Journal of Experimental Social Psychology 66 (2016) 93–99

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

On the scientific superiority of conceptual replications for scientific progress

Christian S. Crandall [a,*], Jeffrey W. Sherman [b]

[a] University of Kansas, USA
[b] University of California, Davis, USA

ARTICLE INFO

Article history:
Received 1 November 2014
Revised 1 October 2015
Accepted 11 October 2015
Available online 24 March 2016

Keywords:
Replication
Conceptual replications
Social psychology
Philosophy of science

ABSTRACT

There is considerable current debate about the need for replication in the science of social psychology. Most of the current discussion and approbation is centered on *direct* or exact replications, the attempt to conduct a study in a manner as close to the original as possible. We focus on the value of *conceptual* replications, the attempt to test the same theoretical process as an existing study, but that uses methods that vary in some way from the previous study. The tension between the two kinds of replication is a tension of values—exact replications value confidence in operationalizations; their requirement tends to favor the status quo. Conceptual replications value confidence in theory; their use tends to favor rapid progress over ferreting out error. We describe the many ways in which conceptual replications can be superior to direct replications. We further argue that the social system of science is quite robust to these threats and is self-correcting.

> Any working scientist is more impressed with 2 replications in each of 6 highly dissimilar experimental contexts than he is with 12 replications of the same experiment. (Meehl, 1990, p. 111.)

Scientific ideas must be robust before they can be endorsed. They must be testable and they must inspire the confidence of a skeptical audience. There are many ways that these ideas acquire a reliable place in the marketplace of ideas: elegance, intuitiveness, explanatory power, rigor, and so on. But the ability of a phenomenon to be replicated is a necessary condition for widespread acceptance by scientists (Schmidt, 2009). Catchy, interesting, telling, and surprising results can all have currency, but if an effect proves unreliable or impossible to replicate, support for the idea will not—must not—persist.

> Only when certain events recur in accordance with rules or regularities, as is the case with repeatable experiments, can our observations be tested—in principle—by anyone. We do not take even our own observations quite seriously, or accept them as scientific observations, until we have repeated and tested them. (Popper, 1959, p. 23).

Time and chance occur for every data set—it is only through replication that we can be confident. Over the past few years, a concern about the repeatability and replicability of experiments has spread throughout the social psychological community. In its wake, a variety of articles, blog posts, research programs, and non-profit organizations have come forward with prescriptions for a more replicable science of social psychology (e.g., Open Science Collaboration, 2012). Some of this work has been well funded and well covered by science journalists (e.g., Meyer & Chabris, 2014). There is a broad consensus in favor of robust findings, for reliability in the scientific record, for high quality research with dependable reporting and replicability, and for progress in scientific knowledge. But there are sharp differences among scientists in (1) which scientific goals should take priority over others and (2) the best way to meet those respective goals.

1. Two kinds of replications

One of the most prominent fault lines among scientists is in their approaches to replication. Following standard discourse, we divide replication into two kinds—exact replication and conceptual replication. An *exact* or *direct* replication is an attempt to conduct a study, usually published in a peer-reviewed journal, in a manner as close to the original as possible. An exact replicator seeks to use the same materials, the same manipulations, the same dependent variables, and the same kind of participants as the originally published article. A replication is considered "successful" when the exact replication results in a pattern of data that mimics—or is close to—the original article's findings.

The second class of replications is known as *conceptual* replications. A conceptual replication is an attempt to test the same fundamental idea or hypothesis behind the original study, but the operationalizations of the phenomenon, the independent and dependent variables, the type and design of the study, and the participant population may all differ substantially. (Others have called this a distinction between replicability [exact replication] and repeatability [conceptual

* Corresponding author.

http://dx.doi.org/10.1016/j.jesp.2015.10.002
0022-1031/

replication], e.g., Casadevall & Fang, 2010; Drummond, 2009) The critical difference between an exact and a conceptual replication is whether or not they share the same operationalizations of the theory. Although this distinction is common in discussions among social psychologists, the scientific and philosophical literature on the matter is surprisingly scant (cf. McGrath, 1981).

In some fields of science, an exact replication is a sensible proposition. In physics, chemistry, biology or geology, the processes that affect an outcome are usually transhistorical and transcultural—language, politics, and social history rarely affect gravity, electron weight, the structure of proteins, or water flow through limestone. The meaning of the operationalization is consensual among scientists; the value of mass or acceleration, bone density, or the Mohs hardness scale, for example.

But in matters of social psychology, one can never step in the same river twice—our phenomena rely on culture, language, socially primed knowledge and ideas, political events, the meaning of questions and phrases, and an ever-shifting experience of participant populations (Ramscar, 2015). At a certain level, then, all replications are "conceptual" (Stroebe & Strack, 2014), and the distinction between direct and conceptual replication is continuous rather than categorical (McGrath, 1981). Indeed, many direct replications turn out, in fact, to be conceptual replications. At the same time, it is clear that direct replications are based on an attempt to be as exact as possible, whereas conceptual replications are not.[1]

2. Replications and theoretical consequences

> *The meaning of theoretical terms cannot be totally exhausted by operational definitions, but the ways in which theoretical terms function in science cannot be understood in the absence of the ways in which they are operationalized.* (Hull, 1988, pp. 516–517).

There is no controversy over the need for replication; virtually all scientists and philosophers of science endorse the notion that replication of one sort or another is absolutely essential. The controversy is largely over the degree to which different kinds of replications advance scientific knowledge. Historically, research in psychology has favored conceptual over direct replications. Most researchers were trained to value the pursuit of robustness and generality of theoretical ideas over the repeatability of a particular study. However, recent observations that direct replications may reproduce original findings at a lower rate than expected have led to calls for increasing the frequency and publication of direct replications (Open Science Collaboration, 2015). At the same time, conceptual replications have been increasingly criticized for biasing research toward confirmation and impeding the possible disconfirmation of research findings and the theories they support (e.g., Nosek, Spies, & Motyl, 2012; Pashler & Harris, 2012; Roberts, 2014).

In terms of both published commentary (Pashler & Wagenmakers, 2012) and public social media discussions on the "replication crisis," those favoring a shift toward direct replication have received a great deal of attention. As described above, it is certainly true that, historically, the potential advantages of direct replication have received relatively little attention. As such, the recent discussions can be seen as an important corrective to a historical imbalance. At the same time, the virtues of conceptual replication have been overlooked or directly challenged in these venues. Given the field's historical predilection toward conceptual replication, this may be of little consequence. However, it is our experience that psychological researchers, particularly those in the early stages of training, are increasingly prone to dismiss the potential benefits of conceptual replication in favor of direct replication. Moreover, the recent focus on direct replication seems to have created the perception of consensus that direct replications are of greater value than conceptual replications. Again, we have noticed these tendencies particularly among younger members of our research guild. In this context, we believe that it is important to provide a corrective and to articulate the importance of conceptual replication. The primary purpose of this paper is to do that and, at the same time, offer suggestions to increase the scientific value of conceptual replication.

3. The purposes of direct and conceptual replication

Whereas direct replications enhance one's confidence in operationalizations, conceptual replications enhance one's confidence in theoretical hypotheses. In 1906, the physicist Pierre Duhem (1906/1954) pointed out that every empirical scientific test was comprised of a conjunction of the theoretical hypothesis and its operationalizations—every empirical test conflates *ideas* with *methods*. A "failure" of an empirical test is always ambiguous, because the failure may indicate that the idea is incorrect (e.g., a failed conjecture or "falsification," Popper, 1963/2002), or it may indicate that the operationalization process failed, or both. Exact replications can never speak to this ambiguity, they can only perpetuate it; this makes straightforward falsification a logical impossibility (see also Meehl, 1990; Quine, 1980).

Conceptual replications disperse this ambiguity, and as a result, can contribute more to theoretical development and scientific advance. If an idea replicates *across* operationalizations, then the idea is substantially more likely to be correct than if it replicates using the exact same operationalizations, no matter how many times or with whatever precision. As such, conceptual replications are critical for establishing the *generalizability* of an initial observation and the theory it purports to support.

The history of science is replete with examples in which an original demonstration is met with substantial skepticism, with specific complaints about confounds, alternative explanations, or concerns about the effectiveness of methods. In these cases, direct replications meet with exactly the same complaints, but conceptual replications can prove far more persuasive. Mackay and Oldford (2000) reviewed critical studies about whether the speed of light was infinite or merely very, very fast. In 1671, Ole Rømer measured anomalies in the timing of eclipses of Io, one of Jupiter's moons, and showed that these anomalies could be accounted for by the distance between Earth and Jupiter; when Earth and Jupiter were closer, the eclipse came sooner. Because Rømer knew the size of the Earth's orbit, he was able to make a fairly accurate estimate of the speed of light. But this research did not convince many of the leading scientists of the day, including Descartes and Giovanni Cassini, the director of the Paris Observatory, who offered many criticisms and alternative explanations for the data, despite the fact that Rømer replicated his findings over many years (Soter & deGrasse Tyson, 2000). It wasn't until 1729, when James Bradley used the parallax motion of stars (because stars are different distances from Earth, the Earth's movement causes an apparent shifting of the stars' relative position) to calculate the speed of light, and calculated a very similar number to Rømer's. Following the use of this technique the scientific community agreed that the speed of light was not infinite, but rather about 300,000,000 m per second.

Exact replications are often unconvincing to the scientific community—skeptics require a different method for test of the same hypothesis. The second operationalization often dispels spurious criticisms of the first study's method (MacKay & Oldford, 2000). Exact replication could create precision in estimation, but it would not convince the scientific community of *the meaning of the observation*; a conceptual replication did.

[1] If one prefers the continuous approach, the reader might interpret this paper to favor movement in the direction of greater conceptual replication and away from more exact replications.

4. Increasing the robustness of operations versus theories

4.1. Direct replications

One way to frame the key differences between direct and conceptual replication is to consider the unit of analysis in each. In direct replications, the unit of analysis is the previously observed effect. The question to be addressed is *to what extent can that effect be repeated*? Such a replication cannot tell us any more about the meaning of the effect or any more about why the effect is obtained. The effect must be assumed to apply only to the specific operationalizations of independent and dependent variables. In this case, the replication tells us little about an idea or theory (see Table 1). It can only increase or decrease our confidence about the relations between a particular operation and a particular dependent variable. Because operationalization always limits a concept to its representations and measurement (Bridgman, 1927), an exact replication maintains the narrow focus of the original empirical test; it breaks no new ground, it adds little new information, it cannot establish generality.

This delimited repetition of a particular operationalization can create replications that perpetuate a flawed design. The history of social psychology is replete with examples in which direct replication led to repeated error. Brehm's (1956) landmark paper on cognitive dissonance and the "spreading of alternatives" was the first empirical publication on cognitive dissonance. This effect has been directly replicated many times in different labs, across different decades, and using different kinds of subjects, including monkeys (Egan, Santos, & Bloom, 2007). This finding represents a triumph of direct replication. However, as it turns out, there was an important flaw in the design of the study that compromised random assignment and, therefore, the interpretation of the results. The many direct replications of the study included this exact same flaw. Chen and Risen (2010) identified the flaw and reported findings that raised important questions about the conditions under which the spreading of alternatives will be observed. Directly replicating Brehm (1956) perpetuated the flaw and, because of that flaw, provided almost nothing worthwhile about the nature of dissonance.

James Stoner's (1961) MIT master's thesis found that group discussion of a series of vignettes led participants to support riskier choices than they had made as individuals. This effect came to be known as the "risky shift" phenomenon. What followed was a decade of extremely vigorous research by dozens of social psychologists investigating risky shift (see Brown, 1965). Because the Stoner materials were created in a way that favored risk, everyone who used the materials exactly (as so many did) failed to discover that a conservative shift was equally likely, if provided stimulus materials that generated an initial consensus in a conservative direction. The failure to recognize that risky shift was simply a small—and biased—portion of the group polarization effect was hampered by the repeated use of the exact materials Stoner (1961) developed. Instead, with a conceptual replication, Moscovici and Zavalloni (1969) showed that the phenomenon was *actually* group polarization—groups with initial consensus tend to become more extreme through discussion. Their work greatly expanded our understanding of the effect. Similar concerns about bias repeated in exact replications can also be found in confounding politics with racial prejudice (Sniderman, Piazza, Tetlock, & Kendrick, 1991, cf. Henry & Sears, 2002) and the measure of gender stereotypes and the structure of gender roles (Bem, 1981; Spence, 1993)

An emphasis on direct replications focuses a field's attention and resources toward upholding versus upending specific findings, rather than toward building a theoretical framework that is robust across specific operationalizations. Direct replications take a careful look at the status quo, but rather than provide and test an alternative, they must accept the choices of a prior scientist or lab. They can support the finding or they can question the finding, but they cannot replace the finding.

There is a certain irony to the limited focus and structure of exact replications. One of the primary instigators for growth in direct replication was the disbelief of counter-intuitive, flashy, high-profile effects (e.g., exotic embodiment and priming effects). Rather than contributing to an understanding of the proposed theories behind those effects, direct replication emphasizes the significance of the effect, in and of itself. If psychologists care more about theories than specific effects, then they should focus on conducting research that adds to understanding of theory, and not focus obsessively on particular instantiations of a theory.

Table 1
The value of exact and conceptual replications: Success and failure.

Kind of replication	Replication "Succeeds"	Replication "Fails"
Exact	• Increases confidence in methods of previous study • Theoretical reach unaffected • Confidence in theory modestly improved	• Methods of first study come under some suspicion • Original support for theory weakened
Conceptual	• Increases confidence in reach of theory • Supports *theoretical* interpretation of previous study • Methods of both studies earn confidence	• Focus on methods of replication • Support for theory diminishes • Alternative explanations for previous study considered

4.2. Conceptual replications

By contrast, *ideas* are the unit of analysis in conceptual replication. The question becomes not whether a specific finding may hold, but whether a theory can be retained in the face of multiple and variable tests of its hypotheses. If the predictions of a theory are supported across a range of operationalizations of independent and dependent variables, then we gain confidence that we have learned something about a theory, rather than a single effect.

The testing of multiple operations is important in any field in which the key constructs are difficult to operationalize with high consensus. What does it mean to be intelligent? What is prejudice? What is the best way to induce a positive mood? In comparison to key constructs in the physical and biological sciences, the manipulation and measurement of our constructs is extraordinarily ambiguous, and often changes across contexts, time, and population. Conceptual replication is a critical means to create consensus about the meaning of our results. To the extent that exact replications require time, energy, money, imagination, journal pages, editor and reviewer time, and substantial draw on expert human labor, they will compete with conceptual replications and innovative research that is substantially more likely to test theoretical hypotheses and advance scientific progress.

Consider an example of a demonstration rooted in conceptual replication by one of the authors and colleagues (Eidelman, Crandall, Goodman, & Blanchar, 2012). Their theoretical notion was that political conservatism (including tendencies to view individuals as responsible for their own outcomes, prefer the status quo, and accept of hierarchy) comes directly from the architecture of cognition. That is, there are non-ideological cognitive structures and processes that lead people to prefer conservative ideology, "restricting people to simple and basic modes of thought will lead to the acceptance of conservative attitudes and values" (p. 809). Eidelman et al. tested this hypothesis in four separate conceptual replications. Restricting thought to the "simple and basic" was operationalized by (1) high blood alcohol content, (2) cognitive load through a simultaneous listening task, (3) reducing time available to make judgments, and (4) simple instructions not to think too deeply. Moreover, several different measures of conservatism were employed. In each case, reducing people's ability or motivation to think deeply led participants to a higher level of endorsed conservatism (meta-analytic effect size, $r = .34$).

In Eidelman et al., the exact same theoretical hypothesis was tested four times using different methods (in three different states, with different gender ratios, among students and townies, across a couple of years); these were conceptual replications. The strength of the empirical contribution is in the demonstration that the operationalizations *didn't matter*. The paper had no direct replications. Indeed, direct replications would have made the paper *weaker*, not stronger, if, in the place of one of the operationalizations, the authors had conducted an exact replication of one of the operationalizations. This would have increased the likelihood that the results were due to specific instantiations of the independent and dependent variables. But, because the authors varied the independent and dependent variables across studies, they could make a strong claim to having learned something general about the relationship between cognitive processes and political ideology (Van Berkel, Crandall, Eidelman, & Blanchar, 2015, used the same strategy to show that low cognitive effort leads to endorsement of hierarchy and reduced endorsement of egalitarian values.).

4.3. Science is a social phenomenon

If science were to turn on a single experiment, a single statistical demonstration, or a single p-value, it would follow error almost endlessly. Much of the criticism and panic in psychology in recent years has been about the accuracy of individual statistical hypothesis tests. Simmons, Nelson, and Simonsohn (2011) showed that researchers without scruples can nearly always find a $p < .05$ in a data set if they set their minds to it. Ioannidis (2005), in his over-titled paper "Why Most Published Research Findings are False," argued that small samples, research that looks into small effects, data mining, and the biasing of financial interest can all lead to Result sections that may not reflect nature's own structure.

Underpowered studies and small effects are common enough (Maxwell, 2004), and the rising popularity of power analysis and study planning (e.g., Funder et al., 2014) is likely to change this. But of these complaints, only financial interest is likely to have a sustained effect across multiple studies and different labs. Science tolerates error and it tolerates chance. Published findings are refuted, alternative explanations proposed, and, most effectively, anomalous data are ignored.

4.4. Individual p-values and science as community

Scientific progress cannot be assigned to solitary p-values. The current standard for judging progress is science is multiple publications, using multiple operations, by multiple labs, over multiple years. This is what good scientific practice is (Fraley & Vazire, 2014; Lakatos & Musgrave, 1970; Laudan, 1978; Meehl, 1967), and this is what social psychologists do. Before scientists accept a phenomenon, they want substantial reliability, replicability, and variation and variety in the demonstrations. In this context, individual p-values are a nearly insignificant component of scientific progress (or its problems).

The practice of p-hacking is a good example of the disconnect between individual behavior and group results. P-hacking and its close relations have been discussed for decades (Crandall, 2001; Simmons et al., 2011), and several practices potentially relevant to the strategy have been identified (John, Loewenstein, & Prelec, 2012). Certainly some p-values reported in articles are made smaller by scientists seeking the opportunity to publish (Kicinski, 2013). But the critical issue is how much bias—and consequent error—is introduced by individual p-hacking? In an ambitious review that looked at 2844 scientific journals across a broad range of science, with 114,720 articles representing over a million statistical tests, Head and colleagues (Head, Holman, Lanfear, Kahn, & Jennions, 2015) found evidence that p-hacking takes place in 4 out of the 14 scientific domains (including "psychology and cognitive sciences"). But they also found that "while p-hacking is probably common, its effect seems to be weak relative to the real effect sizes being measured." As social psychologists, we are particularly sensitive to how phenomena can take very different shape when found at the individual and group levels. Even in the presence of occasional misbehavior with probability reporting, Head et al. (2015) concluded, "p-hacking probably does not drastically alter scientific consensuses drawn from meta-analyses."[2]

4.5. Individual scientists as replicators

> *Particular kinds of social arrangements make good epistemic use of the grubbiest motives.* (Kitcher, 1993, p. 305).

Each scientist must consider whether they wish to engage in exact or conceptual replication. The rewards for innovation are far greater than for replication (Fuller, 1997; Open Science Collaboration, 2012, cf. Greenwald, 2015) and, we have argued, the addition to scientific progress is greater for conceptual replications than exact ones. Kitcher (1995) in his essay on "Organization of Cognitive Labor" discusses how individual scientists should consider engaging in exact replication. Kitcher points out that scientists' goals are for "credit" for making discoveries, and that credit is given for being perceived to be right; thus, one must be both accurate and first to the post. Exact replications can never be first *by definition*, and so there is small credit assigned to direct replications. And so, though direct replication is good for science, it can be bad for individual scientists (little credit is given, regardless of the outcome of an exact replication).

Kitcher (1995, pp. 348–352) provides a mathematical analysis of the value of pursuing existing (replicative) versus alternative scientific methods. He demonstrates that, for any individual scientist, choosing either strategy is risky and potentially suboptimal for the scientist (and that only a minority is likely to pursue direct replications). But, some scientists pursue innovation as a strategy, and others pursue replication and work within existing procedures and paradigms. As a result, risk is distributed throughout the scientific endeavor, and the field ultimately behaves more optimally than the individual scientists within it. Our point is not that exact replications should not be done or that conceptual replications are the *only* strategy to pursue. Instead, we argue that not all individual scientists need to pursue exact replications—this is almost certainly inefficient, suboptimal, and against the wishes and desires of many scientists. Kitcher's (1995) analysis reminds us that the field as a whole is better served when *some* (and only some) scientists pursue replication.

4.6. Progress is an outcome of collective activity

The most famous argument that scientific progress results from collective activity comes from David Hull's (1988) *Science as a Process*. Hull followed biological taxonomists and analyzed science from an evolutionary biological point of view. According to this view, science works as a collective human activity in which people compete for credit. One consequence is that scientists have an ambivalent relationship with trusting one another's work.

> *Each scientist has only a few decades to contribute to science. Time cannot be wasted checking every single knowledge-claim* [by one's self] *before it is accepted. Accepting without testing makes scientific progress possible, but it also increases the likelihood that some of the knowledge-claims accepted by science will be mistaken. However, one should not forget that knowledge by acquaintance is also far from fallible. Some of the erroneous views that scientists come to hold are of their own making. Whether the knowledge acquired is first, second, or third hand,*

[2] This does not remove the necessity to estimate publication bias or p-hacking in meta-analysis. But it suggests that science, in general, and meta-analyses, in particular, can often be fairly robust to minor misbehavior on the part of individual scientists (see also Kicinski, 2013).

it can always be mistaken. If science required infallibility of absolute certainty, it would be in trouble. We can be happy that it does not (Hull, 1988, p. 439).

The current focus of criticism in social psychology has been directed toward the individual paper, failing to understand science as a system and failing to recognize that individual papers riven with error, bias, or fraud do not, by themselves, undermine progress in science. The collective process of science is far more robust than the individual study, scientist, or p-value. The most important thing is not whether an individual paper is "false" but rather, as a field, we can make good judgments about research areas, phenomena, or techniques. Indeed, authors, readers, reviewers, and editors all have a responsibility toward modesty when considering the results of a single study or paper. Authors' claims should be modest, readers' conclusions should be modest, expectations about replications should be modest (Stanley & Spence, 2014) and reviewers' and editors' expectations should be modest. The critics' focus on single studies does not target the right level. Because we must be skeptical of every paper, as scientists, we need to look across studies and labs and methods. And, when possible, we should conduct cumulative meta-analyses of research programs involving both direct and conceptual replications.

4.7. What we do not claim

A number of philosophers and scientists have been attracted to a picture of science as a dialogue between an imaginative voice and a critical voice (Godfrey-Smith, 2009, p. 165).

On occasion, the Great Replication Debate has been framed in terms of doing *good* vs. *bad science*. Both direct and conceptual replication can be good or bad science. Replications can be done well or they can be done poorly; we do not invoke the quality of the research itself. Instead, we argue there is a difference in valuing operational reproduction compared to theoretical generalization. Regardless of whether research is entirely novel, pursues conceptual replication, or pursues direct replication, studies conducted with methodological precision that are highly powered are more informative.

The debate has also been framed in terms of *Type I* versus *Type II error*. Our argument is framed in terms of what is being learned from an empirical test. In the case of direct replications, Type I and II error refer to false positives and negatives about a particular set of operations. In the case of conceptual replications, Type I and II error refer to false positives and false negatives about theoretical hypotheses. It is perfectly reasonable that a scientist may be particularly concerned about Type I error at the operational level and Type II error at the theoretical level. A similar argument applies to discussion of prevention versus promotion motives (Crowe & Higgins, 1997) and need for certainty and structure (Neuberg & Newsom, 1993; Webster & Kruglanski, 1994). Do scientists seek theoretical or empirical certainty? Do they promote operational or theoretical innovation? Preferences about replications for individual scientists will come from answers to these values-based questions.

4.8. When do we most need exact replications?

There are many cases in which careful attention to exact replication is essential. In social psychology, we may wish to establish the reliability of a particular operation, especially if that manipulation might lead to policy recommendations. For matters of policy, for program evaluation, for psychotherapy outcome studies—for many practical application matters—the operationalizations can be at least as important as the theoretical idea itself. In these cases, exact replications are necessary and ethically mandated. In such cases, it also can be important to get precise estimates of effect sizes in order to calibrate and predict practical outcomes.

4.9. Leveraging failed conceptual replications

We certainly do not mean to imply that our scientific practices are ideal and that there is no room for improvement; that is not the case. One important opportunity for advancement is putting failed conceptual replications to better use. Though a great deal has been said about the importance of publishing failed direct replications, and a number of journals have established practices for doing so, little to nothing has been said about doing the same for conceptual replications. However, just as we believe that successful conceptual replications are critical for scientific advance, so, too, can be failed conceptual replications. A number of authors have suggested that conceptual replications are useful only when they confirm a theory, but are not useful for disconfirming a theory (e.g., Nosek et al., 2012; Pashler & Harris, 2012). The problem, according to critics, is that failed conceptual replications are dismissed as not providing a good test of the theory in question. That is, when a conceptual replication fails to support a theory, rather than reduce our belief in the theory, we are tempted to explain the failure in terms of methodological problems with the operationalization of the key variables. As such, conceptual replication has been described by critics as solely a mechanism for confirmation bias.

In considering this argument, it is worth noting that the same problem confronts failed direct replications. A motivated party can always find some reason to dismiss a failed direct replication. Indeed, in the past few years, a number of reports of failed replications of priming studies in social psychology have been challenged based on methodological differences between the original study and the direct. Here, we are again confronted with the unfortunate reality that direct replication in social psychology is practically impossible, given fluctuations in the meanings of independent and dependent variables across time, contexts, and populations. The point here is that the temptations of confirmation bias exist for both direct and conceptual replications, and this potential should not be used as an argument in favor of direct and in opposition to conceptual replication.

Another criticism of conceptual replications is that, when they fail, it is too easy to attribute the failure to an undetected moderator. Of course, this is a problem for failed direct replications, as well, and has been a common and controversial response to such failures. The temptation to explain failed conceptual replications in this way is even stronger, given the many potential differences in the operationalizations of independent and dependent variables found among conceptual replications of the same hypothesis. These differences provide ready-made candidates for undetected moderators. In our view, in the case of both direct and conceptual replications, claims of undetected moderators must be supported by additional research that directly tests those moderators. Indeed, replication failures can offer important opportunities for theoretical advance when potential moderators can be identified and tested (for a compelling example from chemistry, see Retraction Watch, 2015).

Due to the increased ambiguity surrounding failed conceptual replications, at this point, such results are much less likely to be published than are failed direct replications. Much progress has been made in the creation of mechanisms through which failed direct replications may be published. In order to extract the most value from conceptual replication, a similar movement is needed to encourage the publication of failed conceptual replications. The publication of such results is critical if we wish to avoid "Conceptual Type I error"—the adoption of false theories. And, as above, the publication of failed conceptual replications is critical for identifying important variables that moderate the conditions under which a theory is likely to hold. If we remain unaware of the various attempts at conceptual replication, identifying these moderators is very difficult.

If we are to publish failed conceptual replications, then we must be attentive to the ambiguities surrounding those failures. If it is too easy

to dismiss the findings due to flawed independent or dependent variables, then their informational value is reduced. To counter this problem, conceptual replications should include careful pilot testing of variables to ensure that they are manipulating/measuring the constructs that they are supposed to. In addition, conceptual replications should include robust manipulation checks. These practices reduce the degrees of freedom available to dismiss failed conceptual replications.

4.10. Envoi: Replication in the context of human scientific practice

We are not arguing to trade higher confidence in a single set of operations for lower confidence in multiple operations. We are arguing to trade higher confidence in a single set of operations for higher confidence in theory. Some scientists want—and need—exact replications to generate confidence in their experiments. These scientists should conduct them, they should recommend publication of them in the review process, and they should read, cite, and enjoy them. But these same scientists should not require the same behaviors of others, make a narrow type of scientific practice necessary for publication, nor expect all others to share their values. We recommend the same tolerance for those who prefer conceptual replications. A lack of dissent and diversity in a scientific community is a sure prescription for lethargic progress.

Science has been remarkably successful in the face of diversity of interests, and can tolerate modest levels of noise among the signal in social psychological research. As an institution that organizes human behavior, it is remarkably robust to such threats as modest levels of p-hacking, publication bias, and false positive papers.

> *Flawed people, working in complex social environments, moved by all kinds of interest, have collectively achieved a vision of parts of nature that is broadly progressive and that rests on argument meeting standards that have been refined and improved over centuries.* (Kitcher, 1993, p. 390).

If scientific progress can be conceptualized as an evolutionary process (Campbell, 1974), then variety is necessary for proper selection. As Hull (1988) notes, "selection processes, though they are highly effective, are also extremely inefficient. Selection requires waste, and if biological evolution is any sign, a great deal of waste. How efficient can science be made without decreasing its rate of conceptual growth?" (p. 521). The rate of scientific progress requires material to select from—better and more effective ideas and theories should persist over incoherent and ineffective ideas. To maximize progress means tolerating—even celebrating—innovate theories and operationalizations, knowing full well that many of them will be ineffective or wrong. If the fastest and most effective way to generate effective progress is through selection, then providing alternative views, multiple operations, and conceptual replications will speed scientific progress at a faster rate than exact replications.

References

Bem, S. L. (1981). Androgyny and gender schema theory: A conceptual and empirical integration. *Nebraska Symposium on Motivation, vol. 32*. (pp. 179–226). Lincoln: University of Nebraska Press.

Brehm, J. (1956). Postdecision changes in the desirability of alternatives. *Journal of Abnormal and Social Psychology, 52*, 384–389.

Bridgman, P. (1927). *The logic of modern physics*. New York: Macmillan.

Brown, R. (1965). *Social psychology*. New York: Free Press.

Campbell, D. T. (1974). Evolutionary epistemology. In P. A. Schilpp (Ed.), *The Philosophy of Karl Popper, Book 1* (pp. 413–463). LaSalle, IL: Open Court Publishing.

Casadevall, A., & Fang, F. C. (2010). Reproducible science. *Infection and Immunity, 78*, 4972–4975.

Chen, M. K., & Risen, J. L. (2010). How choice affects and reflects preferences: revisiting the free-choice paradigm. *Journal of Personality and Social Psychology, 99*, 573–594.

Crandall, C. (2001). Scientific progress: A need for trust, a need for skepticism. *Dialogue, 16*(2), 20–21 (Available for download at: http://www.academia.edu/2705937/Scientific_Progress_A_need_for_trust_a_need_for_skepticism).

Crowe, E., & Higgins, E. T. (1997). Regulatory focus and strategic inclinations: Promotion and prevention in decision-making. *Organizational Behavior and Human Decision Processes, 69*, 117–132.

Drummond, C. (2009). Replicability is not reproducibility: Nor is it good science. *Proc. Eval. Methods Mach. Learn. Workshop 26th ICML, Montreal, Quebec, Canada* (Downloaded from http://www.csi.uottawa.ca/~cdrummon/pubs/ICMLws09.pdf, Mar 25, 2015).

Duhem, P. (1906/1954). *La théorie physique. Son objet, sa structure [translated as The aim and structure of physical theory]*. Princeton, NJ: Princeton University Press.

Egan, L. C., Santos, L. R., & Bloom, P. (2007). The origins of cognitive dissonance: Evidence from children and monkeys. *Psychological Science, 18*, 978–983.

Eidelman, S., Crandall, C. S., Goodman, J. A., & Blanchar, J. C. (2012). Low-effort thought promotes political conservatism. *Personality and Social Psychology Bulletin, 38*(6), 808–820.

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), e109019.

Fuller, S. (1997). *Science*. Minneapolis: University of Minnesota Press.

Funder, D. C., Levine, J. M., Mackie, D. M., Morf, C. C., Sansone, C., Vazire, S., & West, S. G. (2014). Improving the dependability of research in personality and social psychology: Recommendations for research and educational practice. *Personality and Social Psychology Review, 18*, 3–12.

Godfrey-Smith, P. (2009). *Theory and reality: An introduction to the philosophy of science*. Chicago: University of Chicago Press.

Greenwald, A. G. (2015). *The extremely limited value of isolated failures to replicate. In "What do direct replications say about the reproducibility of psychological phenomena?"*. New York: Symposium at Association for Psychological Science (May 22, 2015).

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology, 13*(3), e1002106. http://dx.doi.org/10.1371/journal.pbio.1002106.

Henry, P. F., & Sears, D. O. (2002). The symbolic racism 2000 scale. *Political Psychology, 23*, 253–283.

Hull, D. L. (1988). *Science as a process: An evolutionary account of the social and conceptual development of science*. Chicago: University of Chicago Press.

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

John, L., Loewenstein, G. F., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological Science*. http://dx.doi.org/10.2139/ssrn.1996631.

Kicinski, M. (2013). Publication bias in recent meta-analyses. *PloS One, 8*(11), e81823.

Kitcher, P. (1995). *The advancement of science: Science without legend, objectivity without illusions*. New York: Oxford University Press.

Lakatos, I., & Musgrave, A. (1970). *Criticism and the growth of knowledge*. Cambridge: Cambridge University Press.

Laudan, L. (1978). *Progress and its problems: Towards a theory of scientific growth, vol. 282*, Berkeley, CA: University of California Press.

MacKay, R. J., & Oldford, R. W. (2000). Scientific method, statistical method and the speed of light. *Statistical Science, 15*, 254–278.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9*, 147–163.

McGrath, J. E. (1981). Dilemmatics: The study of research choices and dilemmas. *American Behavioral Scientist, 25*, 179–210.

Meehl, P. E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*, 103–115.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1*, 108–141.

Meyer, M. N., & Chabris, C. (2014). Why psychologists' food fight matters: Important findings haven't been replicated, and science may have to change its ways. *Slate.com*. http://www.slate.com/articles/health_and_science/science/2014/07/replication_controversy_in_psychology_bullying_file_drawer_effect_blog_posts.html

Moscovici, S., & Zavalloni, M. (1969). The group as a polarizer of attitudes. *Journal of Personality and Social Psychology, 12*, 125–135.

Neuberg, S. L., & Newsom, J. T. (1993). Personal need for structure: Individual differences in the desire for simpler structure. *Journal of Personality and Social Psychology, 65*, 113–131.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*, 615–631.

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science, 7*(6), 657–660.

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

Pashler, H., & Harris, C. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7*, 531–536.

Pashler, H., & Wagenmakers, E. (2012). Special section on replicability in psychological science: A crisis of confidence? (special section). *Perspectives on Psychological Science, 7*, 528–654.

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

Popper, K. R. (1963/2002). *Conjectures and refutations: The growth of scientific knowledge*. London: Routledge.

Quine, W. V. O. (1980). *From a logical point of view: Nine logico-philosophical essays Cambridge*. MA: Harvard University Press.

Ramscar, M. (2015, Aug. 15). The unspeakable in the pursuit of the unrepeatable (web log post). (Retrieved from) https://ramscar.wordpress.com/2015/08/05/the-unspeakable-in-pursuit-of-the-unrepeatable/#more-704

Retraction Watch (2015). 50 years later, is it time to retract a retraction by a Nobel prize-winning author? Downloaded from http://retractionwatch.com/2015/09/25/five-decades-later-is-it-time-to-retract-a-nobelists-retraction/ (September 25, 2015)
Roberts, B. W. (2014). The Deathly Hallows of psychological science. Downloaded October 31, 2014 from http://pigee.wordpress.com/2014/03/10/the-deathly-hallows-of-psychological-science/
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.
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.
Sniderman, P. M., Piazza, T., Tetlock, P. E., & Kendrick, A. (1991). The new racism. *American Journal of Political Science*, *35*, 423–447.
Soter, S., & deGrasse Tyson, N. (2000). *Cosmic horizons: Astronomy at the cutting edge*. New York: American Museum of Natural History.
Spence, J. T. (1993). Gender-related traits and gender ideology: Evidence for a multifactorial theory. *Journal of Personality and Social Psychology*, *64*, 624–635.
Stanley, D. J., & Spence, J. R. (2014). Expectations for replications are yours realistic? *Perspectives on Psychological Science*, *9*, 305–318.
Stoner, J. A. (1961). A comparison of individual and group decision involving risk. Unpublished master's thesis, Massachusetts Institute of Technology, School of Industrial Relations.
Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives in Psychological Science*, *8*, 59–71.
Van Berkel, L., Crandall, C. S., Eidelman, S., & Blanchar, J. C. (2015). Hierarchy, dominance, and deliberation egalitarian values require mental effort. *Personality and Social Psychology Bulletin*, *41*, 1207–1222.
Webster, D. M., & Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *Journal of Personality and Social Psychology*, *67*, 1049–1067.