INTERNATIONAL REVIEW OF SOCIAL PSYCHOLOGY

Świątkowski, W. and Dompnier, B. (2017). Replicability Crisis in Social Psychology: Looking at the Past to Find New Pathways for the Future. *International Review of Social Psychology*, 30(1), 111–124, DOI: https://doi.org/10.5334/irsp.66

RESEARCH ARTICLE

Replicability Crisis in Social Psychology: Looking at the Past to Find New Pathways for the Future

Wojciech Świątkowski and Benoît Dompnier

Over the last few years, psychology researchers have become increasingly preoccupied with the question of whether findings from psychological studies are generally replicable. The debates have originated from some unfortunate events of scientific misconduct in the field, and they have reached a climax with the recent discovery of a relatively weak rate of replicability of published literature, leading to the so-called replicability crisis in psychology.

The present paper is concerned with examining the issue of replicability in the field of social psychology. We begin by drawing a state of the art of the crisis in this field. We then highlight some possible causes for the crisis, discussing topics of statistical power, questionable research practices, publication standards, and hidden auxiliary assumptions of context-dependency of social psychological theories. Finally, we argue that given the absence of absolute falsification in science, social psychology could greatly benefit from adopting McGuire's perspectivist approach to knowledge construction.

Keywords: Perspectivism; Publication standards; Questionable research practices; Replicability crisis; Social psychology; Statistical power

What does it mean to do *good* research? How do we define criteria of research quality? Arguably, the answer may be twofold. On the one hand, one can define research quality from a purely scientific and methodological standpoint, with reference to validity of drawn conclusions and contribution to existing knowledge. Here, the standard of quality is defined with reference to internal, external, and construct validity (cf. Brewer & Crano, 2014) associated with studied psychological effects. On the other hand, one can also adopt a more evaluative perspective and judge the quality of a given piece of research based on the reputation of the journal in which it was published. The assumption underlying this point of view is that the more prestigious the outlet, the better the research is considered. The goal of empirical science is to discover truths about the world using scientific methodology, the quality of which can be assessed solely with reference to the first definition. In practice, however, many of the decisions that guide scientists' careers in academia (e.g., grant attributions, hiring process) are based on the second standard of research quality. This twofold definition of research quality creates incentives that are often conflicting for researchers: striving for research validity on one hand, and research publishability on the other (Nosek, Spies, & Motyl, 2012).

Department of Social Psychology, University of Lausanne, CH

Corresponding authors: Wojciech Świątkowski (wojciech.swiatkowski@unil.ch), Benoît Dompnier (benoit.dompnier@unil.ch)

The reality is that the better a journal's reputation – based on indices such as the journal impact factor (Garfield, 2006) – the harsher the competition to publish in this journal becomes. The spirit of competitiveness could somehow relate to the quality of research being published in a positive way, for instance, by favoring papers of the highest scientific validity. However, strong competition associated with the publication process may also promote the search for "perfect data" (Giner-Sorolla, 2012), thereby encouraging a variety of questionable practices (John, Loewenstein, & Prelec, 2012; Kerr, 1998; Schimmack, 2012; Simmons, Nelson, & Simonsohn, 2011) that ultimately cast doubts on the validity of what is being published and the reproducibility of such findings.

For the last few years, psychology researchers have become increasingly preoccupied with the question of whether the findings that are typically published in the literature are replicable. As a matter of fact, two special issues in high-level journals have recently been devoted to discussing this so far neglected topic (Pashler & Wagenmakers, 2012; Stangor & Lemay Jr., 2016). Undoubtedly, this heightened interest follows the recent debates that called into question conventional practices of research conduct and data analysis (e.g., Ioannidis, 2005). Especially impactful was the recent collaborative Psychology Reproducibility Project led by Brian Nosek (Open Science Collaboration, 2015), which indicated that many – if not the majority of – published findings in psychology are indeed not replicable.

The present paper is concerned with the so-called replicability crisis in psychology that originated over the last few years, with a focus on social psychology. To begin with, we will present a state of the art of the current crisis in replicability and confidence in the field. Next, we will review causes that may have produced such a phenomenon. Specifically, we will first discuss the issue of low statistical power in psychological studies. Then, we will cover matters related to questionable research practices and current publication standards. Finally, we will address the issue of hidden assumptions of context-dependency of social psychological effects and discuss the potential of perspectivism – an epistemological approach to scientific discovery developed by William McGuire (1983, 1989, 1999, 2004) – on social psychology.

Social Psychology: A Field in Confidence Crisis

About five years ago, social psychology entered a cycle of unfortunate events that considerably undermined both the reputation of the field and the confidence in the knowledge it produced. Undoubtedly, the first event to mention was the discovery of high-profile cases of outright fraud (Crocker & Cooper, 2011), such as Diederik Stapel's case. This scholar admitted to having intentionally fabricated data over a dozen years, which eventually resulted in retracting more than 50 published papers. At the time, this incident could only have added some extra fuel to the already burning fire that was caused by the publication of some surprising findings in one of the top outlets in social psychology. Bem (2011) published an article presenting nine experimental studies in which he made a controversial claim about the existence of extrasensory perception, according to which individuals could experience a sense of future events before they would even occur. This article had been considered as conforming to scientific standards (Judd & Gawronski, 2011). Such an attitude was interpreted by many as the sign that these very standards are flawed at the outset and therefore "allow" the publication of too many findings that are actually false positives (e.g., Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011). Ultimately, the original effect failed to be replicated by other researchers (Galak, LeBoeuf, Nelson, & Simmons, 2012). Yet, this episode raised serious questions about the way researchers in psychology analyze and report their data (e.g., Simmons et al., 2011).

Furthermore, John et al. (2012) assessed the prevalence of *questionable research practices* among academic psychologists and found that the occurrence of these were quite high. For instance, among the 2,155 respondents to their survey, almost 56 percent admitted to having decided to collect more data after seeing that the initial test was not statistically significant, and nearly 46 percent admitted to having selectively reported studies that "worked" in a paper to be published. Last but not least, methodologists have also highlighted a prevalent issue of misreporting statistical analyses in published articles (Bakker & Wicherts, 2011; Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2015). For instance, Nuijten et al. (2015) reported that in a sample of more than 30,000 articles selected from the top eight psychology journals, one in eight possibly contained an inconsistent *p*-value that might have affected the statistical conclusion.

Taken as a whole, one can wonder to what extent the conclusions published in the psychological literature are scientifically valid and sound. Given that one of the key features that defines the scientific quality of a proposed claim is reproducibility, the effects that are published should be more than likely replicable. Until recently, replicating existing studies has been no more than just a rare phenomenon in psychological science (Schmidt, 2009). Indeed, Makel, Plucker, and Hegarty (2012) estimated the rate of replicated studies in the field to be as low as 1.07 percent.

Recently, however, a substantial collaborative research effort – the Psychology Reproducibility Project – was deployed to assess the extent to which one could replicate psychological effects from the published literature (Open Science Collaboration, 2015). Independent research teams attempted to replicate 100 published effects from various fields of psychology. The results seemed to be somewhat in line with what one might have expected based on the above review: Only 39 percent of findings were considered to be successfully replicated. Regarding the effects specifically from the field of social psychology that were under scrutiny, only 25 percent were replicated. Likewise, other widely known effects from the social psychological literature, such as ego-depletion (Baumeister, Bratslavsky, Muraven, & Tice, 1998) or unconscious behavioral priming (Bargh, Chen, & Burrows, 1996) were challenged in replication studies that proved unsuccessful (Doyen, Klein, Pichon, & Cleeremans, 2012; Hagger et al., 2016).

While this state of affairs is certainly not one any scientist would like to identify with, it should be viewed as an opportunity and as a strong incentive to reconsider some of the current practices that might lie at the origin of this crisis. In the sections below, we review practices that we deem to be the most relevant to the following question: why do so many published studies in psychology fail to be replicated? Our ambition here is not to be exhaustive; our overview will rather aim to highlight some credible answers.

The Important (Neglect of) Statistical Power

The first candidate answer to the question of why a replication study would not succeed in finding an original effect is statistical power, which refers to the probability of correctly rejecting a tested hypothesis (in current practices, H_0). In other words, statistical power determines the chance when performing a test to declare an effect as "statistically significant" if the tested effect does genuinely exist. High statistical power is therefore a desirable property one should seek to achieve and should be of greatest concern when designing a study. If a study is to address a research question appropriately, the minimal requirement is that it should have the means to detect a supposedly existing psychological effect. Stated otherwise, running a study whose outcome is very likely to be negative at the outset – even though the effect under investigation truly exists – is clearly pointless and represents a waste of valuable research resources.

Power can be formally defined with reference to Type-II error, which is the decision of not rejecting the null hypothesis H_0 when it is actually false. If β refers to the long-term frequency of committing the Type-II error, then statistical power is defined as $1 - \beta$. Following widely accepted recommendations (e.g., Cohen, 1965; Maxwell, 2004), $1 - \beta$ should be at least as high as 0.80. With the power set at 0.80, running five studies assessing a true effect will on average yield four statistically significant results. Alternatively, statistical power can also be considered as a long-term *p*-value distribution (Cumming, 2012). If $1 - \beta = 0.80$, it then means that if one were to run 100 replication studies of the exact same true effect, approximately 80 percent of these studies would yield a *p*-value significant at 0.05 level, and approximately 20 percent would yield a *p*-value comprised between 0.05 and 1 (see Cumming, 2012, p. 323). As statistical power is a function of the alpha level (long-term frequency of committing the Type-I error of rejecting H_0 when it is true), sample size, and effect size (Keppel, 1991), it is possible to derive one when the other three are fixed. To increase power, one can increase the study's sample size.

Methodologists have long urged scientists to work systematically with substantial sample sizes to draw inferences from highly powered studies. These detect more reliable and stable effects and yield fewer false positives than those with lesser power (Cohen, 1962, 1990, 1992; Maxwell, 2004; Maxwell, Kelley, & Rausch, 2008). In other words, high-powered studies hold higher informational value than low-powered studies (see Lakens & Evers, 2014, for a review). All other things being equal, a significant effect from a high-powered study has a lower chance of being a false-positive finding than if it was obtained with a low-powered study. Likewise, a non-significant effect from a high-powered study has a lower chance to be a false-negative than if it was obtained with lower power.

Power is also particularly important in studies where measurement error is to be expected. Low-powered studies that assess "noisy effects" (i.e., where there are systematic random variations associated with the way variables are measured) can often yield overestimated effect sizes, whose achieved statistical significance capitalizes on random error rather than on the true parameter value (Loken & Gelman, 2017). As it is reasonable to assume measurement error to be present in psychological research, this should give an extra incentive to consider power issues so as to limit the proportion of published studies whose effect sizes are spuriously inflated. It is therefore recommended to perform a priori power analyses indicating the minimum sample size necessary to achieve a desirable level of statistical power when designing a study.[1]

Unfortunately, despite its crucial scientific importance and numerous efforts by leading methodologists in the field, statistical power has often been neglected and underused in social psychological science, leading to many low-powered studies whose chances of detecting an effect have been lower than one could get by literally flipping a coin (Cohen, 1990). Indeed, for years, systematic literature reviews have pointed out the issue of low statistical power across studies in psychology (Bakker, van Dijk, & Wicherts, 2012; Maxwell, 2004; Rossi, 1990; Sedlmeier & Gigerenzer, 1989). Most likely, this state of affairs is due to the fact that researchers use sample sizes that are too small given the small effect sizes typical to the field (see Richard, Bond Jr., & Stokes-Zoota, 2003, for a review). Thus, with regard to the issue of the prevalence of unsuccessful replication studies, it is possible that some of these simply lacked the statistical power needed to properly detect the original effects (Maxwell, 2004; Tressoldi, 2012). For instance, Tressoldi (2012) noted that out of six recent meta-analyses aimed at assessing the validity of controversial effects, four meta-analyses included studies with low and very low statistical power, ranging from 0.07 to 0.55 levels.

However, as much as the statistical power of replication studies is relevant for this issue, it cannot solely account for the low rate of replicability in social psychology in general. For instance, the mean power achieved in the Psychology Reproducibility Project (Open Science Collaboration, 2015) was estimated to be 0.92, and the researchers expected to detect around 89 effects as statistically significant based on the 100 they tested, assuming these effects were true. This prediction scores far beyond the 25 percent of actually replicated findings, which indicates that the statistical power of the replication studies is clearly not the only issue at stake. Yet, what about the original studies? As stated above, social psychological studies are frequently underpowered given the small effect sizes observed in the field. This likely inflates the rate of false-positive findings in the published literature that are later unreplicable (Forstmeier, Wagenmakers, & Parker, 2016; Ioannidis, 2005; Lakens & Evers, 2014).

In line with this argument, there seems to be an apparent inconsistency between the actual power typically achieved in psychological studies and their capacity to detect statistically significant results (Francis, 2012; Schimmack, 2012). Specifically, Schimmack (2012) observed a substantial gap between the number of reported significant results in multiple study articles and their respective level of power. Indeed, testing several hypotheses in such research programs typically involves performing a high number of statistical tests. As noted by Maxwell (2004), increasing the number of statistical tests to be conducted decreases the overall probability of declaring all outcomes from those tests as statistically significant (i.e., their overall statistical power), even though the probability of finding at least one significant result increases. Consequently, many such published multi-study articles should theoretically yield fewer significant results, and the reported statistical tests should be consistent with their respective power level. In the following section, we will provide some explanations for this discrepancy, which will further account for the low rate of replicability among psychological studies.

Questionable Research Practices Cause Questionable Research Conclusions

Although even highly powered studies do not demonstrate the absence of an effect when they fail the attempt of replication, they do raise the possibility that the effect under investigation might actually be a false

positive. Indeed, many research practices that have been commonly deployed by the scientific community might be at the origin of false positives in the psychological literature. Practices that directly cause the Type-I error rate to inflate include the *researcher's degrees of freedom* (Simmons et al., 2011), that is, all kinds of unjustifiable flexibility in data analysis, such as working with several undisclosed dependent variables, collecting more observations after initial hypothesis testing, stopping data collection earlier than planned because of a statistically significant predicted finding, controlling for gender effects a posteriori, dropping experimental conditions, and so on. Likewise, other procedures known as *p-hacking* (undisclosed multiple testing without adjustments) and *cherry picking* (dropping observations to reach a significance level) lead to the same problematic consequences: using such techniques when analyzing data increases the Type-I long-term error rate, especially when applied in combination. For instance, combining three of the aforementioned practices could inflate the alpha level to as high as 60 percent (see Simmons et al., 2011)! In practical terms, this means that more than half of the findings declared as "statistically significant" could be merely false positives. Simmons et al. (2011) illustrated the danger of such procedures by showing that one could literally find evidence for any claim, no matter how absurd, when using these questionable analytical practices, such as the finding that listening to certain types of songs could change people's actual age. To the extent that the *p*-value represents the long-term rate of false decisions of rejecting the null hypothesis H_0, any procedure that changes this error rate should be adjusted for, if one wishes to keep the alpha level at the desired, conventional level of 0.05 (Wagenmakers, 2007).

Based on recent reports, the high prevalence of these questionable research practices among researchers could explain the abundance of false positives in the social psychological literature (John et al., 2012). Unfortunately, the problem of a high rate of false positives in published articles is further galvanized by what Kerr (1998) described as "HARKing" (Hypothesizing after the results are known), a practice that casts serious doubt on the validity of published findings. HARKing involves presenting any kind of post-hoc hypotheses in the introduction of a published article as if they were a priori from the beginning. At the time, Kerr (1998) reported that HARKing was a commonly used strategy to increase the publishability of one's findings. Although from the standpoint of current research standards HARKing is considered as an unacceptable and condemnable practice, it should be noted that it was once explicitly encouraged. Most notably, in a practical guide for researchers, Bem (2003) advised that

> the data may be strong enough to justify recentering your article around the new findings and subordinating or even ignoring your original hypotheses [. . .]. If your results suggest a compelling framework for their presentation, adopt it and make the most instructive finding your centerpiece. (pp. 187–188)

Obviously, there is nothing wrong with conducting exploratory research per se, which should in fact occupy an integral and important part in a research program for purposes such as hypothesis generation or testing auxiliary theories (McGuire, 1997). What is actually harmful, scientifically speaking, is disguising exploratory and other unexpected findings as confirmatory results. Performing extensive exploratory data analysis in a search for interesting yet unpredicted findings always presents the inherent risk of capitalizing on random sampling chance and consequently detecting a statistically significant false positive, especially when researchers' degrees of freedom or other kinds of questionable research practices are applied. Once such exploratory effects are detected – regardless of whether they are actually false positives – they are easily transformed into post-hoc hypotheses given the strength of the hindsight bias that allows one to always find a plausible rationale for almost any hypothesis (Fischhoff, 1975). In the worst-case scenario, HARKing causes a false positive to become a new (yet wrong from the outset) theory, which ultimately undermines the quality of produced science (see Ferguson & Heene, 2012, for a review). Given the relative high prevalence of HARKing among researchers (John et al., 2012; Kerr, 1998), it is likely that many supposedly "unsuccessful" studies were actually devoted to replicating invalid, yet presented as confirmatory, effects. Ideally, when an interesting, but unexpected, finding emerges from a study, it should be addressed in a follow-up confirmatory one.

The current debate on research practices might implicitly convey an indictment toward researchers and impute the responsibility for poor research quality on their shoulders, assuming that they purposely applied dubious procedures and hence sacrificed scientific validity over scientific publishability. We do not call into question the ethical responsibility any scientist should endorse with regard to the way he or she manages data. Yet, in our view, it is more reasonable to assume that the large-scale prevalence of questionable research practices among academic psychologists should be traced to a more structural cause. As a matter of fact, we suggest in the following sections that current publication standards actually promote these kinds of questionable research practices and thus may be at the origin of the replicability crisis in social psychological science.

Publication Standards Promote Bad Research Practices

Many journals have been sensitive to replication issues that came along with the crisis and have consequently developed new criteria as standards for publication. For instance, increasingly more journals require addressing issues of sample size and statistical power, and they have also become open to new approaches to data analysis (e.g., meta-analysis, Bayesian statistics). Still, some of the current publication standards encourage bad research practices in several ways. Arguably, the first to mention is the pervasive dominance of the null hypothesis significance testing (NHST) as the main tool of scientific inference in researchers' statistical toolbox, along with the sancti-

fied level of $p < 0.05$.[2] With regard to editorial practices, maintaining statistical significance at the $p < 0.05$ level as the *sine qua non* condition for publication has several deleterious consequences on the quality of published research.

First, it encourages the kind of dichotomous thinking in which the alpha level of 0.05 sets the boundary between the strict existence and non-existence of an effect. Cumming (2012, 2014) extensively criticized this reasoning by arguing that the consistency of the conclusions drawn from research is better when one focuses on continuous effect size estimation instead of on hypothesis testing (see Coulson, Healey, Fidler, & Cumming, 2010). This is because hypothesis testing involves making black-and-white decisions (i.e., an effect is statistically significant or not) that do not take into account the numerical value of a point estimate of the tested effect. For instance, two studies addressing the same research issue can make close estimations of an effect size, yet only one of them reaches the significance threshold. In this case, the perceived consistency of the studies would depend on the adopted analytical strategy.

Furthermore, the requirement of statistical significance fixed at the $p < 0.05$ level tends to undermine the evidential value of the published literature. On the one hand, recent advances in Bayesian statistics have pointed out that a *p*-value significant at 0.05 may often pertain to only weak evidence against the null hypothesis (Johnson, 2013; Wetzels et al., 2011). From this point of view, publishing studies associated with a *p*-value as "high" as 0.05 results in cumulating fairly poor and unreliable evidence for the generated claims, which could further account for the low replicability rate in psychology. A recent Bayesian re-interpretation of the Reproducibility Project (Etz & Vandekerckhove, 2016) backs up this claim. While the authors claimed that 75 percent of replication studies among those that they considered provided a qualitatively similar amount of evidence as the original studies, the evidence was often weak. Indeed, out of 72 re-analyzed original studies, 43 (60%) represented indecisive and ambiguous evidence for the published effects.

On the other hand, the overemphasis on $p < 0.05$ as a prerequisite for publication ultimately biases the estimated magnitude of published effects. This phenomenon is widely known as the "file drawer" problem (Fanelli, 2012; Ferguson & Heene, 2012; Rosenthal, 1979): Positive, statistically significant results get more easily published than negative, statistically non-significant results. To provide with an extreme, hypothetical version of this problem, Rosenthal somewhat cynically stated that "the journals are filled with the 5% of the studies that show Type I errors, while the file drawers back at the lab are filled with the 95% of the studies that show non-significant (e.g., $p > 0.05$) results". (1979, p. 638). This issue is a straightforward consequence of one of the most severe limitations of the NHST procedure and has to do with asymmetrical decision making based on the *p*-value. Indeed, a significant *p*-value leads to rejection of the null hypothesis, while a non-significant *p*-value does not lead to acceptance of the null hypothesis (Cohen, 1990, 1994), despite common but improper practices of doing so (Hoekstra, Finch, Kiers, & Johnson, 2006). A non-significant *p*-value is difficult to interpret, providing evidence neither against nor in favor of the null hypothesis. For this reason, authors, reviewers, and editors prefer focusing on supposedly more conclusive, positive, and statistically significant results. This ultimately results in the disappearance of negative results from the literature (Fanelli, 2012). Consequently, this bias creates a structural incentive to search for positive results. It should then come as no surprise that researchers might be motivated to use a variety of questionable research practices, further engaging in self-confirmatory biases and eventually convincing themselves that their findings are genuine, in a situation where research validity conflicts with research publishability.

The file drawer problem and the resulting publication bias also directly distort scientific knowledge. This is because whenever there is enough data available in the literature relative to an effect, it is a common strategy to use meta-analytic procedures to try to see a "clearer image" in the fog of sometimes conflicting evidence. However, if negative results are systematically omitted from the publication process, the "clearer image" must necessarily be distorted, especially when the omitted "negative" results are declared as such because of insufficient statistical power and are actually Type-II errors. One can only speculate about the extent of this phenomenon, but in view of the low statistical power in many psychological studies (e.g., Maxwell, 2004), it is reasonable to assume that this situation is fairly common. As a consequence, to the extent that considerably more positive findings are available in the literature than negative ones, parameter estimations from meta-analyses might often be spuriously inflated (Rosenthal, 1979). Again, the vicious circle continues as the worst-case scenario might occur when these artificially inflated parameter estimations serve for power analyses. The overestimation of effect size in power analyses can lead to underestimating the sample size needed to detect the effect of interest (leading to an underpowered study), which increases the likelihood of committing a Type-II error.

Furthermore, it should also be noted that many scientific practices that easily result in false-positive findings have often been encouraged by unrealistic standards of perfection for publication (Giner-Sorolla, 2012). Indeed, authors are too often demanded to present an almost perfect match between the theoretical predictions they test and the empirical evidence they find in their studies. It is essential to keep in mind that regardless of the validity of a theory, sampling variability will always sprinkle data with some random noise. Cooper's (2016) recent editorial made it clear: "Real data are messy, and imperfection (not perfection) should be expected" (p. 433).

Last but not least, current editorial standards overemphasize the need for novelty, which also impedes the construction of cumulative science. Indeed, top-ranked journals are more inclined to accept studies highlighting new and original psychological effects, with a serious cost of neglecting the importance of replications

(e.g., Neuliep & Crandall, 1993). The latter are still too often deemed as lacking prestige and inspiring little interest among researchers, and, therefore, until now, the studies have only been rarely conducted (Makel et al., 2012). Likewise, the excellence of publishing new effects as a result of confirmatory research tends to underestimate the role of conducting exploration in one's research program. This might further contribute to HARKing and other questionable strategies in data analysis.

Many concrete solutions have already been put forward to encourage researchers to adopt research practices that could help in building cumulative science and contribute to overcoming the crisis. Among these, Open Science Collaboration (2012, 2015) and the Many Labs Replication Project are collaborative projects that specifically focus on replicability issues. Likewise, the Open Science Framework helps researchers preregister their studies (openscienceframework.org), which could be a useful tool to clearly distinguish between exploratory and confirmatory findings. Besides these "large-scale" projects, good individual research practices are the *sine qua non* condition for the quality of psychological studies to increase. Among such practices, one should remember to take into account the importance of statistical power in study design (e.g., Maxwell, 2004; Maxwell et al., 2008). In this regard, we can think of Cohen's (1990) "less is more, except of course for sample size" (p. 1304). Furthermore, instead of relying solely on hypothesis testing along with the *p*-value, adopting the meta-analytical thinking with parameter estimation is more than desirable (e.g., Stukas & Cumming, 2014). As the American Statistical Association has recently acknowledged, "data analysis should not end with the calculation of a *p*-value when other approaches are appropriate and feasible" (Wasserstein & Lazar, 2016, p. 132). As such, Bayesian statistics seem to be an interesting tool that can provide the strength of evidence for one's theory (e.g., Mulder & Wagenmakers, 2016; Świątkowski, submitted) and corroborate the null hypothesis when of theoretical interest (e.g., Gallistel, 2009; Rouder, Speckman, Sun, Morey, & Iverson, 2009). Crucially, however, one should always keep in mind that no single statistical index can ever substitute for well-informed scientific thinking (Gigerenzer & Marewski, 2015; Wasserstein & Lazar, 2016).

Hidden Auxiliary Assumptions and Psychological Theories' Replicability

The last perspective on the replicability crisis we want to discuss, one that we often find surprisingly neglected in the debates on replicability issues – especially in fields such as social psychology – pertains to the generalizability of psychological theories and the external validity of studied effects. Within a purely confirmatory research paradigm, the demonstration of an effect involves assessing the validity of some predictions derived from the theory under investigation. In a nutshell, a theory can be defined as a set of logical propositions that posit causal relationships attempting to explain observable, naturally occurring phenomena (Fiske, 2004). These logical propositions are initially broad and abstract, but they give rise to more concrete and specific predictions that are empirically testable.

Supposedly, the whole point of doing science is deriving testable predictions from theories and consequently ruling out those that did not "work out" in line with the Popperian principle of refutability (Popper, 1959). Thus, if a theoretical conjecture predicts a specific observation, then, with respect to the *modus tollens* principle, empirically falsifying this relationship leads to falsifying the conjecture. This approach is often regarded as the golden standard of scientific inquiry and is indeed very popular among experimental social psychologists (Jost & Kruglanski, 2002). It enables making strong inference as to which of several competing theories should be rejected as being false and which deserve further investigation.[3]

In practice, however, whenever a theoretical conjecture is subjected to empirical scrutiny, some "hidden" assumptions are also implicitly tested. These involve auxiliary theories on the one hand (Lakatos, 1978; McGuire, 1983) and the empirical realization of specific conditions describing the experimental particulars on the other (Meehl, 1978). Consequently, failing to observe a predicted outcome does not necessarily mean that the theory itself is wrong, but rather that the conjunction of the theory and the underlying assumptions at hand are invalid (Lakatos, 1978; Meehl, 1978, 1990, 1997).

For instance, imagine a study testing the impact of competition on cognitive performance. Based on a given theory, a specific prediction is made that making participants solve a cognitive task (e.g., mathematical operations) in competitive settings will have a negative impact on their individual performance. Having eventually failed to observe such a relationship in the study, one might end up wondering whether this result falsifies the theory, or whether it may actually be imputed to the nature of the instrument assessing the subjects' performance (i.e., auxiliary theory N°1), to the kind of population from which the participants were drawn (i.e., auxiliary theory N°2), or to the quality of the induction of the experimental treatment (i.e., empirical realization of the specific condition), and so on (see Meehl, 1978).

Quite surprisingly, most social-psychological theories hardly ever make any explicit mention of the auxiliary assumptions that are present in the process of their elaboration (McGuire, 1983; see also Klein et al., 2012). Given the rationale outlined above, these assumptions could in fact be necessary to verify the predictions of a theory. This state of affairs might indeed seem surprising, insofar as social psychologists are best positioned to acknowledge the importance of context-dependency when studying human behavior. Arguably, context-dependency could actually play an active role in the "hidden assumptions" underlying many social-psychological theories. It is therefore reasonable to assume that context-dependency could also account for the unexpectedly low rate of replicability in the social-psychological literature (Open Science Collaboration, 2015; Van Bavel, Mende-Siedlecki, Brady, & Reinero, 2016a, 2016b; but see also Inbar, 2016). Thus, apart from considerations relative to statistical power and false positives, it is also likely that at least some replication

studies fail to replicate because the auxiliary assumptions regarding the context or the population-related specificities of original effects were simply not met.

Van Bavel et al. (2016b) put forward an argument giving some credit to this claim. The authors have recently pointed out that the replication rate discrepancy between social and cognitive psychology (25% vs. 53%) in the Open Science Collaboration's (2015) Reproducibility Project could be attributed to a greater context-sensitivity of the former over the latter, rather than to other methodological factors (e.g., effect sizes, sample sizes). While this argument does not put the validity of social-psychological theories in jeopardy *stricto sensu*, it highlights the sour fact that many of them may just not be as universal as we (social psychologists) would like them to be. Likewise, Sears (1986) had already emphasized long ago the fact that many studies in social psychology are conducted with college students, which may limit the extent to which the conclusions drawn are generalizable across other populations. Other widely reported cross-cultural differences (e.g., Hofstede & Hofstede, 2001; Markus & Kitayama, 1991; Nisbett, Peng, Choi, & Norenzayan, 2001) could also limit the universality and ecological validity of social-psychological theories.

As a matter of fact, the conflict between the search for psychological universals and the social constructivist approach that circumscribes human behavior to its historical and cultural context already has a long-standing tradition in our field (see Jost & Kruglanski, 2002). In some way, current debates on replicability issues related to context-dependency echo the intellectual crisis from the 1970s that called into question the epistemological foundations of the mainstream experimental social psychology (e.g., Gergen, 1973). Once again, we are faced with the same dilemma of whether the goal of establishing universal laws that experimental social psychology strives to pursue (Norenzayan & Heine, 2005) is ultimately attainable, given that the object of study – human beings – is necessarily situated in a particular context.

Bringing Social Psychology Back to its Name: Toward a Perspectivist Social Psychology

At this stage, one may wonder about the objective of performing replication studies – or any study at all, actually(!) – since they can never lead to rejecting the tested hypothesis. This is because a failure to observe the predicted outcome can always give rise to some "ad hoc explanations" (Meehl, 1967, p. 114) pertaining to hidden assumptions that moderate the validity of the effects under scrutiny. Indeed, the popperian falsifiability requirement is hardly fulfilled in a field where observing a theory-consistent outcome corroborates the theory, but where it is a priori known that observing a theory-inconsistent outcome will not imply theory rejection (Meehl, 1978). The ambiguous nature of a non-significant *p*-value, based on which psychologists predominantly make their inferences, may be at least partly responsible for this problem. A non-significant *p*-value provides neither evidence for nor against the null hypothesis, and even when the latter is true, the *p*-value can take almost any value (Cumming, 2012; Dienes, 2011).[4] Thus, even when a theory lacks empirical support and when failed attempts at replication accumulate, it is still hardly possible to properly operate the falsification mechanism, and this without even mentioning the fact that publication bias may prevent failed replications to be published (Ferguson & Heene, 2012).

Clear criteria about the most crucial aspect in theory construction – falsifiability – are indeed not consensual in social psychology (e.g., Trafimow, 2009; Wallach & Wallach, 2010). In the long run, the lack of a viable falsification procedure seriously undermines the quality of scientific knowledge psychology produces. Without a way to build a cumulative net of well-tested theories and to abandon those that are false, social psychology risks ending up with a confused mixture of both instead. With regard to this matter, Meehl (1978) had already emphasized that

> in the developed sciences, theories tend either to become widely accepted and built into the larger edifice of well-tested human knowledge or else they suffer destruction in the face of recalcitrant facts and are abandoned, perhaps regretfully as a "nice try". But in fields like personology and social psychology, this seems not to happen. There is a period of enthusiasm about a new theory, a period of attempted application to several fact domains, a period of disillusionment as the negative data come in, a growing bafflement about inconsistent and unreplicable empirical results, multiple resort to ad hoc excuses, and then finally people just sort of lose interest in the thing and pursue other endeavors. (p.807)

Before the temptation gets too strong to say that psychological science is in the end a pointless endeavor – as it has apparently no ultimate tool for distinguishing "good" from "bad" theories – one may wish to consider a possible response to this paradox that was put forward by William McGuire. In his contextualist approach to knowledge construction (McGuire, 1983), later labeled as *perspectivism* (McGuire, 1989, 1999, 2004), the problem of auxiliary assumptions in empirical assessment of theoretical conjectures is handled by acknowledging their integral role in the process of theory elaboration. Here, the "hidden" auxiliary assumptions that condition the extent to which a theory is generalizable are not considered as a problem, but rather as a means in the "discovery process to make clear the meaning of the [theory], disclosing its hidden assumptions and thus clarifying circumstances under which the [theory] is true and those under which it is false" (McGuire, 1983, p. 7). McGuire's perspectivist epistemology is strongly rooted in logical empiricism[5] as it builds upon some of its core postulates, but it also radically departs from others.

On the a priori side, perspectivism agrees with logical empiricism that researchers should conduct their research based on theoretically embedded hypotheses that guide their observations and organize data (McGuire, 1999). However, the perspectivist innovation stresses the

importance of accounting the same hypothesis by multiple theories and also formulating contrary hypotheses along with the theories from which they can be derived. This innovation stems from McGuire's (1983, 1999) basic assumption that every form of knowledge must necessarily be an incomplete picture of reality because its very nature of representation cannot completely account for the complexity of environment. Therefore, one single phenomenon could always be accounted for by multiple hypotheses driven from multiple theories, even contradictory ones. On the a posteriori side, perspectivism follows logical empiricism's premise regarding the importance of confronting theoretical predictions with empirical reality. However, perspectivism goes further by considering empirical assessment not only as test of whether the initial hypothesis is true, but also as a tool in strategic research planning (McGuire, 1999).

On the other hand, perspectivism diverges from logical empiricism in two crucial instances. First, instead of traditionally assuming that some theories are true and others false, perspectivism holds that all theories are *both* true and false, depending precisely on the conjunction that is met between the theory and the underlying auxiliary assumptions that are involved in empirical testing (McGuire, 1983). This position radically departs from the commonly endorsed point of view in which a researcher seeks to establish theoretical universality by managing to fail all possible attempts of theory falsification. Perspectivism rather asserts that identifying and making clear the "hidden" auxiliary assumptions that circumscribe a theory in a more explicit context of validity and generalizability contributes to theory understanding and full appreciation of its richness.

The second peculiarity of the perspectivist approach that follows is continuity – and not contrast – between the exploratory stage (i.e., hypothesis generation) and the confirmatory stage (i.e., empirical assessment) within a research program (McGuire, 1983). Conventionally, hypotheses are generated at an exploratory step of a research program to later be put under scrutiny and an empirical falsification, hence yielding a black-or-white outcome. Instead, perspectivism asserts that the confrontation between a hypothesis and empirical assessment "is not so much a testing of the hypothesis as it is a continuing revelation of its full meaning made apparent by its pattern of confirmations and disconfirmations in a strategically programmed set of observable situations" (McGuire, 1983, p. 14). Thus, the hypothesis-testing phase also contributes to theory construction by providing empirical insight through exploratory research. The perspectivist approach emphasizes the fact that exploration and confirmation should be equally important: One should follow the other within the same research program so as to make the process of scientific discovery more circular. The following section provides an illustration on how perspectivism can give rise to systematic strategic planning and help develop a positive research agenda.

At the conceptual level, building a perspectivist research program (McGuire, 1983, 1989, 2004) starts with the formulation of an initial hypothesis about a possible relationship between two variables of interest. This initial hypothetical insight is then explored using a stepwise procedure. As a first step, one can explore the meaning of the variables, for instance, by using word games and synonyms, to identify alternative labels as well as several possible working definitions for each of them. Whereas the initial hypothesis relies to some extent on intuitive choices, this verbal exploration enables the researcher to define consciously and precisely the nature of the variables under investigation among a set of possible alternatives.

Once the appropriate labels and working definitions are selected, the relationship between the two variables is expressed through various modalities (e.g., verbal, graphical, statistical) that allow the researcher to think about the relationship without relying on a unique and too restrictive representation. Then, multiple explanations for the relationship are formulated to create several mediational hypotheses, assuming that only one process cannot fully explain it. In addition, several moderators can be identified with the goal of defining the boundaries of the relationship and those of its explanations.

Finally, since perspectivism assumes that any hypothesis is true within some specific contexts, the hypothesis contrary to the initial one also has to be explored by searching for several possible mediators and moderators that could apply to it. McGuire (1997) developed a set of 49 creative heuristics to help researchers go through this step of planning and developing original hypotheses for their research.

Taken as a whole, this procedure enables the researcher to create a set of well-defined hypotheses that predict the conditions in which, as well as the processes through which, the initial hypothesis and its contrary are expected to be valid. **Figure 1** illustrates the general structure of a theoretical model based on these guidelines.

Even if endorsing perspectivism in theory construction may appear at first glance as a potential source of confusion

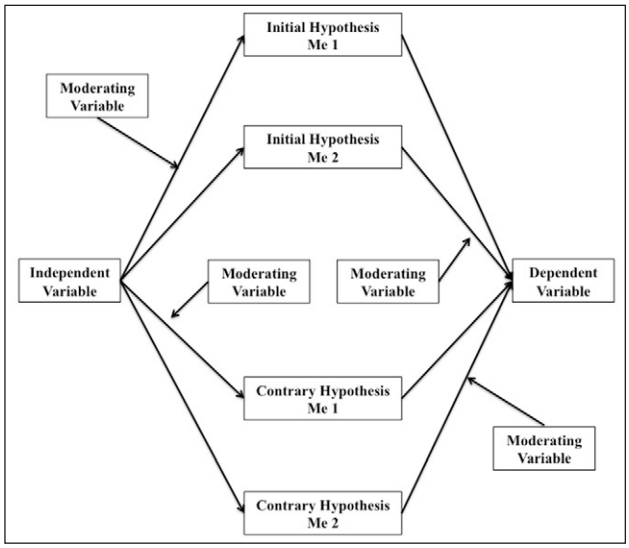

Figure 1: A hypothetical example of a theoretical model derived through the perspectivist approach (based on Jost, Banaji, & Prentice, 2004, pp. 319–332).

due to the introduction of a high level of complexity in theoretical reasoning, it offers a broader view of the conditions of validity of expected effects and thus impacts the type of research to be conducted empirically. Rather than running research based on a vague theoretical conceptualization containing an unknown number of hidden auxiliary assumptions, perspectivism invites researchers to first start by conducting highly selective studies that focus on some key parts of the overall model. Then, subsequent studies should be guided by this a priori framework, but also by the findings obtained by previous research. In this respect, empirical confrontation serves not only to test a priori hypotheses but plays an active role in the discovery process through exploratory research. Importantly, if this exploratory step leads to discovering new, relevant, and meaningful relationships (e.g., a post-hoc moderator of a hypothesis), these relationships should then be addressed in follow-up confirmatory studies. In this way, one can clearly see that the perspectivist process of knowledge construction is a continuous interplay between confirmation and exploration.

In sum, we believe that a wider endorsement of the perspectivist approach in social psychology would positively relate to the replicability rate of published studies. First, as we already mentioned, emphasizing auxiliary assumptions that are present throughout the process of theory construction and empirical assessment would allow making clear boundary conditions to a theory's validity. Thus, contexts where a theory is expected to hold true can be anticipated for attempts of replication, which is likely to increase the chances of success. Furthermore, because perspectivism gives researchers legitimate reasons to expect that a theory might not hold true in a particular context, theoretically meaningful predictions on the absence of an effect can be made. In other words, the perspectivist agenda could increase the extent to which researchers focus on corroborating the null hypothesis. This, in turn, could provide an extra argument for publishing "null" results and hence bring some balance in the published literature between "positive" and "negative" findings (see also Dienes, 2016).

A possible criticism of perspectivism would be to argue that the endorsement of this approach can lead to an extremist position where one considers that literally every proposition can be considered as true if the appropriate context had been found. For instance, if we push the perspectivist argument to its extreme, should we not consider that even highly implausible propositions – such as Bem's (2009) results on extra-sensory perceptions or Simmons et al.'s (2011) spurious effect of age change by listening to music – could be true in some very particular contexts? Actually, should some kinds of hypotheses not be discarded a priori instead of retained in the hope that some context might be found in which they hold true? Is there not a threat to perspectivism's validity?

We do not think this is the case. Perspectivism asserts that every hypothesis is both right and wrong as a function of context. For this context to be investigated, a hypothesis must be deductively derived from a coherent theoretical system so that speculations can be made about the mechanism involved. An explanatory mechanism for a hypothesis can then be predicted to hold in some contexts and not in others. After a closer inspection, one realizes that such an inquiry is not possible with propositions like those of Bem (2009) or Simmons et al. (2011) simply because they do not rely on a hypothesis based on a theory. In other words, these are not genuine hypotheses that could be studied with the perspectivist approach, which applies only to hypothetico-deductive propositions. Without a theory, no claim about the underlying mechanism accounting for a proposition can be made. Hence, there is no way to identify contexts in which it could or not hold. Consequently, an a priori rejection of an ungrounded claim's validity – regardless of its plausibility – does not conflict with the endorsement of perspectivism.

On the other hand, we do not believe that hypotheses consistent with the existing body of knowledge should be ruled out for being a priori – that is, without empirical investigation – deemed false in every context. First, because science should be about constructing knowledge based on confrontation between hypothetical conjectures with empirical tests and not based on personal or ideological beliefs. Second, because as long as one deals with a genuine theory, that is, a coherent system of causal relationships that obeys the rules of formal logic, there is no possible a priori way to establish its falsehood (i.e., without subjecting it to empirical test). Moreover, as we argued throughout the article, it is not even possible to establish absolute falsehood based on empirical assessment (hence the value of perspectivism). While perspectivism asserts that all hypotheses hold both true and false depending on contexts under consideration, it does not mean that every theory should be endlessly pursued if one fails to find a sufficiently interesting and relevant context for the theory to be true. Abandoning theories that "do not work out" is part of doing science. However, such decisions cannot be solely accounted by purely logical criteria. They depend on sociological and psychological factors proper to a scientific community (Kuhn, 1962). Each scientific community endorses its own set of values that are relevant in evaluating theories (e.g., degree of precision, range of application, simplicity, fruitfulness, etc.) and that guide scholars in their choices of which theory is worth investigating and which is not (Chalmers, 1982).

Finally, one should not forget that, in the perspectivist approach, social psychology could ultimately be viewed as a form of philosophy that has no pretense of describing the outside world. Every hypothesis could be regarded as being limited by a specific set of "perspectives", outside of which the hypothesis would no longer hold. Perspectivist social psychology could thus be considered as a contemplative approach to knowledge construction, where the richness of the theory – rather than its validity – can become an end in itself. We hope however to have made a compelling case that adopting this approach would be beneficial for our field. In essence, it urges the researcher to anticipate and make overt necessary auxiliary assumptions within a strategically planned research program.

Conclusion

In this article, we have sought to address the current issue of the replicability crisis in social psychology. First, we made a state of the art of this crisis and exposed some of the most emphatic and illustrative examples of its occurrence. We then shed light on the causes that are, in our sense, the most likely to have spawned the crisis. Specifically, we pointed out that the recurrent issue of low statistical power in studies in social psychology may be at least partly responsible for the low rate of replicability in the domain. We argued that the prevalent use of questionable research practices (John et al., 2012), researcher's degrees of freedom (Simmons et al., 2011), and HARKing (Kerr, 1998) can also account for the replicability crisis. We then emphasized the fact that current publication standards may promote such bad practices, leading to deleterious consequences on the published literature. Among these, we discussed the problematic use of the null hypothesis significance testing and the associated requirement of a $p < 0.05$ as a condition of publication, along with unrealistic expectations regarding researchers' data and the need for novelty (Giner-Sorolla, 2012). Finally, we also pointed out that the neglect of possible auxiliary assumptions relative to context-dependency variables could also account for an unexpectedly low rate of replicability in the domain of social psychology (Van Bavel et al., 2016b). With regard to the latter issue, we argued that shifting toward a more perspectivist social psychology could be beneficial to improve the replicability of studies in social psychology.

In our view, a greater reliance on perspectivism in social psychology could be greatly beneficial for the field. Arguably, making explicit the assumptions and the contexts in which postulated theories hold and those in which they do not could be a great step toward improving the replicability rate of social psychological theories. As a first step toward this improvement, authors should remember to always include relevant information about their samples when reporting their studies, beyond participants' sex and age (e.g., country of origin, academic section for students). Likewise, the social context in which the studies were conducted (e.g., type of social interactions between experimenters and participants during or before the experiment, presence of incentive or reward) should also be described with more details than what is currently done in most papers. In this regard, Klein et al. (2012) provided a list of guidelines for specifying some methodological information useful for determining boundary conditions of assessed effects.

Obviously, *there is no such thing as a free lunch*. Acknowledging the importance of auxiliary assumptions in theory elaboration and testing puts into question one's theory's universality. On the other hand, however, it is precisely the desire and aspirations for theoretical universality that might have partly led to the apparent crisis in our field in the first place. In our view, it is more reasonable to assume that many of our theories are subject to contextual and cultural dependency. This should, after all, come as no surprise for researchers from the field that includes "social" in its name.

Notes

[1] Power analysis can be easily performed with the freely available G*Power 3 software (Faul, Erdfelder, Lang, & Buchner, 2007).

[2] We hasten to make clear that this statistical tool is not *bad* per se, contrary to what might be thought based on the recent debates on the difficulties associated with *p*-values (e.g., Branch, 2014; Trafimow & Marks, 2015). Rather, it is the way NHST is typically used in social sciences in general that would require a thorough reconsideration (see Gigerenzer, 2004; Gigerenzer, Krauss, & Vitouch, 2004).

[3] According to the Popperian epistemology, scientific theories cannot be proven to be true but only to be false (see Chalmers, 1982, and Dienes, 2008, for a review). This asymmetry stems from the fact that observing a theory-consistent observation cannot deductively lead to confirming the theory, but a theory-inconsistent observation can deductively lead to disconfirming the theory. Thus, a tested theory can either be rejected if it is falsified by empirical test or not be rejected (i.e., corroborated but not confirmed) if it survives the attempt of falsification. Popper argued that focusing on refuting theories would ensure the progress of science because theories that survive attempts of falsification tend to converge to the truth, even though the latter cannot be ultimately attained (Popper, 1969, as cited in Chalmers, 1982).

[4] *P*-values can be used to falsify predictions in equivalence testing (Rogers, Howard, & Vessey, 1993), yet even those are still hardly ever performed in the psychological literature.

[5] Logical empiricism – also referred to as logical positivism – is a movement in philosophy of science that was founded at the beginning of the 20th century by a group of scientists and philosophers called Vienna Circle (*Wienerkreis*). Within this approach to scientific construction of knowledge, researchers should derive their hypotheses from empirically anchored theoretical systems to then be subjected to empirical tests (see Dienes, 2008, and McGuire, 1983, 1999, for a review). Although logical empiricism is not the dominant approach in philosophy of science anymore, it still has been very influential in psychology across the second half of the last century (McGuire, 1999).

Acknowledgement

We would like to thank two anonymous reviewers for useful comments on the previous version of this article. We also thank Jessica Gale for her help in proofreading.

Competing Interests

The authors have no competing interests to declare.

Author's Note

This article was supported by the Swiss National Science Foundation (SNF project n°100014_159464).

References

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: https://doi.org/10.1177/1745691612459060

Bakker, M., & **Wicherts, J. M.** (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*, 666–678. DOI: https://doi.org/10.3758/s13428-011-0089-5

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. DOI: https://doi.org/10.1037/0022-3514.71.2.230

Baumeister, R. F., Bratslavsky, E., Muraven, M., & **Tice, D. M.** (1998). Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology, 74*, 1252–1265. DOI: https://doi.org/10.1037/0022-3514.74.5.1252

Bem, D. J. (2003). Writing the empirical journal article. In: Zanna, M. P., Darley, J. M., & Roediger, H. L., III. (Eds.), *The compleat academic: A career guide* (2nd ed., pp. 185–219). Washington, D.C.: American Psychological Association.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100*, 407–425. DOI: https://doi.org/10.1037/a0021524

Branch, M. (2014). Malignant side effects of null-hypothesis significance testing. *Theory & Psychology, 24*, 256–277. DOI: https://doi.org/10.1177/0959354314525282

Brewer, M. B., & **Crano, W. D.** (2014). Research design and issues of validity. In: Reis, H. T., & Judd, C. M. (Eds.), *Handbook of research methods in social and personality psychology* (pp. 11–26), New York: Cambridge University Press.

Chalmers, A. F. (1982). *What is this thing called science?: An assessment of the nature and status of science and its methods* (2nd ed.). St. Lucia: University of Queensland Press.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65*, 145–153. DOI: https://doi.org/10.1037/h0045186

Cohen, J. (1965). Some statistical issues in psychological research. In: Wolman, B. B. (Ed.), *Handbook of clinical psychology* (pp. 95–121). New York: McGraw-Hill.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304–1312. DOI: https://doi.org/10.1037/0003-066X.45.12.1304

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159. DOI: https://doi.org/10.1037/0033-2909.112.1.155

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997–1003. DOI: https://doi.org/10.1037/0003-066X.49.12.997

Cooper, M. L. (2016). Editorial. *Journal of Personality and Social Psychology, 110*, 431–434. DOI: https://doi.org/10.1037/pspp0000033

Coulson, M., Healey, M., Fidler, F., & **Cumming, G.** (2010). Confidence intervals permit, but do not guarantee, better inference than statistical significance testing. *Frontiers in Psychology, 1*(26), 1–9. DOI: https://doi.org/10.3389/fpsyg.2010.00026

Crocker, J., & **Cooper, M. L.** (2011). Addressing scientific fraud. *Science, 334*, 1182. DOI: https://doi.org/10.1126/science.1216775

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*, 7–29. DOI: https://doi.org/10.1177/0956797613504966

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. New York: Palgrave Macmillan.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science, 6*, 274–290. DOI: https://doi.org/10.1177/1745691611406920

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology, 78*, 78–89. DOI: https://doi.org/10.1016/j.jmp.2015.10.003

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: https://doi.org/10.1371/journal.pone.0029081

Etz, A., & **Vandekerckhove, J.** (2016). A Bayesian Perspective on the Reproducibility Project: Psychology. *PLOS ONE, 11*(2), e0149794. DOI: https://doi.org/10.1371/journal.pone.0149794

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics, 90*, 891–904. DOI: https://doi.org/10.1007/s11192-011-0494-7

Faul, F., Erdfelder, E., Lang, A.-G., & **Buchner, A.** (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191. DOI: https://doi.org/10.3758/BF03193146

Ferguson, C. J., & **Heene, M.** (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science, 7*, 555–561. DOI: https://doi.org/10.1177/1745691612459059

Fischhoff, B. (1975). Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance, 1*(3), 288–299. DOI: https://doi.org/10.1037/0096-1523.1.3.288

Fiske, S. T. (2004). Mind the gap: In praise of informal sources of formal theory. *Personality and Social Psychology Review, 8*(2), 132–137. DOI: https://doi.org/10.1207/s15327957pspr0802_6

Forstmeier, W., Wagenmakers, E.-J., & **Parker, T. H.** (2016). Detecting and avoiding likely false positive findings – a practical guide. *Biological Reviews.* DOI: https://doi.org/10.1111/brv.12315

Francis, G. (2012). The psychology of replication and replication in psychology. *Perspectives on Psychological Science, 7*(6), 585–594. DOI: https://doi.org/10.1177/1745691612459520

Galak, J., LeBoeuf, R. A., Nelson, L. D., & **Simmons, J. P.** (2012). Correcting the past: Failures to replicate psi. *Journal of Personality and Social Psychology, 103*, 933–948. DOI: https://doi.org/10.1037/a0029709

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review, 116*, 439–453. DOI: https://doi.org/10.1037/a0015251

Garfield, E. (2006). The history and meaning of the journal impact factor. *Journal of the American Medical Association, 295*(1), 90–93. DOI: https://doi.org/10.1001/jama.295.1.90

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

Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics, 33*, 587–606. DOI: https://doi.org/10.1016/j.socec.2004.09.033

Gigerenzer, G., Krauss, S., & **Vitouch.** (2004). The null ritual: What you always wanted to know about significance testing but were afraid to ask. In: Kaplan, D. (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 391–408). Thousand Oaks, CA: Sage.

Gigerenzer, G., & **Marewski, J. N.** (2015). Surrogate science: The idol of a universal method for scientific inference. *Journal of Management, 41*, 421–440. DOI: https://doi.org/10.1177/0149206314547522

Giner-Sorolla, R. (2012). Science or art?: How aesthetic standards grease the way through the publication bottleneck but undermine science. *Perspectives on Psychological Science, 7*, 562–571. DOI: https://doi.org/10.1177/1745691612457576

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

Hoekstra, R., Finch, S., Kiers, H. A., & **Johnson, A.** (2006). Probability as certainty: Dichotomous thinking and the misuse of *p* values. *Psychonomic Bulletin & Review, 13*, 1033–1037. DOI: https://doi.org/10.3758/BF03213921

Hofstede, G. H., & **Hofstede, G.** (2001). *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. Thousand Oaks, CA: Sage.

Inbar, Y. (2016). The association between "contextual dependence" and replicability in psychology may be spurious. *Proceedings of the National Academy of Sciences of the United States of America, 113*(34), E4933–E4934. DOI: https://doi.org/10.1073/pnas.1608676113

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine, 2*(8), e124. DOI: https://doi.org/10.1371/journal.pmed.0020124

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: https://doi.org/10.1177/0956797611430953

Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences, 110*, 19313–19317. DOI: https://doi.org/10.1073/pnas.1313476110

Jost, J. T., Banaji, M. R., & **Prentice, D. A.** (2004). *Perspectivism in social psychology: The yin and yang of scientific progress*. Washington, D.C.: American Psychological Association. DOI: https://doi.org/10.1037/10750-000

Jost, J. T., & **Kruglanski, A. W.** (2002). The enstrangement of social constructionism and experimental social psychology: History of the rift and prospects of reconciliation. *Personality and Social Psychology Review, 6*(3), 168–187. DOI: https://doi.org/10.1207/S15327957PSPR0603_1

Judd, C. M., & **Gawronski, B.** (2011). Editorial comment. *Journal of Personality and Social Psychology, 100*(3), 406. DOI: https://doi.org/10.1037/0022789

Keppel, G. (1991). *Design and analysis: A researcher's handbook* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*, 196–217. DOI: https://doi.org/10.1207/s15327957pspr0203_4

Klein, O., Doyen, S., Leys, C., Magalhães de Saldanha da Gama, Miller, S., Questienne, L., & **Cleeremens, A.** (2012). Low hopes, high expectations: Expentancy effects and replicability of behavioral experiments. *Perspectives on Psychological Science, 7*(6), 572–584. DOI: https://doi.org/10.1177/1745691612463704

Kuhn, T. S. (1962). *The structure of scientific revolutions.* Chicago: University of Chicago Press.

Lakatos, I. (1978). *The methodology of scientific research programmes*. London and New York: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511621123

Lakens, D., & **Evers, E. R. K.** (2014). Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science, 9*, 278–292. DOI: https://doi.org/10.1177/1745691614528520

Loken, E., & **Gelman, A.** (2017). Measurement error and the replication crisis. *Science, 355*(6325), 584–585. DOI: https://doi.org/10.1126/science.aal3618

Makel, M. C., Plucker, J. A., & **Hegarty, B.** (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science, 7*, 537–542. DOI: https://doi.org/10.1177/1745691612460688

Markus, H. R., & **Kitayama, S.** (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review, 92*, 224–253. DOI: https://doi.org/10.1037/0033-295X.98.2.224

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes,

consequences, and remedies. *Psychological Methods, 9*, 147–163. DOI: https://doi.org/10.1037/1082-989X.9.2.147

Maxwell, S. E., Kelley, K., & **Rausch, J. R.** (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology, 59*, 537–563. DOI: https://doi.org/10.1146/annurev.psych.59.103006.093735

McGuire, W. J. (1983). A contextualist theory of knowledge: Its implications for innovation and reform in psychological research. In: Berkowitz, L. (Ed.), *Advances in experimental social psychology, 16*, 1–47. Orlando, FL: Academic Press. DOI: https://doi.org/10.1016/s0065-2601(08)60393-7

McGuire, W. J. (1989). A perspectivist approach to the strategic planning of programmatic scientific research. In: Gholson, B., Shadish, W. R., Jr., Neimeyer, R. A., & Houts, A. C. (eds.), *Psychology of science: Contribution to metascience* (pp. 214–245). New York: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9781139173667.012

McGuire, W. J. (1997). Creative hypothesis generating in psychology: Some useful heuristics. *Annual Review of Psychology, 48*, 1–30. DOI: https://doi.org/10.1146/annurev.psych.48.1.1

McGuire, W. J. (1999). *Constructing social psychology: Creative and critical processes.* Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511571206

McGuire, W. J. (2004). A perspectivist approach to theory construction. *Personality and Social Psychology Review, 8*, 173–182. DOI: https://doi.org/10.1207/s15327957pspr0802_11

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*, 103–115. DOI: https://doi.org/10.1086/288135

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*, 806–834. DOI: https://doi.org/10.1037/0022-006X.46.4.806

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1*, 108–141. DOI: https://doi.org/10.1207/s15327965pli0102_1

Meehl, P. E. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In: Harlow, L., Mulaik, S. A., & Steiger, J. H. (Eds.), *What if there were no significance tests?* (pp. 393–425). Mahwah, NJ: Erlbaum.

Mulder, J., & **Wagenmakers, E.-J.** (2016). Editors' introduction to the special issue "Bayes factors for testing hypotheses in psychological research: Practical relevance and new developments". *Journal of Mathematical Psychology, 72,* 1–5. DOI: https://doi.org/10.1016/j.jmp.2016.01.002

Neuliep, J. W., & **Crandall, R.** (1993). Reviewer bias against replication research. *Journal of Social Behavior and Personality*, *8*, 21–29.

Nisbett, R. E., Peng, K., Choi, I., & **Norenzayan, A.** (2001). Culture and systems of thought: Holistic versus analytic cognition. *Psychological Review, 108*, 291–310. DOI: https://doi.org/10.1037//0033-295X.108.2.291

Norenzayan, A., & **Heine, S. J.** (2005). Psychological universals: What are they and how can we know? *Psychological Bulletin, 131*(5), 763–784. DOI: https://doi.org/10.1037/0033-2909.131.5.763

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. DOI: https://doi.org/10.1177/1745691612459058

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & **Wicherts, J. M.** (2015). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods, 48*, 1205–1226. DOI: https://doi.org/10.3758/s13428-015-0664-2

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. DOI: https://doi.org/10.1177/1745691612462588

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*, aac4716. DOI: https://doi.org/10.1126/science.aac4716

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. DOI: https://doi.org/10.1177/1745691612465253

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

Richard, F. D., Bond, C. F., Jr., & **Stokes-Zoota, J. J.** (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7*, 331–363. DOI: https://doi.org/10.1037/1089-2680.7.4.331

Rogers, J. L., Howard, K. I., & **Vessey, J. T.** (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin, 113*(3), 553–565. DOI: https://doi.org/10.1037/0033-2909.113.3.553

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*, 638–641. DOI: https://doi.org/10.1037/0033-2909.86.3.638

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology, 58*, 646–656. DOI: https://doi.org/10.1037/0022-006X.58.5.646

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & **Iverson, G.** (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*, 225–237. DOI: https://doi.org/10.3758/PBR.16.2.225

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles.

Psychological Methods, 17, 551–566. DOI: https://doi.org/10.1037/a0029487

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: https://doi.org/10.1037/a0015108

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology, 51*, 515–530. DOI: https://doi.org/10.1037/0022-3514.51.3.515

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

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. DOI: https://doi.org/10.1177/0956797611417632

Stangor, C., & **Lemay, E. P., Jr.** (2016). Introduction to the special issue on methodological rigor and replicability. *Journal of Experimental Social Psychology, 66*, 1–3. DOI: https://doi.org/10.1016/j.jesp.2016.02.006

Stukas, A. A., & **Cumming, G.** (2014). Interpreting effect sizes: Toward a quantitative cumulative social psychology. *European Journal of Social Psychology, 44*, 711–722. DOI: https://doi.org/10.1002/ejsp.2019

Świątkowski, W. (submitted). On the use of subjective probabilities in data analysis: An introduction to Bayesian statistics for social psychology starters.

Trafimow, D. (2009). The theory of reasoned action: A case study of falsification in psychology. *Theory & Psychology, 19*(4), 501–518. DOI: https://doi.org/10.1177/0959354309336319

Trafimow, D., & **Marks, M.** (2015). Editorial. *Basic and Applied Social Psychology, 37*(1), 1–2. DOI: https://doi.org/10.1080/01973533.2015.1012991

Tressoldi, P. E. (2012). Replication unreliability in psychology: elusive phenomena or "elusive" statistical power? *Frontiers in Psychology, 3,* 1–5. DOI: https://doi.org/10.3389/fpsyg.2012.00218

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & **Reinero, D. A.** (2016a). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences, 113*, 6454–6459. DOI: https://doi.org/10.1073/pnas.1521897113

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & **Reinero, D. A.** (2016b). Reply to Inbar: Contextual sensitivity helps explain the reproducibility gap between social and cognitive psychology. *Proceedings of the National Academy of Sciences, 113,* E4935–E4936. DOI: https://doi.org/10.1073/pnas.1609700113

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review, 14*, 779–804. DOI: https://doi.org/10.3758/BF03194105

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & **van der Maas, H.** (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology, 100*, 426–432. DOI: https://doi.org/10.1037/a0022790

Wallach, L., & **Wallach, M. A.** (2010). Some theories are unfalsifiable: A comment on Trafimow. *Theory & Psychology, 20*, 703–706. DOI: https://doi.org/10.1177/0959354310373676

Wasserstein, R. L., & **Lazar, N. A.** (2016). The ASA's statement on *p*-values: Context, process, and purpose. *The American Statistician, 70*, 129–133. DOI: https://doi.org/10.1080/00031305.2016.1154108

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & **Wagenmakers, E.-J.** (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t-tests. *Perspectives on Psychological Science, 6*(3), 291–298. DOI: https://doi.org/10.1177/1745691611406923

How to cite this article: Świątkowski, W., & Dompnier, B. (2017). Replicability Crisis in Social Psychology: Looking at the Past to Find New Pathways for the Future. *International Review of Social Psychology*, 30(1), 111–124, DOI: https://doi.org/10.5334/irsp.66

Published: 02 May 2017