapel. Retrieved from https://www.tilburguniversity.edu/upload/3ff904d7-547b-40ae-85fe-bea38e05a34a_Final%20report%20Flawed%20Science.pdf

Luttrell, A., Petty, R. E., & Xu, M. (2017). Replication and fixing failed replications: The case of need for cognition and argument quality. *Journal of Experimental Social Psychology*, *69*, 178–183. doi:10.1016/j.jesp.2016.09.006

Marsh, A. A., Rhoads, S. A., & Ryan, R. (2018). A multi-semester classroom demonstration yields evidence in support of the facial feedback effect. *Emotion*. doi:10.1037/emo0000532

Noah, T., Schul, Y., & Mayo, R. (2018). When both the original study and its failed replication are correct: Feelings observed cancels the facial-feedback effect. *Journal of Personality and Social Psychology*, *114*(5), 657–664. doi:10.1037/pspa0000121

Nosek, B. A., & Lakens, D. (2014). A method to increase the credibility of published results. *Social Psychology*, *45*(3), 137–141.

O'Donnell, M. (2018). Registered replication report: Dijksterhuis and van Knippenberg (1998). *Perspectives on Psychological Science*, *13*, 268–294. doi:10.1177/1745691618755704

Open Science Collaboration (OSC; (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251)

Payne, B. K., Burkley, M. A., & Stokes, M. B. (2008). Why do implicit and explicit attitude tests diverge? The role of structural fit. *Journal of Personality and Social Psychology*, *94*(1), 16–31. doi:10.1037/0022-3514.94.1.16

Petty, R. E., & Cacioppo, J. T. (1986). *Communication and persuasion: Central and peripheral routes to attitude change*. New York, NY: Springer.

Petty, R. E., & Cacioppo, J. T. (2016). Methodological choices have predictable consequences in replicating studies on motivation to think: Commentary on Ebersole et al. (2016). *Journal of Experimental Social Psychology*, *67*, 86–87. doi:10.1016/j.jesp.2015.12.007

Petty, R. E., Cacioppo, J. T., & Goldman, R. (1981). Personal involvement as a determinant of argument-based persuasion. *Journal of Personality and Social Psychology*, *41*(5), 847–855. doi:10.1037//0022-3514.41.5.847

Popper, K. R. (1959). *The logic of scientific discovery*. London, UK: Hutchinson.

Price, T., & Harmon-Jones, E. (2015). Embodied emotion: The influence of manipulated facial and bodily states on emotive responses. *Wiley Interdisciplinary Reviews: Cognitive Science*, *6*(6), 461–473. doi:10.1002/wcs.1370

Risen, J. L., & Gilovich, T. (2008). Why people are reluctant to tempt fate. *Journal of Personality and Social Psychology*, *95*(2), 293–307. doi:10.1037/0022-3514.95.2.293

Roediger, H. L. (2012). Psychology's woes and a partial cure: The value of replication. Observer. Retrieved from http://www.psychologicalscience.org/index.php/publications/observer/2012/february-12/psychologys-woes-and-a-partial-cure-the-value-of-replication.html

Schweinsberg, M., Madan, N., Vianello, M., Sommer, S. A., Jordan, J., Tierney, W., … Uhlmann, E. L. (2016). The pipeline project: Prepublication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology*, *66*, 55–67. doi:10.1016/j.jesp.2015.10.001

Shnabel, N., & Nadler, A. (2008). A needs-based model of reconciliation: Satisfying the differential emotional needs of victim and perpetrator as a key to promoting reconciliation. *Journal of Personality and Social Psychology*, *94*(1), 116–132. doi:10.1037/0022-3514.94.1.116

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on power studies?. *Psychological Bulletin*, *105*(2), 309–316. doi:10.1037//0033-2909.105.2.309

Strack, F. (2016). Reflection on the smiling registered replication report. *Perspectives on Psychological Science*, *11*(6), 929–930. doi:10.1177/1745691616674460

Strack, F. (2017). From data to truth in psychological science. A Personal Perspective. *Frontiers in Psychology*, *8*, 702.

Strack, F., Martin, L. I., & 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*(5), 768–777. doi:10.1037/0022-3514.54.5.768

Strack, F., & Stroebe, W. (2017). What have we learned?. *What Can we Learn? Behavioral and Brain Sciences*, *41*, e151.

Stroebe, W. (2012). The truth about, But Nobody Seems to Care. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *7*(1), 54–57. doi: 10.1177/1745691611427306

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

Stroebe, W. (2016b). Populist science reporting at the Volkskrant. Retrieved from https://dub.uu.nl/nl/opinie/populist-science-reporting-volkskrant

Stroebe, W., Gadenne, V., & Nijstad, B. A. (2018). Do our laws only apply to students: The problem with external validity. *Basic and Applied Social Psychology*, *40*, 384–395. doi:10.1080/01973533.2018.1513362

Stroebe, W., Postmes, T., & Spears, R. (2012). Scientific misconduct and the myth of self-correction in science. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *7*(6), 670–688. doi: 10.1177/1745691612460687

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, *9*(1), 59–71. doi:10.1177/1745691613514450

Trafimow, D. (2012). The role of auxiliary assumptions for the validity of manipulations and measures. *Theory & Psychology*, *22*, 486–498. doi:10.1177/0959354311429996

Trafimow, D. (2018). An a priori solution to the replication crisis. *Philosophical Psychology*, *31*(8), 1188–1214. doi: 10.1080/09515089.2018.1490707

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences, USA*, *113*(23), 6454–6459.

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., … Zwaan, R. A. (2016). Registered replication report: Strack, Martin, and Stepper (1988). *Perspectives on Psychological Science*, *11*(6), 917–928. doi:10.1177/1745691616674458

Yong, E. (2012). Nobel Laureate challenges psychologists to clean up their act. Nature. Retrieved from https://www.nature.com/news/nobel-laureate-challenges-psychologists-to-clean-up-their-act-1.11535