The Alleged Crisis and the Illusion of Exact Replication

Stroebe, Wolfgang; Strack, Fritz

Published in:
Perspectives on Psychological Science

DOI:
10.1177/1745691613514450

Publication date:
2014

The Alleged Crisis and the Illusion of Exact Replication

Wolfgang Stroebe and Fritz Strack

Perspectives on Psychological Science 2014 9: 59
DOI: 10.1177/1745691613514450

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

The Alleged Crisis and the Illusion of Exact Replication

Perspectives on Psychological Science
2014, Vol 9(1) 59–71

DOI: 10.1177/1745691613514450
pps.sagepub.com
SAGE

Wolfgang Stroebe[1,2] and Fritz Strack[3]
[1]Department of Psychology, Utrecht University, the Netherlands; [2]Department of Social and Organizational Psychology, University of Groningen, the Netherlands; and [3]Department of Psychology, University of Würzburg, Germany

Abstract
There has been increasing criticism of the way psychologists conduct and analyze studies. These critiques as well as failures to replicate several high-profile studies have been used as justification to proclaim a "replication crisis" in psychology. Psychologists are encouraged to conduct more "exact" replications of published studies to assess the reproducibility of psychological research. This article argues that the alleged "crisis of replicability" is primarily due to an epistemological misunderstanding that emphasizes the phenomenon instead of its underlying mechanisms. As a consequence, a replicated phenomenon may not serve as a rigorous test of a theoretical hypothesis because identical operationalizations of variables in studies conducted at different times and with different subject populations might test different theoretical constructs. Therefore, we propose that for meaningful replications, attempts at reinstating the original circumstances are not sufficient. Instead, replicators must ascertain that conditions are realized that reflect the theoretical variable(s) manipulated (and/or measured) in the original study.

Keywords
replication, replicability crisis, null findings, scientific fraud, priming, epistemology, critical rationalism

At a time when social psychologists believed that they had every reason to be proud of their discipline came the shattering news that Diederik Stapel, a prominent researcher in social psychology, had committed scientific fraud on a major scale. Social psychologists had hardly recovered from this shock when two more colleagues were accused of fraud and resigned from their positions. These events were particularly damaging, because they coincided with the start of a discussion of trust in psychological data (see Special Section on Replicability in Psychological Science: A Crisis of Confidence? *Perspectives on Psychological Science*, 2012). Even though this discussion focused on methodological issues that were unrelated to fraud, this distinction was not always maintained by the popular press. According to the introduction to the special section, there is "currently a crisis of confidence in psychological science reflecting an unprecedented level of doubt among practitioners about the reliability of research findings in the field" (Pashler & Wagenmakers, 2012, p. 528). Nosek and colleagues started the "Reproducibility Project," a large-scale, collaborative effort to estimate the reproducibility of psychological science, which involves replicating all studies published in three psychology journals in 2008 (http://openscienceframework.org). In contrast to the prevalent sentiment, we will argue that the claim of a replicability crisis is greatly exaggerated and that the hope that such a crisis (if it ever existed) could be solved by increasing the number of exact replications is misplaced.

Is the Claim of a Replicability Crisis Exaggerated?

There seems to have been two sets of events that fueled the crisis perception. First, there have been claims that some psychological researchers engage in "questionable research practices" that result in "false positive" findings (e.g., Bakker, van Dijk, & Wicherts, 2012; John, Loewenstein, & Prelec, 2012; LeBel et al., 2013; Simmons, Nelson, & Simonsohn, 2011; Vul, Harris, Winkielman, & Pashler, 2009). Simmons et al. (2011) reported simulations that showed that a number of research practices (e.g., stopping data collection on the basis of interim data analysis;

Corresponding Author:
Wolfgang Stroebe, Department of Psychology, Utrecht University, P.O. Box 80140, 3508 TC, Utrecht, the Netherlands
E-mail: w.stroebe@uu.nl

dropping experimental conditions from published reports) can result in an erroneous rejection of the null hypothesis. To be sure, methodological discussions are important for any discipline, and both fraud and dubious research procedures are damaging to the image of any field and potentially undermine confidence in the validity of social psychological research findings. Thus far, however, no solid data exist on the prevalence of such research practices in either social or any other area of psychology. In fact, the discipline still needs to reach an agreement about the conditions under which these practices are unacceptable.

Second, there have been a number of failures to replicate high-profile experiments on social priming (Doyen, Klein, Pichon, & Cleeremans, 2012; Pashler, Coburn, & Harris, 2012; Shanks et al., 2013). For example, Doyen et al. (2012) reported an exact replication of a study by Bargh, Chen, and Burrows (1996; Experiment 2a, 2b) that failed to reproduce the original results. Bargh et al. had repeatedly found that students who had been primed with words that triggered the stereotype of elderly walked more slowly down a corridor than students primed with words unrelated to the elderly stereotype. The Doyen et al. (2012) failure was soon followed by a report by Shanks et al. (2013) of a series of nine studies that failed to replicate the findings of another iconic experiment, the "professor study" of Dijksterhuis and van Knippenberg (1998). These authors had found that participants who were primed with a category of persons who are considered highly intelligent (e.g., professors) performed better on a task of trivial pursuit than did participants primed with a category of persons who are considered less intelligent (e.g., hooligans). These replication failures were the more astounding because of earlier publications of successful replications of both of these findings (summarized in the Appendix).

Do these failures to replicate amount to a crisis of replicability, as Pashler and Harris (2012) claimed in their contribution to the special section of *Perspectives on Psychological Science* on replications? We would argue that such a conclusion is premature. Failures to replicate are puzzling, but in social psychology, as in most sciences, empirical findings cannot always be replicated (this was one of the reasons for the development of meta-analytic methods). It is therefore surprising that the failures to replicate some social priming studies received such disproportionate attention. Furthermore, one must wonder whether the response by Kahneman, who in a widely circulated letter to "priming researchers" warned that there was a "train wreck looming" because of a "storm of doubt about the robustness of priming results," was really justified given the state of the published literature where priming is an entirely undisputed method that is widely used to test hypotheses about associative memory (e.g., Higgins, Rholes, & Jones, 1977; Meyer & Schvaneveldt, 1971; Tulving & Schacter, 1990). In this tradition, its impact on behavior was studied in a minor subarea of research, whereas most social psychological priming studies investigated the impact of subliminal or supraliminal primes on judgment. A meta-analysis of studies that investigated how trait primes influence impression formation identified 47 articles based on 6,833 participants and found overall effects to be statistically highly significant (DeCoster & Claypool, 2004).

Are Exact Replications the Answer?

The claim of a replicability crisis in psychology is based on a major misunderstanding. Particularly, the myopic focus on "exact" replications neglects basic epistemological principles. Exact replications are replications of an experiment that operationalize both the independent and the dependent variable in exactly the same way as the original study. (In contrast, conceptual replications try to operationalize the underlying theoretical variables using different manipulations and/or different measures.)

In evaluating the usefulness of exact replications, one has to distinguish between applied and basic research. A scientist who wants to establish the efficiency of a specific treatment or intervention is well advised to repeatedly apply exactly the same procedure. This is particularly relevant for clinical trials where a lack of reliability may have fatal consequences. However, matters are different in basic research where empirical outcomes are meaningful only with respect to the theory being tested. In the postbehaviorist era, psychological theories are based on internal mechanisms such that replications must be directed at the internal antecedents of such theories. Although reproducibility of scientific findings is one of science's defining features, the ultimate issue is the extent to which a theory has undergone strict tests and has been supported by empirical findings. It would be a mistake to assume that estimates of the reproducibility of empirical findings are the same as estimates of the validity of a specific theory. A finding may be eminently reproducible and yet constitute a poor test of a theory.

The fact that good experimental research is typically conducted with the aim to test theories throws a different light on the discussion of replicability. We will therefore briefly discuss the notion of theory and what researchers do when they test hypotheses derived from a theory.

Theories consist of a set of abstract constructs and of hypotheses about the relationship between these constructs. In conducting experiments to test such a hypothesis, we develop empirical operationalizations to translate these constructs into variables that can be manipulated or measured. Because most theoretical constructs are fairly abstract and can be operationalized in multiple ways,

researchers can never be sure whether they have chosen a realistic or even optimal operationalization of a given construct. Because researchers can never be certain that they properly operationalized the theoretical constructs they are assessing and that they were successful in controlling for all third variables that might have been responsible for their findings, a theory can never be proven to be "true" (Popper, 1959). However, as we will discuss later, this same feature may also create problems for falsifying a theory.

This reservation is less relevant for studies that are conducted with the aim of merely testing the efficacy of a specific treatment. For example, an exact replication of a study to demonstrate the efficacy of a drug or psychological intervention is informative, because with drugs or interventions (i.e., treatments), the main issue is that they work and that they have no negative side effects. Although one might want to vary the subject population receiving the treatment to have a broader basis for one's evaluation of its efficacy, it would make no sense to use a different drug or alter the intervention. After all, the question to be addressed is whether the original treatment was effective.

Exact replications are also important when studies produce findings that are unexpected and only loosely connected to a theoretical framework. Thus, the fact that priming individuals with the stereotype of the elderly resulted in a reduction of walking speed was a finding that was unexpected. Furthermore, even though it was consistent with existing theoretical knowledge, there was no consensus about the processes that mediate the impact of the prime on walking speed. It was therefore important that Bargh et al. (1996) published an exact replication of their experiment in the same paper. Similarly, Dijksterhuis and van Knippenberg (1998) conducted four studies in which they replicated the priming effects. Three of these studies contained conditions that were exact replications. In the fourth, they primed "intelligent" directly with the trait rather than indirectly with the word "professor."

Because it is standard practice in publications of new effects, especially of effects that are surprising, to publish one or two exact replications, it is clearly more conducive to the advancement of psychological knowledge to conduct conceptual replications rather than attempting further duplications of the original study, unless plausible conceptual replications have failed. Given that both research time and money are scarce resources, the large-scale attempts at duplicating previous studies seem to us misguided (http://openscienceframework.org).

The main criticism of conceptual replications is that they are less informative than exact replications (e.g., Pashler & Harris, 2012). This raises two questions, namely (a) what information do we gain from a successful replication of the original finding after faithfully repeating the original experiment, and (b) what do we learn when we have failed to replicate the original finding? Let us address each of these issues in the following sections.

A comparison of direct and conceptual replications

The illusion of exact replication. If one accepts that the true purpose of replications is a (repeated) test of a theoretical hypothesis rather than an assessment of the reliability of a particular experimental procedure, a major problem of exact replications becomes apparent: Repeating a specific operationalization of a theoretical construct at a different point in time and/or with a different population of participants might not reflect the same theoretical construct that the same procedure operationalized in the original study. This is less of a problem in studies where both the independent and the dependent variables are not culturally or socially mediated, for example, when the size or the brightness of stimuli is assessed by participants (who work in isolation) or when weight perception is related to the size of an object. Obviously, in such studies, one does not have to worry whether these variables have been properly realized, and exact replications should yield the findings of previous studies. However, in social psychological studies, the faithful replication of an operationalization of a theoretical construct at a different point in time and with a different subject population may be dissociated from the theoretical construct of the original study.

Let us illustrate this point with some classic social psychological experiments. In their study of the effect of the severity of initiation to a group on liking for that group, Aronson and Mills (1959) operationalized the severe initiation by having female participants read aloud "12 obscene words, e.g., fuck, cock, and screw" as well "two vivid descriptions of sexual activity from contemporary novels" (p. 178). If repeated with today's female students, this manipulation might trigger amusement rather than embarrassment. Similarly, it is likely that a researcher who tried to induce fear about toothbrushing in high school students by telling them that improper care of their teeth might result in "cancer, paralysis or other secondary diseases" (Janis & Feshbach, 1953) might arouse disbelief rather than fear.

Why outcomes of exact replications are often uninformative. Let us now return to the failed replications of behavior priming studies described earlier. To discuss these failures, readers need to be reminded of the theoretical rationale behind these studies. According to current theorizing (e.g., Loersch & Payne, 2011; Schröder & Thagard, 2013; Strack & Deutsch, 2004), the priming activates concepts that spread activation to other concepts that are episodically or semantically linked (e.g., "elderly" → "walking slowly"; "professors" → "intelligent"). Then,

priming may affect behavior in a controlled fashion if a behavioral decision is based on concepts whose activation potential has been increased. Alternatively, these concepts may be directly linked with behavioral schemata. Of course, such subtle influences depend on other conditions, such as factors that facilitate or constrain the execution of such behaviors and the awareness of the priming episode, which may cause an active correction (see Strack & Hannover, 1996; Strack, Schwarz, Bless, Kübler, & Wänke, 1993). Thus, subtly priming or even subliminal priming procedures may be more effective than blatantly directing people's attention toward some content.

The theoretical variable "activation of concept X" is manipulated by exposing the person to some element of "X" or an element that is closely associated. Although the experimenter has control over the prime, this is not true for the concept it activates. People differ in their beliefs about the elderly. Furthermore, different beliefs are differently accessible in different contexts (e.g., an old athlete vs. an old professor). Therefore, the same prime can activate different concepts in different people and/or under different conditions.

It is crucial for replicating behavior priming studies that the prime is successful in increasing the accessibility of the cognitive representation that is assumed to induce the behavior. In priming the stereotype of the elderly, the theoretically targeted cognitive representation is that of "walking slowly." It is therefore possible that the priming procedure used in the Doyen et al. (2012) study failed in this respect, even though Doyen et al. faithfully replicated the priming procedure of Bargh et al. (1996). First, the French translation of the words Bargh et al. used in their scrambled sentence test might have been associated with different meanings in Belgium. Second, it is also possible that the concept of "walking slowly" is not a central part of the stereotype of elderly in Belgium some 20 years later. With life expectancies increasing nearly every year and people staying active to a much higher age (Stroebe, 2011), this construct might no longer form part of the elderly stereotype even in New York. Although Doyen et al. (2012) tested whether they succeeded in priming the stereotype of elderly people, they failed to assess whether this prime increased the accessibility of the cognitive representation of "walking slowly."[1]

This criticism also applies to the failed replications of the Dijksterhuis and van Knippenberg (1998) "professor studies" reported by Shanks et al. (2013). Although it seems likely that professors are considered more intelligent than soccer hooligans by the students who served as subjects in Shank's research, it is still possible that even if the participants considered professors as more intelligent than hooligans, the priming manipulation might have failed to increase the cognitive representation of the concept "intelligence." It is even possible that the fact that these findings are reported in most social psychology textbooks and are therefore widely known among student participants could have affected the results. Another likely reason for their failure could be their selection of knowledge items. If the effect of professor priming is motivational, then the knowledge items have to be selected in a way as to best reflect motivation effects. This is unlikely to be the case for questions that are so difficult that students are unlikely to know the answer or so easy that practically everybody can answer them easily.

To be sure, these possibilities are speculative, but they illustrate that nonreplications are uninformative unless one can demonstrate that the theoretically relevant conditions were met. Faithfully replicating the original conditions of an experiment does not guarantee that one addresses the same theoretical construct as in the original study. In order to check if a given operationalization is successful in manipulating the intended theoretical construct, manipulation checks are necessary that are independent of the dependent variable. Therefore, instead of conducting nine "exact" replications, Shanks et al. (2013) would have been well advised to conduct at least some empirical tests of whether their manipulation succeeded in increasing the cognitive accessibility of the theoretically relevant concept (e.g., using a lexical decision task). Such an approach would have provided the information that exact replications are lacking.

The effective use of replications. Because experiments are typically conducted with the aim of testing a theoretical hypothesis, the important question is not whether the original finding can be duplicated but whether it constituted a rigorous test of the postulated mechanism. To take frustration–aggression theory as an example, any experimental test of that theory is based on the assumption that the experimental manipulation is a valid operationalization of the theoretical construct "frustration" and that the measure of aggression is a valid operationalization of the theoretical construct "aggression." If we replicate the findings of an earlier frustration–aggression experiment by exactly repeating the procedure of that experiment, we have demonstrated that the study is reproducible, but we have only marginally increased our trust in the validity of the underlying theory. Conversely, if a conceptual replication using a different operationalization of both constructs had succeeded in supporting the theoretical hypothesis, our trust in the validity of the underlying theory would have been strengthened. "With every difference that is introduced the confirmatory power of the replication increases," because we have shown that the phenomenon does not hinge on a particular operationalization but "generalizes to a larger area of application" (Schmidt, 2009, p. 93).

An even more effective strategy to increase our trust in a theory is to test it using completely different

manipulations. Let us illustrate this with an example from the social psychology of persuasion. According to dual process theories of persuasion (i.e., the elaboration likelihood model of Petty and Cacioppo [1986] or the heuristic-systematic model of Chaiken [1980]), the impact of the quality of the arguments contained in a communication is greater the more thoughtfully and deeply the communication is processed by a recipient. This prediction has been supported with very different manipulations of the processing depth, such as distraction (Petty, Wells, & Brock, 1976), personal relevance (Petty, Cacioppo, & Goldman, 1981), expectation to have to discuss the communication at a future meeting (Chaiken, 1980), and need for closure (Klein & Webster, 2000). These models further predict that when recipients are unable or unmotivated to process a communication, they will rely more on heuristic cues than on argument quality. This prediction was supported by Petty et al. (1981) using communicator credibility as heuristic cue and by Chaiken (1980) as well as by Klein and Webster (2000) using the number of arguments. Thus, even though the various experiments were very different and used different experimental manipulations, different attitude issues, and different dependent measures, they all tested the same underlying theory. If this theory had been less valid, further empirical studies guided by them would have been likely to fail, even if original studies had yielded (false) positive results.

Thus, one reason why exact replications are not very interesting is that they contribute little to scientific knowledge. If an exact replication reinstates the finding of the original study, we have learned that the original outcome was reproducible. We are not any wiser as to whether the original study was a good test of the theory to be tested because even though the experiment may have been poorly designed, a faithful replication might result in the same finding. Conversely, if we succeed in supporting the theoretical hypothesis with an experiment that operationalized both the independent and the dependent variable differently and thus sampled different parts of the same theoretical concept, we have gained additional information and increased our trust in the underlying theory.

Why null findings are not always that informative

Related to the drive for an increase in exact replications is the argument that we should publish more null findings. This idea has a long history. Already in 1975, Greenwald published an article on the "Consequences of prejudice against the null hypothesis" in which he concluded "that research traditions and customs of discrimination against accepting the null hypothesis may be very detrimental to research progress" (p. 1). Recognizing this problem, a group of social psychology graduate students at the University of North Carolina at Chapel Hill in 1970 had started the journal *Representative Research in Social Psychology*, which gives special consideration to null findings and replications (Chamberlin, 2000). Since 2002, there has also been the free-access *Journal of Articles in Support of the Null Hypothesis*. Another free-access journal, *PLoS One*, also publishes articles reporting null results. Most recently, Pashler and colleagues created the Web site www.psychfiledrawer.org, where short reports of null findings can be posted. Thus, there is no shortage of outlets for the publication of negative findings.

The problem is that these journals lack the prestige of some of the high-impact journals in social psychology that have typically rejected articles reporting null findings. Therefore, strategies have been suggested to encourage exact replications of published research findings through changes in the incentive system of our discipline (e.g., Koole & Lakens, 2012). For example, the editor of the *Journal of Research in Personality* now encourages and publishes high-quality replications in the hope that "JRP will be able to provide critical information about which initial discoveries really hold up over time" (http://www.journals.elsevier.com/journal-of-research-in-personality/news/professor-richard-lucas-encouraging-replication-studies). Furthermore, the Open Science Collaboration—a group of more than 175 volunteer researchers—initiated the Reproducibility Project to replicate a sample of studies published in 2008 in three major psychology journals. The aim is to obtain an "estimate of the reproducibility of current psychological science" (Open Science Collaboration, 2012, p. 658). As one scientist whose supportive comment is cited in that document argued,

> If [our] discoveries are true, other people in other places should be able to rely on them, and build on them, to push further, and build a cumulative body of knowledge. These values are central to what it is, for me, to be a scientist, and the Reproducibility Project expresses those values. . . . The project exclusively attempts *direct* replications "repetition of an experimental procedure" in order to "verify a piece of knowledge." (Schmidt, 2009, pp. 92–93)

The illusion of theoretical verification. Statements such as this raise the question about what precisely is meant by the "discovery" that needs to be "verified." Is the discovery of the Aronson and Mills (1959) study their empirical finding that girls who had to read aloud four-letter words evaluated a group discussion of the sex life of lower animals more positively than girls who had to

read words that were related to sex but not obscene (e.g., prostitute, virgin, and petting), or is it the support of the hypothesis derived from dissonance theory that members who had gone through a severe initiation should overestimate the attractiveness of their group to reduce dissonance?

If it is the empirical finding, then other people in other places should better not rely on the results, unless the other people are other Americans. It is doubtful that the manipulation would have worked outside the United States or Northern Europe even in 1959 (Arnett, 2008; see also Stroebe & Nijstad, 2009).[2] Furthermore, as we argued earlier, even Americans would be unlikely to reproduce these findings with this particular experimental manipulation today. If the discovery that needs to be "verified" is the theoretical hypothesis, then other people in other places should not rely on it too much either, because the bad news is that theories cannot be verified or even demonstrated as probable (Popper, 1959). And with regard to the effect of the severity of initiation on the attractiveness of a group, some revision of the original hypothesis seems certainly necessary (e.g., Lodewijkx & Syroit, 1997; Lodewijkx, van Zomeren, & Syroit, 2005).

What can we learn from null findings? If knowledge about null findings is so important, why does our discipline seem to be so uninterested? One reason is that not all null findings are interesting. For example, just before his downfall, Stapel published an article on how disordered contexts promote stereotyping and discrimination. In this publication, Stapel and Lindenberg (2011) reported findings showing that litter or a broken-up sidewalk and an abandoned bicycle can increase social discrimination. These findings, which were later retracted, were judged to be sufficiently important and interesting to be published in the highly prestigious journal *Science*. Let us assume that Stapel had actually conducted the research described in this paper and failed to support his hypothesis. Such a null finding would have hardly merited publication in the *Journal of Articles in Support of the Null Hypothesis*. It would have been uninteresting for the same reason that made the positive result interesting, namely, that (a) nobody expected a relationship between disordered environments and prejudice and (b) there was no previous empirical evidence for such a relationship. Similarly, if Bargh et al. (1996) had found that priming participants with the stereotype of the elderly did not influence walking speed or if Dijksterhuis and van Knippenberg (1998) had reported that priming participants with "professor" did not improve their performance on a task of trivial pursuit, nobody would have been interested in their findings. Thus, null findings are interesting only if they contradict a central hypothesis derived from an established theory and/or are discrepant with a series of earlier studies.

If we accept that the ultimate aim of research is to test theories and if we further accept the epistemological position of critical rationalism (Popper, 1959) that theories can only be falsified but never be proven true, then such negative findings should be influential, because they help us to falsify theories. However, our earlier discussion of the conflict surrounding the professor-priming studies also signals a major problem with Popper's position, namely, that it is difficult to decide when a theory should be considered falsified. The theory underlying the professor studies is that priming people with the concept of professor will increase in their minds the cognitive accessibility of "clever" or "intelligent." According to some not yet specified process, the increase in the cognitive accessibility of the cognitive representations of these concepts will improve their performance on a knowledge test.

The nonreplications published by Shanks and colleagues (2013) cannot be taken as a falsification of that theory, because their study does not explain why previous research was successful in replicating the original findings of Dijksterhuis and van Knippenberg (1998). Thus, although the Shanks et al. (2013) study signals a potential problem with the stability of the professor-priming effect, it offers no explanation for the discrepancy between their findings and that of earlier studies. Even multiple failures to replicate an established finding would not result in a rejection of the original hypothesis, if there are also multiple studies that supported that hypothesis. Because failures of exact replications do not tell us why findings cannot be replicated, they are ultimately not very informative. The believers will keep on believing, pointing at the successful replications and derogating the unsuccessful ones, whereas the nonbelievers will maintain their belief system drawing on the failed replications for support of their rejection of the original hypothesis. Thus, one reason why null findings are not very interesting is because they tell us only that a finding could not be replicated but not why this was the case. This conflict can be resolved only if researchers develop a theory that could explain the inconsistency in findings. Such a theory should allow one to identify the factor that determined whether the priming effect occurs. Thus, instead of testing a theory against the null hypothesis, a more successful strategy is to test it against an alternative hypothesis.

Nonreplications as interaction effects. In the ongoing discussion, "failures to replicate" are typically taken as a threat to the existence of the phenomenon. Methodologically, however, nonreplications must be understood as interaction effects in that they suggest that the effect of the crucial influence depends on the idiosyncratic conditions under which the original experiment was conducted. Of course, interaction effects are highly

informative if the additional variables are psychologically meaningful. However, the conditions of "exact replications" are psychologically unspecified and include an infinite number of influences that may come from the cultural circumstances, the experimental setting, characteristics of the participants, and so forth. The resulting ambiguity prevents anything to be learned from such an interaction. As a consequence, the mere coexistence of exact replications that are both successful and unsuccessful is likely to leave researchers helpless about what to conclude from such a pattern of outcomes. Conducting exact replications in a registered and coordinated fashion by different laboratories does not remove the described shortcomings. This is also the case if exact replications are proposed as a means to estimate the "true size" of an effect. As the size of an experimental effect always depends on the specific error variance that is generated by the context, exact replications can assess only the efficiency of an intervention in a given situation but not the generalized strength of a causal influence.

Instead of leaving the circumstances unspecified, the interaction can be informative if the conditions of replication attempts are psychologically defined. This can be illustrated with the classic example of dissonance aroused by counterattitudinal advocacy. Festinger and Carlsmith (1959) argued that a person will experience dissonance if the person believes "X" but has, as a result of pressure brought on the individual, publicly stated that he or she believes "not X." They further proposed that the magnitude of the dissonance would be maximal if the promised rewards or threatened punishments were just barely enough to induce the person to say "not X." Although their theoretical analysis was valid, it took a decade before researchers were able to reliably replicate the findings reported by Festinger and Carlsmith (1959). Replication became possible only once they realized that for counterattitudinal behavior to arouse dissonance, the actor had to be made to feel responsible for telling a lie (i.e., freedom of choice; Linder, Cooper, & Jones, 1967) and that that lie had negative consequences for the other person (Cooper & Worchel, 1970). This strategy was in line with the argument of Greenwald, Pratkanis, Leippe, and Baumgardner (1986), that to reduce a needlessly overgeneralized theory, one needs to identify the conditions under which this theory applies.

Resolving conflicts between competing theories. The search for moderator variables that determine the conditions under which theories apply is also the most effective strategy for resolving conflicts between competing theories that offer explanations for the same psychological phenomena. After discussing the various empirical strategies that researchers used to distinguish between impression management and intrapsychic explanations for a variety of research findings, Tetlock and Manstead (1985) concluded that neither side emerged as clear winner. They suggested that a more profitable strategy was "to abandon the search for crucial experiments and to focus on clarifying the points of similarity and dissimilarity between rival intrapsychic and impression management theories" (p. 71). Both theories should be considered as special cases of an integrative theoretical framework. Similarly, dissonance theory (Festinger, 1957) and self-perception theory (Bem, 1965) were established as competing theories. And yet decades of research attempting to demonstrate the superiority of one over the other resulted in conflicting findings. To resolve this conflict, it was suggested that both theories were valid but applied only under certain conditions (e.g., Fazio, Zanna, & Cooper, 1977; Stroebe & Diehl, 1988).

If we look at the history of social psychology, theories have rarely been abandoned because of failed replications. Theories are often abandoned because researchers simply lose interest (Greenwald, 2012). The reason for this loss has often been that they no longer matched the prevailing theoretical paradigm (Kuhn, 1996; for example, the rise of social cognition research resulted in a loss of interest in the motivations explanations offered by consistency theories). As Hilton (2012) observed, "social psychology does seem to have a disquieting tendency to forget theories rather than disprove them" (p. 69). But more frequently, competing theories are not abandoned but reduced in their generality. That is, what originally seemed to be a simple main effect turned out to be a more complex interaction.

Replications as fraud detectors

This leads us to a last argument in justification of exact replications, namely, that they facilitated the detection of fraudulent research. For example, Crocker and Cooper (2011, p. 1182) argued: "Despite the need for reproducible results to drive progress, studies that replicate or fail to replicate others' findings are almost impossible to publish in top scientific journals. This disincentive means fraud can go undetected, which was the case with Stapel." Similarly, Roediger (2012, p. 27) stated "that if others had tried to replicate his [Stapel's] work soon after its publication, his misdeeds might have been uncovered much more quickly." Along the same lines, Chambers and Sumner (2012) wrote:

> Replication is our best friend because it keeps us honest. In science, false results have a short (albeit potentially damaging) lifespan because regardless of how they come about, other scientists won't be able to reproduce them. On the other hand, true results will be replicated time and time again by different scientists. (para. 10)

Finally, Mummendey (2012, p. 7) went even further and suggested:

> Scientific journals could expand their already high standards of the peer review system by adding the requirement for a thorough external replication. Authors submit their manuscript together with their data. Once the publication has been approved by a preliminary group of reviewers, the editors invite suitable experts to attempt a replication of the results. After this has been accomplished, both the original manuscript and the replication study are published together.

Replications are poor fraud detectors. Existing evidence, however, sheds doubt on the success of such strategies. In a recent article, Stroebe, Postmes, and Spears (2012) analyzed 40 fraud cases to identify the way by which the fraud had been detected. Because this is rarely reported in official reports, they had to rely on fraud cases that were sufficiently significant to be discussed in local or national papers, because unlike agencies such as the U.S. Office of Research Integrity, journalists typically like to know how fraudsters were discovered. To their own surprise, Stroebe and colleagues found that replications hardly played any role in the discovery of these fraud cases. As was the case with Stapel (Levelt, Noort, & Drenth, 2012), most fraud cases are detected by close colleagues or research or graduate student assistants of the fraudster, who act as whistleblowers (Stroebe et al., 2012). Fraudsters are also frequently identified by outside colleagues, who work in the same field and find inconsistencies in journal articles that are often editorial in nature (e.g., the same figure for different outcomes). For example, the fraud of the Norwegian cancer researcher John Sudbo, like Stapel a young star player in his field, was discovered by the head of the epidemiology division at the Norwegian Institute of Public Health, who realized that the cancer patient database used in the study had not yet been available at the time of the study (http://en.wikipedia.org/wiki/Jon_Sudb%C3%B8).

In social psychology, as in many areas of medicine and even physics, replication failures are due to incomplete description of the methodology that was used in a study or because important moderators were not spelled out by a theory. Given that fraud (as far as we know) is extremely rare, it is not surprising that failed replications are not diagnostic about fraud.

Meta-analyses are poor fraud detectors. That this limitation even applies to meta-analyses can be illustrated with the findings of a meta-analysis of priming studies by DeCoster and Claypool (2004). They reported that the sizes of priming effects were larger in studies conducted in countries outside the United States and Canada. The authors wrote:

> We had no a priori expectation that nationality would moderate priming effects; moreover, there is little reason to believe that the underlying psychological principles responsible for assimilation, anchoring, or correction effects would differ cross-culturally. We therefore suspect that these effects simply reflect idiosyncratic differences between U.S./Canadian labs and those elsewhere. (p. 10)

With hindsight, one striking difference is that Stapel was first author on 7 of the 10 assimilation studies, on 6 of the 7 anchoring contrast studies, and on 2 of the 3 correction contrast studies that were responsible for this nationality effect. Should one have suspected fraud? Probably not. It is quite possible that European (student) participants take experiments more seriously or differ in other ways from American undergraduates, who serve as subjects in most social psychological research.

The findings of the meta-analysis of DeCoster and Claypool (2004) indicate another problem with replications as indicator of fraud, namely, that except for their effect sizes, the findings of the priming research reported by Stapel and colleagues were quite consistent with multiple studies conducted in the United States. It is typical for successful fraudsters that they avoid producing unexpected findings and that their results are in line with the expectations of the field. Some of the Stapel articles that had been included in this meta-analysis were identified as fraudulent in the final report of the three committees that had been appointed to investigate the Stapel affair (Levelt et al., 2012). And yet even though Stapel may have been somewhat overenthusiastic in improving the data, his findings have been replicated by other authors. As Stapel wrote in his autobiography, he was always pleased when his invented findings were replicated: "What seemed logical and was fantasized became true" (Stapel, 2012). Thus, neither can failures to replicate a research finding be used as indicators of fraud, nor can successful replications be invoked as indication that the original study was honestly conducted.

Conclusion

Psychologists have no reason for complacency. The fact that one of our colleagues committed one of the greatest research frauds in recent history should be a wake-up call. Therefore, even though we are (or at least should be) aware of the methodological standards of good research, it is helpful to be reminded that certain procedures can substantially distort the results of the statistical tests (Simmons et al., 2011).

As a field, we have reacted (or are at least planning to react) to these challenges by instituting measures that will increase the risk of discovery for people who still engage in these practices. Thus, if the rule will be

instituted that for all published studies the materials used as well as the data are stored at a publicly accessible Web site, many questionable practices will become publicly visible. Although this regulation cannot prevent people from submitting only a selection of their measures and from "cleaning up" their data before submission, it should act as a deterrent. And it will certainly allow interested parties to reanalyze data to see whether the original conclusions are indeed supported. Unfortunately, there has been a great deal of resistance from major journal publishers (e.g., the American Psychological Association, the Association for Psychological Science) against instituting such a rule, but most Dutch and a few U.S. universities have now done so at the university level.

Another promising improvement is the Web site for reporting successful and unsuccessful replications (www.psychfiledrawer.org). But whereas it will certainly be useful to be informed about studies that are difficult to replicate, we are less confident about whether the investment of time and effort of the volunteers of the Open Science Collaboration is well spent on replicating studies published in three psychology journals. The result will be a reproducibility coefficient that will not be greatly informative, because of justified doubts about whether the "exact" replications succeeded in replicating the theoretical conditions realized in the original research.

As social psychologists, we are particularly concerned that one of the outcomes of this effort will be that results from our field will be perceived to be less "reproducible" than research in other areas of psychology. This is to be expected because for the reasons discussed earlier, attempts at "direct" replications of social psychological studies are less likely than exact replications of experiments in psychophysics to replicate the theoretical conditions that were established in the original study.

Although psychologists should not be complacent, there seem to be no reasons to panic the field into another crisis. Crises in psychology are not caused by methodological flaws but by the way people talk about them (Kruglanski & Stroebe, 2012). Although the discussion of research practices in social psychology is healthy and may ultimately lead to an improvement of the rules that guide us on the way we analyze, report, and store our data, magnifying the present problems into a "replicability crisis" is likely to be counterproductive in the long run.

Appendix: Failures to Replicate Social Priming Studies

Bargh and Colleagues' priming walking speed

Bargh, Chen, and Burrows (1996, Experiment 2a, 2b) primed their participants with a scrambled-sentence test that contained either words that evoked the elderly stereotype or words that were unrelated. After the experiment had apparently been finished, a second experimenter measured how fast participants were walking down a corridor. Participants primed with the elderly stereotype walked significantly slower than those primed with unrelated traits. Moreover, a replication of that study by the same authors using different participants resulted in nearly identical results.

Several successful conceptual replications have been published (Aarts & Dijksterhuis, 2002; Cesario, Plaks, & Higgins, 2006; Dijksterhuis, Spears, & Lepinasse, 2001; Kawakami, Young, & Dovidio, 2002; Ku, Wang, & Galinsky, 2010; Macrae et al., 1998). For example, the participants of the study of Ku et al. (2010, Study 3) were shown a photograph of an older individual and were subsequently asked to write a narrative about a day in the life of that person. Half the participants were instructed to take the perspective of that person in their story writing, while the other half were told to write as objectively as possible. In line with the original predictions, participants afterward walked more slowly if they had been given the instructions to take the elderly perspective than to be objective. Dijksterhuis et al. (2001) and Kawakami et al. (2002) showed that elderly primes slowed down reactions to a lexical decision task. Macrae et al. (1998) found that participants read a word list faster when primed with "Michael Schumacher" (at that time Formula One world champion) than with a neutral prime. Finally, Mussweiler (2006) demonstrated the reverse effect, namely, that moving more slowly led participants to ascribe more elderly-stereotypic characteristics to a target (Experiment 2).

Cesario et al. (2006, Experiment 2) conducted another conceptual replication that differed from the original study only in the method of priming (i.e., subliminal presentation of pictures of elderly [and young people] rather than words in scrambled sentences). Cesario and colleagues were only partially successful in replicating the findings of Bargh et al. (1996). Participants were found to walk more slowly and took more time to walk to the exit after having been subliminally primed with pictures of older people than participants primed with pictures of young people, but walking speed in both conditions did not differ significantly from an unprimed control group. Thus, although the study demonstrated that priming with social categories can influence walking speed, it did not replicate the finding of Bargh et al. (1996) that priming with the elderly stereotype reduces walking speed in comparison to an unprimed control group. This difference could have been due to the different type of priming used.

The only exact replication of the study was conducted by Doyen et al. (2012), who repeated the original study, with the only difference that walking speed was

measured objectively by infrared sensor rather than with a stopwatch. They did not find any difference between the two priming conditions in their first experiment. In their second experiment, they did observe an effect but only if experimenters were explicitly instructed to expect the primed participants to walk more slowly. This raised the possibility that the original findings had been caused by the experimenters' expectations and that the second experimenter measuring walking speed with a stopwatch was somehow aware of the experimental condition. However, in their method section, Bargh et al. (1996) clearly stated that experimenters and observers were blind to experimental conditions. Furthermore, experimenter expectation affected walking speed even when it was measured objectively in the Doyen et al. (2012) study. Thus, it remains unclear how experimenter expectations could have influenced the participants' speed of walking.

The "professor studies" of Dijksterhuis and van Knippenberg

Dijksterhuis and van Knippenberg (1998) primed participants either with a category of persons who are usually considered highly intelligent (e.g., professor) or with a category of persons who are considered less intelligent (e.g., hooligan). The priming manipulation consisted of participants having to think of a typical professor (or hooligan) and then list lifestyle, behavior, and appearance attributes of the typical member of the primed category. In an apparently unrelated second experiment, participants had to answer a number of general knowledge questions taken from trivial pursuit. Dijksterhuis and van Knippenberg found that thinking about a category of intelligent persons resulted in better performance of the trivial pursuit task than thinking of a category of less intelligent persons and conducted four replications of the finding. Conceptual replications have been reported by several different laboratories from different countries, with most articles containing multiple studies (e.g., Bry, Follenfant, & Meyer, 2008; Galinsky, Wang, & Ku, 2008; Haddock, Macrae, & Fleck, 2002; Hansen & Wänke, 2009; LeBoeuf & Estes, 2004; Lowery, Eisenberger, Hardin, & Sinclair, 2007; Nussinson, Seibt, Häfner, & Strack, 2010). However, recently Shanks and colleagues (2013) conducted nine exact replication studies. In none of the studies did the manipulation yield a statistically significant effect on performance.

The findings reported by Shanks et al. (2013) are not only inconsistent with the successful replications of the professor studies, they are also inconsistent with other findings showing that stereotypes can influence individual performance on intellectual tasks. For example, Steele and Aronson (e.g., 1995) have repeatedly demonstrated that the mere salience of a stereotype is sufficient to impair stereotype-related performance. Some years later, Wheeler, Jarvis, and Petty (2001) showed this effect to operate not only with self-stereotyping but also with stereotyping of others. One such mechanism has recently been examined in a functional MRI study in which participants had to perform a memory task ($n - 2$ back task for letters) after having been primed with either the word "clever" or "stupid" (Bengtsson, Dolan, & Passingham, 2011). Findings showed that participants who had been primed with "clever" spent more time making a response following an error. At the same time, this resulted in increased activation in the anterior paracingulate cortex on error trials. This finding is consistent with a role of the anterior paracingulate that is considered to be involved in monitoring task performance.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Notes

1. It must be acknowledged that such manipulation checks are missing not only in replications but also in the original experiments.
2. This was one of the major arguments that Gergen (1973) used to argue "that social psychology is primarily an historical inquiry. Unlike the natural sciences, it deals with facts that are largely nonrepeatable and which fluctuate markedly over time" (p. 310).

References

Aarts, H., & Dijksterhuis, A. (2002). Category activation effects in judgment and behaviour: The moderating role of perceived comparability. *British Journal of Social Psychology*, *41*, 123–128.

Arnett, J. J. (2008). The neglected 95%: Why American psychology needs to become less American. *American Psychologist*, *63*, 602–614.

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.

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*, 543–554. doi:10.1177/1745691612459060

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

Bem, D. J. (1965). An experimental analysis of self-persuasion. *Journal of Experimental Social Psychology*, *1*, 199–218.

Bengtsson, S. L., Dolan, R., & Passingham, R. E. (2011). Priming for self-esteem influences monitoring of one's own performance. *Scan*, *6*, 417–425. doi:10.1093/scan/sq048

Bry, C., Follenfant, A., & Meyer, T. (2008). Blonde like me: When self-construals moderate stereotype priming effects

on intellectual performance. *Journal of Experimental Social Psychology*, *44*, 751–757.

Cesario, J., Plaks, J. E., & Higgins, T. E. (2006). Automatic social behavior as motivated preparation to interact. *Journal of Personality and Social Psychology*, *90*, 893–910.

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology*, *39*, 725–766.

Chamberlin, J. (2000, May). A student publishing tradition. *APA Monitor on Psychology*, *31*, 36. Retrieved from http://www.apa.org/monitor/may00/rrsp.aspx

Chambers, C., & Sumner, P. (2012, September 14). Replication is the only solution to scientific fraud. *The Guardian*. Retrieved from http://www.guardian.co.uk/commentisfree/2012/sep/14/solution-scientific-fraud-replication

Cooper, J., & Worchel, S. (1970). Role of undesirable consequences in arousing cognitive dissonance. *Journal of Personality and Social Psychology*, *16*, 199–206.

Crocker, J., & Cooper, L. (2011, December 2). Editorial: Addressing scientific fraud. *Science*, *334*, 1182. doi:10.1126/science.1216775.

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*, 2–27. doi:10.1207/S15327957PSPR0801_1

Dijksterhuis, A., Spears, R., & Lepinasse, V. (2001). Reflecting and deflecting stereotypes: Assimilation and contrast in impression formation and automatic behavior. *Journal of Experimental Social Psychology*, *37*, 286–299. doi:10.1006/jesp.2000.1449

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.

Doyen, S., Klein, O., Pichon, C.-L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE*, *7*, e29081. doi:10.137/journal.pone.0029081

Fazio, R. H., Zanna, M. P., & Cooper, J. (1977). Dissonance and self-perception: An integrative view of each theory's proper domain of application. *Journal of Experimental Social Psychology*, *13*, 464–479.

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. Retrieved from http://ori.hhs.gov/sites/default/files/gallup_finalreport.pdf

Galinsky, A. D., Wang, C. S., & Ku, G. (2008). Perspective-takers behave more stereotypically. *Journal of Personality and Social Psychology*, *95*, 404–419.

Gergen, K. J. (1973). Social psychology as history. *Journal of Personality and Social Psychology*, *26*, 309–320.

Greenwald, A. G. (1975). Consequences of prejudice against the null-hypothesis. *Psychological Bulletin*, *82*, 1–20.

Greenwald, A. G. (2012). There is nothing so theoretical as a good method. *Perspectives on Psychological Science*, *7*, 99–108.

Greenwald, A. G., Pratkanis, A. R., Leippe, M. R., & Baumgardner, M. H. (1986). Under what conditions does theory obstruct research progress? *Psychological Review*, *93*, 216–229.

Haddock, G., Macrae, C. N. & Fleck, S. (2002). Syrian science and smart supermodels: On the when and how of perception–behavior effects. *Social Cognition, 20*, 461–481.

Hansen, J., & Wänke, M. (2009). Think of capable others and you can make it! Self-efficacy mediates the effect of stereotype activation on behavior. *Social Cognition, 27*, 76–88.

Higgins, T. E., Rholes, W. S., & Jones, C. R. (1977). Category accessibility and impression formation. *Journal of Experimental Social Psychology*, *13*, 141–154.

Hilton, D. (2012). The emergence of cognitive social psychology: A historical analysis. In A. Kruglanski & W. Stroebe (Eds.), *Handbook of the history of social psychology* (pp. 45–79). New York, NY: Psychology Press.

Janis, I. L., & Feshbach, S. (1953). Effects of fear-arousing communications. *Journal of Abnormal and Social Psychology*, *48*, 78–92.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*, 524–532. doi:10.1177/095679611430953

Kawakami, K. L., Young, H., & Dovidio, J. F. (2002). Automatic stereotyping: Category, trait, and behavioral activations. *Personality and Social Psychology Bulletin*, *28*, 3–15.

Klein, C. T. F., & Webster, D. M. (2000). Individual differences in argument scrutiny as motivated by need for cognitive closure. *Basic and Applied Psychology*, *22*, 119–129.

Koole, S. L., & Lakens, D. (2012). Rewarding replications: A sure simple way to improve psychological science. *Perspectives on Psychological Science*, *7*, 608–614.

Kruglanski, A. W., & Stroebe, W. (2012). The making of social psychology. In A. W. Kruglanski & W. Stroebe (Eds.), *Handbook of the history of social psychology* (pp. 3–18). New York, NY: Psychology Press.

Ku, G., Wang, C. S., & Galinsky, A. D. (2010). Perception through a perspective-taking lens: Differential effects on judgments and behavior. *Journal of Experimental Social Psychology*, *46*, 792–798.

Kuhn, T. S. (1996). *The structure of scientific revolutions* (3rd ed.). Chicago, IL: University of Chicago Press.

LeBel, E. P., Borsboom, D., Giner-Sorolla, R., Hasselman, F., Peters, K. R., Ratliff, K. A., & Tucker Mith, C. (2013). PsychDisclosure.org: Grassroots support for reforming standards in psychology. *Perspectives on Psychological Science*, *8*, 424–432.

LeBoeuf, R. A., & Estes, Z. (2004). "Fortunately, I'm no Einstein": Comparison relevance as a determinant of behavioral assimilation and contrast. *Social Cognition, 22*, 607–636.

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

Linder, D. E., Cooper, J., & Jones, E. E. (1967). Decision freedom as a determinant of the role of incentive magnitude in

attitude change. *Journal of Personality and Social Psychology*, *6*, 245–254.

Lodewijkx, H. F. M., & Syroit, J. E. M. M. (1997). Severity of initiation revisited: Does severity of initiation increase attractiveness in real groups? *European Journal of Social Psychology*, *27*, 275–300.

Lodewijkx, H. F. M., van Zomeren, M., & Syroit, J. E. M. M. (2005). The anticipation of a severe initiation: Gender differences in effects on affiliation tendency and group attraction. *Small Group Research*, *36*, 237–262. doi:10.1177/1046496404272381

Loersch, C., & Payne, B. K. (2011). The situated inference model: An integrative account of the effects of primes on perception, behavior, and motivation. *Perspectives on Psychological Science*, *6*, 234–252.

Lowery, B. S., Eisenberger, N. I., Hardin, C. D., & Sinclair, S. (2007). Long-term effects of subliminal priming on academic performance. *Basic and Applied Social Psychology*, *29*, 151–157.

Macrae, C. N., Bodenhausen, G. V., Milne, A. B., Castelli, L., Schloerscheidt, A. M., & Greco, S. (1998). On activating exemplars. *Journal of Experimental Social Psychology*, *34*, 330–354.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of dependence between retrieval operations. *Journal of Experimental Psychology*, *90*, 227–234.

Mummendey, A. (2012). Scientific misconduct in social psychology—Towards a currency reform in science. *European Bulletin of Social Psychology*, *24*, 4–7.

Mussweiler, T. (2006). Doing is for thinking! Stereotype activation by stereotypic movement. *Psychological Science*, *17*, 17–21.

Nussinson, R., Seibt, B., Häfner, M., & Strack, F. (2010). Come a bit closer: Approach motor actions lead to feeling similar and behavioral assimilation. *Social Cognition*, *28*, 40–58.

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, *7*, 657–660.

Pashler, H., Coburn, N., & Harris, C. R. (2012). Priming of social distance? Failure to replicate effects on social and food judgments. *PLoS ONE*, *7*, e42510. doi:10.1371/journal.pone.0042510

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, *7*, 531–536. doi:10.1371/journal.pmed.0020124

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*, 528–530.

Petty, R. E., & Cacioppo, J. T. (1986). *Communication and persuasion*. New York, NY: Springer.

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*, 847–855.

Petty, R. E., Wells, G. L., & Brock, T. C. (1976). Distraction can enhance or reduce yielding to propaganda: Thought disruption versus effort justification. *Journal of Personality and Social Psychology*, *34*, 874–884.

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

Roediger, H. L. (2012, February). Psychology's woes and a partial cure: The value of replication. *Observer, 25*(2), 9, 27–29.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, *13*, 90–100. doi:10.1037/a0015108

Schröder, T., & Thagard, P. (2013). The affective meanings of automatic social behaviors: Three mechanisms that explain priming. *Psychological Review*, *120*, 255–280.

Shanks, D. R., Newell, B. R., Lee, E. H., Balakrishnan, D., Ekelund, L., Cenac, Z., . . . Moore, C. (2013). Priming intelligent behavior: An elusive phenomenon. *PLoS ONE*, *8*, e56515. doi:10.1371/journal.pone.0056515

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.

Stapel, D. (2012). *Ontsporing* [Derailment]. Amsterdam, the Netherlands: Prometheus.

Stapel, D. A., & Lindenberg, S. (2011, April 8). Coping with chaos: How disordered contexts promote stereotyping and discrimination. *Science*, *332*, 251–252.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*, 797–811.

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review*, *8*, 220–247.

Strack, F., & Hannover, B. (1996). Awareness of influence as a precondition for implementing correctional goals. In P. Gollwitzer & J. A. Bargh (Eds.), *The psychology of action: Linking cognition and motivation to behavior* (pp. 579–595). New York, NY: Guilford.

Strack, F., Schwarz, N., Bless, H., Kübler, A., & Wänke, M. (1993). Awareness of the influence as a determinant of assimilation versus contrast. *European Journal of Social Psychology*, *23*, 53–62.

Stroebe, W. (2011). *Social psychology and health*. Maidenhead, England: Open University Press.

Stroebe, W., & Diehl, M. (1988). When social support fails: Supporter characteristics in compliance induced attitude change. *Personality and Social Psychology Bulletin*, *14*, 136–144.

Stroebe, W., & Nijstad, B. (2009). Do our psychological laws apply only to Americans? *American Psychologist*, *64*, 569. doi:10.1037/a0016090

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

Tetlock, P. E., & Manstead, A. S. R. (1985). Impression management versus intrapsychic explanations in social psychology: A useful dichotomy? *Psychological Review*, *92*, 59–77.

Tulving, E., & Schacter, D. L. (1990, January 19). Priming and human memory systems. *Science*, *247*, 301–306.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzling high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, *4*, 274–288.

Wheeler, S. C., Jarvis, W. B. G. & Petty, R. E. (2001). Think onto others: The self-destructive impact of negative racial stereotypes. *Journal of Experimental Social Psychology, 37*, 173–180.