DFG Priority Program SPP 2317

META-REP:

A meta-scientific program to analyze and optimize replicability in the behavioral, social, and cognitive sciences

Program Committee:

Coordinator:

Prof. Dr. Mario Gollwitzer — Ludwig-Maximilians-Universität, München

Other Members:

Prof. Dr. Katrin Auspurg — Ludwig-Maximilians-Universität, München

Dr. Susann Fiedler — Max Planck Institute for Research on Collective Goods, Bonn

Dr. Tina Lonsdorf — Universitätsklinikum Hamburg-Eppendorf, Hamburg

PD Dr. Felix Schönbrodt — Ludwig-Maximilians-Universität, München

1 Summary

The behavioral, social, and cognitive sciences are in the midst of an intensive debate about the replicability of their empirical findings. Shaken by the results of many replication projects that have been conducted in recent years, scholars have begun discussing (1) what replicability actually means and when a replication can be regarded as successful vs. failed; (2) whether low replication rates are due to too many false positive findings in the literature, to methodological weaknesses in the replication studies, or to the underappreciated influence of contextual effects, and (3) what can be done to effectively and sustainably secure a high level of replicability in the behavioral, social, and cognitive sciences. So far, most of these discussions have been based on ad-hoc arguments. Thus, a concerted, integrative, and interdisciplinary meta-scientific research program is needed to systematically scrutinize these arguments. "META-REP" is such a program. Projects joining "META-REP" will contribute to the emerging field of replication science. More specifically, projects will (1) define, operationalize, and assess replicability and replication success (vs. failure) in their respective field precisely and validly (the "what" question); (2) comprehensively explain why replication rates vary across and within scientific fields (the "why" question); and (3) evaluate and optimize the effectiveness and suitability of potential strategies aimed at increasing or maintaining replication rates in different fields (the "how" question). The results from this program will be relevant for all sciences that are discussing issues of replicability. On a broader level, this program will enrich the public discourse about the credibility, the value, and the usefulness of science in general.

2 State of the Art and Preliminary Work

Many behavioral, social, and cognitive sciences have been disconcerted by the surprisingly low replication rates that have been reported in recent years. In psychology, for instance, several extensive replication projects (e.g., the "Reproducibility Project: Psychology"; Open Science Collaboration, 2015; or the "Many Labs" projects; Ebersole et al., 2016; Klein et al., 2014, 2018) suggest that many psychological findings – including those that have been published in high-impact journals and presented in standard psychology textbooks – cannot be replicated. But psychology is not an exception. Similar replication projects conducted in behavioral economics, the life sciences, and the neurosciences mirror the findings observed in psychology (e.g., Begley & Ellis, 2012; Button et al., 2013; Camerer et al., 2016, 2018; Ioannidis, Munafò, Fusar-Poli, Nosek, & David, 2014; Poldrack et al., 2017; Prinz, Schlange, & Asadullah, 2011). These findings – as daunting as they may appear – have led researchers to critically assess their scientific practices, their theories, and the incentive systems that have dominated their respective fields. Consequently, we are now seeing (1) an intensifying scholarly exchange about "replicability" particularly in the behavioral, cognitive, and social sciences (e.g., Dettweiler, Hanfstingl, Schröter, & Burzan, 2019; Goodman, Fanelli, & Ioannidis, 2016; note also the steadily growing number of publications on this topic displayed in Figure 1), and (2) a radical paradigm shift regarding the degree of transparency and methodological rigor that funding agencies, scientific journals, and science organizations are currently demanding (see Munafò et al., 2017; Vazire, 2018).

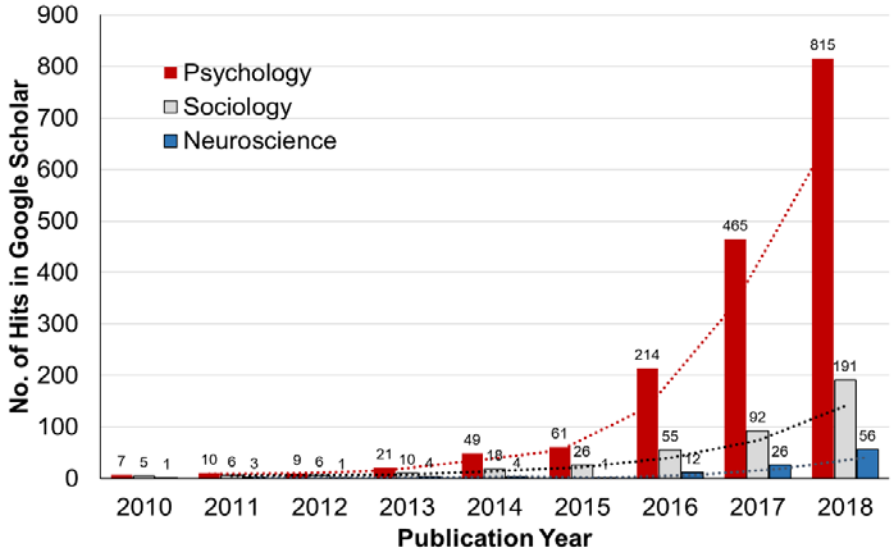

Figure 1: Increase in published articles on the "replication crisis" in psychology, sociology, and the neurosciences between 2010 and 2018 (source: Google Scholar)

Replicability has become a (meta-)scientific research object, and scholars in the behavioral, social, and cognitive sciences are unraveling critical questions such as what replicability actually means, why replication rates are sometimes so low, and how replicability can be improved. Importantly, understanding replicability is also societally relevant given that it is directly linked to laypeople's and decision-makers' trust in science (e.g., Hendriks, Kienhues, & Bromme, 2016; Wingen, Berkessel, & Englich, 2019). The Priority Program "META-REP" proposed here will contribute significantly to this debate by (1) **describing and defining** "replication" (including "successful" vs. "failed" replications) across different scientific disciplines (the "what" question), (2) **explaining** why replication rates vary across and within different disciplines (the "why" question), and (3) **evaluating** measures that have been proposed and implemented to increase replication rates (the "how" question). These three aspects will be discussed in the following sections.

2.1 The "What" Question: What is Replicability and When Can Replications Be Regarded Successful (vs. Failed)?

In quantitative research, the term "replication" has been used to characterize a variety of different activities (e.g., Freese & Peterson, 2017). These activities can be categorized along two dimensions, that is, (1) whether or not new data were collected for the replication, and/or (2) whether the replication differed from the original study regarding design, measures, or data analysis. Combining these two dimensions results in a 2×2 matrix (see Table 1). Although the specific terms sometimes differ across disciplines (Barba, 2018; Duvendack, Palmer-Jones, & Reed, 2017; Goodman et al., 2016; Hamermesh, 2007; National Academies of Sciences, Engineering, & Medicine, 2019) and different disciplines have focused on different cells so far (Christensen, Freese, & Miguel, 2019), the basic idea can be applied to all empirical sciences.

Table 1: Types of "Replications" in the Behavioral, Social, and Cognitive Sciences

	...Using the same data	**...By collecting new data**
Same materials, methods, code (i.e., no deviation)	Reproducibility Analysis	Direct/"Exact" Replication
Different materials, methods, code etc.	Robustness Analysis	Generalizability Analysis (Extension; "Conceptual" Replication)

Apart from the question what "replication" means and how it is understood in different disciplines, an unresolved question is when a replication attempt can be considered "successful" vs. "failed." Regarding *reproducibility* (the upper left cell in Table 1), the question is how much deviation from the originally reported result is acceptable: zero changes, a particular absolute or percentage deviation? For instance, in brain imaging research, which is typical in the cognitive neurosciences, the question how much spatial overlap (i.e., in voxels) is sufficient to count as a successful "spatial replication" is unclear (Hong, Yoo, Han, Wager, & Woo, 2019). For *robustness* analyses, *direct replications*, and *conceptual replications*, defining "success" vs. "failure" is even more difficult. The problem starts with a proper definition of the "replicandum" (see Earp & Trafimow, 2015; Wong & Steiner, 2018). The question is: what is it that needs to be replicated exactly?

For example, in their *Reproducibility Project: Psychology*, which conducted "direct" replications (the upper right cell in Table 1), the Open Science Collaboration (2015) analyzed whether the effect observed in the replication study was statistically significant in the same direction as the effect observed in the original study, and whether the original point estimate of the population effect size was within the replication studies' confidence interval. In addition, the authors used a subjective assessment of the replication success. Other authors proposed alternative approaches to classify (LeBel, McCarthy, Earp, Elson, & Vanpaemel, 2018) or quantify replication success, building, for instance, on "small-telescope testing" (Simonsohn, 2015), Bayesian inference (Patil, Peng, & Leek, 2016; Verhagen & Wagenmakers, 2014), or estimates of the *a priori* probability of replicating a "true" effect (Trafimow, 2018). To date, there is no single accepted definition or operationalization of "replication success" within or across disciplines (see also Nosek & Errington, 2019). Notably, replicability base-rates may vary considerably across disciplines and sub-fields within a discipline: the cognitive sciences are more likely to deal with "anthropologically constant" effects than, for instance, social psychology (Gergen, 1976, 1996). In order to evaluate and compare replication rates across disciplines, replicability base-rates (or priors) and discipline-specific sampling errors need to be taken into account (see Bird, 2018; Ioannidis, 2005; Ulrich et al., 2016).

In addition, there has been some debate about whether previous replication projects (such as the *Reproducibility Project: Psychology*) might have underestimated the "true" replicability of behavioral, social, and cognitive science effects. For instance, several scholars have highlighted that the studies included in the *Reproducibility Project: Psychology* – the last study in a multiple-study paper published in one of three high-impact journals in psychology – might be particularly susceptible to an upward-biased effect size estimate (e.g., Fiedler, 2018a; Gilbert, King, Pettigrew, & Wilson, 2016; Ingre & Nilsonne, 2018). One consequence of this upward bias in the original study is that replication studies are often underpowered (Erdfelder & Ulrich, 2018; Etz & Vandekerckhove, 2016). Also, the finding that effect size estimates in the replication studies were considerably lower than in the original studies might represent a regression artifact (Fiedler & Prager, 2018).

The Priority Program proposed here – "META-REP" – will scrutinize the conceptual and empirical tenability of these arguments and, thus, bring some order into the current debate about replicability and replication success in the behavioral, social, and cognitive sciences. More specifically, projects included in "META-REP" will

1. **evaluate, compare, and refine methodological and statistical approaches** to estimate reproducibility, robustness, "direct" replicability, and generalizability, respectively, and assess their applicability in different disciplines – these approaches include, for instance, "systematic forking paths"/"specification curve"/"multiverse" analyses (see Simonsohn, Simmons, & Nelson, 2015; Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016; for examples, see Lonsdorf, Klingelhöfer-Jens, et al., 2019; Rohrer, Egloff, & Schmukle, 2017; Wacker, 2017; or the #narps Project in NeuroImaging), base-rate corrected replication rates (Bird, 2018), conditional causal effects estimates (Wong & Steiner, 2018), Bayesian estimates (Patil et al., 2016; Verhagen & Wagenmakers, 2014), or "replication closeness" indicators (Trafimow, 2018);
2. **assess the extent** to which reproducibility, robustness, "direct" replicability, and generalizability indices (in replication projects and meta-analyses) may provide a biased picture of the "true" replicability in different sciences (e.g., Fiedler & Prager, 2018; Stanley & Spence, 2014; Carter, Schönbrodt, Gervais, & Hilgard, 2019) and develop methods to correct for these biases;
3. by using best-practice approaches, **provide a valid and systematic picture** of the reproducibility, the robustness, the "direct" replicability, and the generalizability of behavioral, social, and cognitive science findings, while taking a more heterogeneous and more representative range of effects and methods – as well as specific historic and political backgrounds of different research areas – into account.

2.2 The "Why" Question: Why Do Replications Fail, and Why Do Replication Rates Vary Within and Across Disciplines?

Accepting that the behavioral, social, and cognitive sciences do have a problem with the replicability[1] of their effects inevitably leads to the "why" question. What are the causes and conditions that have led to (or amplified) this problem? Much has been speculated about these causes and conditions (e.g., De Boeck & Jeon, 2018; Nelson, Simmons, & Simonsohn, 2018; Nosek, Ebersole, DeHaven, & Mellor, 2018, just to name a few), but a systematic empirical investigation of these causes and conditions across disciplines is lacking. Two reasons for low replication rates have been discussed in depth so far: inflated false-positive rates and the underappreciated context dependency of effects.

[1] For reasons of simplicity, we will use "replicability" as an umbrella term here (instead of mentioning all four types of "replications" listed in Table 1). Yet, the arguments discussed here apply, in principal, equally well to reproducibility, the robustness, the "direct" replicability, and the generalizability of findings in the behavioral, social, and cognitive sciences.

2.2.1 Inflated False-Positive Rates

The assumption underlying this argument is that many effects are not reproducible, robust, and "directly" replicable (see Table 1) because they never existed in the first place – stated differently, that the literature is full of "false positive" findings (e.g., Ioannidis, 2005; Simmons, Nelson, & Simonsohn, 2011). Needless to say, effects that do not exist are unlikely to be replicated.

Causes for inflated false-positive rates can exist both on the individual (i.e., researcher) level as well as on the system (i.e., scientific community) level. On the *individual* level, "Questionable Research Practices" (QRPs; see Bakker, van Dijk, & Wicherts, 2012; John, Loewenstein, & Prelec, 2012; Simmons et al., 2011; see also Fiedler & Schwarz, 2016)[2] and unintended errors ("honest mistakes") may explain the non-replicability of a specific finding (Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2016). Notably, QRPs are not necessarily applied in bad faith; their prevalence may also reflect the perpetuation of (inappropriate) norms in a field (Smaldino & McElreath, 2016), the adoption of practices that are no longer applicable in a new context, or by motivated reasoning (Ioannidis et al., 2014). On the *system* level, publication bias (i.e., a system that favors the publication of hypothesis-congruent / "significant" over hypothesis-incongruent or "non-significant" results; e.g., Franco, Malhotra, & Simonovits, 2014; Nosek, Spies, & Motyl, 2012; Szucs & Ioannidis, 2017; Wilson & Wixted, 2018) or conflicts of interest (Lieb, van der Osten-Sacken, Stoffers-Winterling, Reiss, & Barth, 2016) can inflate false-positive rates and cause upward-biases regarding reported effects in the literature.

On the one hand, publication bias itself is sufficient to produce high false-positive rates, even if no QRPs are executed on the individual level. Low base rates of true effects and low statistical power (due to small sample sizes) amplify this problem (Ioannidis, 2005). On the other hand, publication bias and other aspects of the academic incentive structure (e.g., hiring decisions based on quantitative output in terms of publications and external funding, Dougherty & Horne, 2019) amplify individual-level processes by incentivizing researchers to exploit their analytical flexibility or to abandon "double-checking" for errors.

Notably, the discussion about QRPs (and the extent to which they result in false positives) has focused on experimental lab research, which is dominant in (some sub-disciplines within) psychology, behavioral economics, or the neurosciences. For instance, neuroscientists typically deal with high-dimensional data and an excessive work-flow flexibility allowing thousands of different options to analyze their data (Carp, 2012; Luck & Gaspelin, 2017; Poldrack et al., 2017) and massive requirements for correction of multiple comparisons (e.g., Eklund, Nichols, & Knutsson, 2016). At the same time, sample sizes (Turner, Paul, Miller, & Barbey, 2018) and, hence, statistical power is typically low (Button et al., 2013; Cremers, Wager, & Yarkoni, 2017). In the social sciences (such as quantitative sociology), by contrast, observational data are prevailing (Freese & Peterson, 2017), and there are often hundreds of justifiable possibilities to specify analytical models (e.g., in terms of which subgroups, variables, or interaction terms to be included in regressions; and which regression models to be used; see e.g. Christensen et al., 2019). At the same time, the number of cases is very large, meaning that there is a high statistical power to detect even very small effects. Thus, in quantitative sociology and other social science disciplines, the large analytical flexibility in combination with non-transparent reporting practices may be a more severe cause for non-replicability than the selective publishing of research observed in experimental psychology (Christensen et al., 2019; Christensen & Miguel, 2018; Munoz & Young, 2018; Wuttke, 2019).

[2] Such QRPs include (a) the selective reporting of statistically significant findings in a dataset (while neglecting non-significant ones; cf. Renkewitz, Fuchs, & Fiedler, 2011), (b) the strategic inclusion vs. exclusion of outliers (in order to decrease the *p*-value of one's focal effect), (c) the strategic inclusion of covariates and/or interaction terms (cf. Christensen et al., 2019) into the model, (d) strategic decisions to sample more participants (in order to make the focal effect statistically significant) or to stop data collection (if the *p*-value is below .05), and (e) "HARKing" – hypothesizing after the results are known (Kerr, 1998).

Currently, meta-science theories on the incentive structures fostering QRPs are largely missing (see Auspurg & Hinz, 2017; Cottrell, 2014; Feigenbaum & Levy, 1993). There is some theoretical work on incentive structures in science (see already Merton, 1957), but hardly any systematic research that links these incentive structures to replicability. However, only a clear understanding of the norms and the mechanisms according to which academic institutions and scientific communities operate will eventually help us produce robust and replicable research (Wuttke, 2019). By inviting research projects that specifically analyze the impact of individual- and system-level causes as well as their interactions on replicability, the proposed "META-REP" program will contribute to a better understanding and cross-disciplinary exchange on how false-positive rates in the literature can be explained. More precisely, projects included in "META-REP" will

1. **systematically document and compare** research practices, incentive systems, journal policies, and (implicit and explicit) normative expectations in specific areas (e.g., publication pressure; see Dougherty & Horne, 2019) in the behavioral, social, and cognitive sciences and their respective impact on replication rates (e.g., by correlating changes in norms to replication rates or by experimentally manipulating incentive systems);
2. **compare the impact** of system-immanent incentive structures with other, more generic societal factors (such as societal expectations towards scientists, political pressure, conflicts of interest, media attention, etc.) on individual practices and norms in different disciplines;
3. **re-analyze existing replication projects and meta-analyses** (i.e., meta-meta-analysis) with regard to individual- and system-level factors that produce biases in original and/or replication research (e.g., investigating publication bias in publication bias research; see Dubben & Beck-Bornholdt, 2005).

2.2.2 Context Dependency

The second argument for low replication rates is built on the assumption that empirical effects in the behavioral, social, and cognitive sciences are much more context-dependent than expected. Context dependency means that the original effect exists under the specific conditions under which it has been tested, but not under any other conditions. Notably, the context dependency argument differs from the false-positive argument described in Section 2.2.1: while the false-positive argument assumes that the reported effect size in an original study was considerably upward-biased (or that the effect did not even exist in the first place), the context-dependency argument acknowledges that the original effect had been estimated correctly, but that it existed only under specific conditions that were unfortunately absent in the replication study, such as sample characteristics, time/zeitgeist, location/culture, experimenter, materials, and methods.

Compared to the false-positive argument, the context-dependency argument has received less attention so far, although some scholars have used the argument to explain (post-hoc) why a particular effect could not be replicated in a "direct" replication project. Thus, context dependency as a possible explanation for the non-replicability of an effect is a double-edged sword: On the one hand, it is extremely informative to elucidate the boundary conditions under which an effect replicates or generalizes across studies. On the other hand, the notion of context dependency can be strategically used to "explain" the non-replicability of an effect, that is, to shield the effect from being challenged. This strategy is scientifically inappropriate because it immunizes a theoretical claim and makes it unfalsifiable (Lakatos, 1976; Meehl, 1990; Zwaan, Etz, Lucas, & Donnellan, 2018). The latter problem may explain why some scholars are so skeptical about the context dependency argument in the replicability debate. Nonetheless, it is theoretically plausible to assume that context dependency does predict replicability for some content domains.

Empirically, the context-dependency argument has received inconclusive support so far. On the one hand, by re-analyzing the *Replication Project: Psychology* data (Open Science Collaboration, 2015), Van Bavel, Mende-Siedlecki, Brady, and Reinero (2016) showed that the extent to which the effects were rated as context-sensitive (by experts) reliably predicted their replicability (for a critique on Van Bavel et al.'s approach, however, see Inbar, 2016). On the other hand, findings from the "ManyLabs 2" project (Klein et al., 2018), a concerted attempt to replicate 28 psychological effects, suggest that the specific context characteristics investigated here (i.e., culture, task order, lab vs. online setting) did not substantially moderate effect sizes. Of course, this does not rule out the possibility that other context variables moderate the effects. More systematic research is therefore needed to elucidate "context dependency" both conceptually and empirically (Pettigrew, 2018).

One important question that needs to be answered is whether variation in observed effect sizes (e.g., as observed in a ManyLabs study) is due (a) to a "true" heterogeneity of the focal effect as a function of moderators that are theoretically meaningful (because they qualify the effect on a conceptual level) or (b) to moderator effects on the operational or measurement level (i.e., construct validity; see Gollwitzer & Schwabe, 2019). For instance, studies suggesting that an experimental effect exists in some cultures, but not in others may either signify that the effect is truly culture-specific (e.g., Henrich, Heine, & Norenzayan, 2010; Kerr, Ao, Hogg, & Zhang, 2018), that a manipulation has "worked" in some cultures, but not in others (e.g., Magraw-Mickelson & Gollwitzer, 2018), or that measurement properties of the observed dependent variable differ across cultures (i.e., measurement invariance; see Hussey & Hughes, 2018).

Notably, context variables that appear "operational" at first glance may moderate the effect systematically and constitute substantive boundary conditions for an effect to emerge. A very good example in that regard is a series of projects that aimed at replicating the classic "facial feedback" effect (Strack, Martin, & Stepper, 1988), which is based on the hypothesis that activating the *Zygomatic major* (i.e., the "smiling muscle"), for instance, by holding a pen sidewise in one's mouth, makes participants find cartoons more amusing than, for instance, holding the pen lengthwise in one's mouth. A registered replication project involving 17 labs was unable to replicate the original effect (Wagenmakers et al., 2016), but the replication studies differed in a number of aspects from the original study. One difference was that, in the replication studies, participants were monitored via webcams (which did not exist in the 1980s). Recently, Noah, Schul, and Mayo (2018) were able to show that exactly this difference can account for the non-replicability of the original effect, and that being monitored is a substantive (and psychologically reasonable) boundary condition of the "facial feedback effect." This example shows the (hitherto unexploited) value of elucidating the context-dependency of effects and of differentiating between theoretically meaningful (i.e., substantive) and theoretically irrelevant (i.e., operational) moderator effects of "context."

In their critical commentary to the *Reproducibility Project: Psychology*, Gilbert et al. (2016) discuss other examples of context-related boundary conditions that may explain failed replications (for a response to their comment, however, see Anderson et al., 2016). Accepting that effects are context-specific and systematically depend on boundary conditions, and that these boundary conditions can be more or less theoretically relevant is important to correctly interpret "failed" replications, but only if these boundary conditions can be correctly estimated, are specified *a priori*, and are replicable themselves (Hardwicke & Shanks, 2016). The context-dependency argument also underscores the need for conceptual replications (or "generalizability analyses;" see Table 1). While there is now a certain consensus about the methodological criteria for "direct" replications (e.g., Erdfelder & Ulrich, 2018), the conceptual and empirical basis for conceptual replications is currently debated on various levels (i.e., philosophy of science, methodology, theory-building).

In sum, while there have been some valuable attempts to elucidate the context dependency of observed effects in the social, behavioral, and cognitive sciences, much more needs to be done across disciplines to fully understand when and why effects can or cannot be replicated and whether this is

due to variations of the effect itself or to methodological conditions under which the effect is investigated. Thus, individual projects included in "META-REP" will

1. **re-analyze existing data** (from meta-analyses and replication projects) with regard to the effects of (1) manipulation heterogeneity, (2) measurement invariance, and (3) sample characteristics on the heterogeneity / non-replicability of effects in different areas;
2. **analyze the usefulness of conceptual replications** from (1) a philosophy of science perspective (Feest, 2019), (2) a conceptual ("theory-building") perspective (i.e., how does a theory need to be specified in order to allow for appropriate conceptual replications; see Glöckner & Betsch, 2011), (3) a pragmatic perspective (i.e., what are standards for useful "close" or conceptual replications; see Brandt et al., 2014; Crandall & Sherman, 2016), and (4) a cost-benefit perspective (Miller & Ulrich, 2016);
3. **evaluate the impact** of underspecified theories, design characteristics, and invalid/questionable measurement models ("pseudo-measurements") on replication rates in different research areas.

2.3 The "How" Question: How Can a Sufficiently High Level of Replicability be Maintained Within and Across Disciplines?

Many of the remedies that have been proposed in the literature so far focus almost exclusively on methodological practices, such as implementing "open science" and stricter data management standards; demanding more highly-powered studies; asking researchers to report their statistical results in more detail (e.g., effect size estimates and confidence intervals); encouraging Bayesian hypothesis testing in lieu of null hypothesis significance testing; or banning inferential statistics from being reported in a paper altogether (e.g., Trafimow & Marks, 2015; for an evaluation of this practice, see Fricker Jr., Burke, Han, & Woodall, 2019). Recently, De Boeck and Jeon (2018) argued that although the literature is rife with suggestions on how to increase replication rates – including statistical "corrections" for biases, errors, and questionable research practices –, many of these suggestions are built on untested assumptions, and some of the suggestions may have unintended side effects. Thus, a systematic evaluation of these suggestions is needed, with regard to (a) the plausibility of their underlying assumptions, (b) their acceptability and implementability in different research areas and disciplines, (c) their effectiveness with regard to the primary outcome variable (i.e., replication rates), and (d) their unintended side effects.

Among the methodological remedies mentioned before, the implementation of "open science" standards in the context of article submissions and funding proposals has been most enthusiastically embraced, but also most heatedly discussed in the scientific community.[3] By motivating researchers to commit to these "open science" standards, the proponents of this view hope to decrease the prevalence of questionable research practices and honest mistakes (see Section 2.2.1) and, thus, to improve replication rates (e.g., Grahe, 2018; Jussim, Crawford, Anglin, Stevens, & Duarte, 2016; Kidwell et al., 2016; Sakaluk & Graham, 2018; Wicherts, Bakker, & Molenaar, 2011; Wicherts, Veldkamp, Augusteijn, Bakker, Van Aert, & Van Assen, 2016). However, three issues are worth discussing. *First*, we currently do not know whether these measures have any effect on increased replication rates. For example, even if there were more open data and analysis code available, QRPs

[3] Although the umbrella term "open science" is defined much more broadly (e.g., OECD, 2015), the following six facets of "open science" are particularly relevant (see LeBel et al., 2018; Spellman, Gilbert, & Corker, 2017): (1) Preregistered specification of the empirical hypotheses, the sampling and recruitment procedure, the analytical procedure, etc., (2) Compliance with reporting standards, such as adhering to established or newly developed reporting guidelines (e.g., the CONSORT guidelines for randomized controlled trials established in the life sciences; Schulz, Altman, & Moher, 2010), (3) Open materials, including study materials, videos or protocols showing mock trials of the experimental procedure, etc., (4) Data sharing in compliance with the FAIR principles (findable, accessible, interoperable, re-useable; see Wilkinson et al., 2016) and with current data documentation standards, such as DDI (https://www.ddialliance.org/), (5) Reproducible analysis code, and (6) Sharing research output (open access).

and errors might still exist, simply because researchers are lacking incentives to reproduce or replicate others' work (as replications are more difficult to publish and also less rewarding in terms of scientific recognition than "original" findings; see already Merton, 1957; Feigenbaum & Levy, 1993). Experiences in economics and political science have shown that published data and materials alone are not sufficient to make results reproducible; journals might also have to hire "data managers" who carefully check all files for errors and inappropriate statistical procedures (Vilhuber, 2019).

Second, while information about low replication rates (unsurprisingly) decreases laypeople's trust in science, recent research also suggests that information about "open science" reforms and research transparency does *not* increase public trust in science (Anvari & Lakens, 2019; Wingen et al., 2019; but see Hendriks et al., 2016). This demonstrates that seemingly plausible consequences of reforms do not necessarily occur and, hence, warrant empirical evaluation.

And, *third*, some scholars have speculated that enforcing new standards of openness and transparency might have undesired side effects, such as a one-sided focus on data analysis at the cost of proper theory development (Fiedler, 2017, 2018b), a one-sided focus on false positives at the cost of preventing "false negatives" (Fiedler, Kutzner, & Krueger, 2012), the creation of a social dilemma in which individual researchers that adhering to "open science" standards are (or, at least, feel) exploited by those who do not (Abele-Brehm, Gollwitzer, Steinberg, & Schönbrodt, 2019), or an unnecessary bureaucratization of the scientific process with undesirable side effects on researchers' creativity (Brainerd & Reyna, 2018; Kaufman & Glăveanu, 2018).

As of now, there are first preliminary attempts to evaluate the effects of an increasing "open science" culture with regard to its desirable and undesirable side effects (e.g., Vazire, 2018; Wai & Halpern, 2018), but a systematic research program in this regard is lacking. A collaborative effort to evaluate the effects of changes in incentive structures and norms on the systemic and the individual level is necessary; a Priority Program such as "META-REP" would provide the ideal platform for such a collaborative transdisciplinary effort.

Another potential cause of low replicability is a weak logical link between theories and their empirical tests (Glöckner & Betsch, 2011; Muthukrishna & Henrich, 2019; Oberauer & Lewandowsky, 2019). This also relates to the context-dependency argument (see Section 2.2.2), according to which substantive moderator variables may account for heterogeneity in effect sizes. In order to test the effects of these moderators, they need to be specified. Ideally, such a specification builds on a priori theorizing; however, it is more reasonable to assume that theory specification is an iterative process that requires both reasoning and interpreting empirical data. In line with this notion, Glöckner, Fiedler, and Renkewitz (2018) recently suggested that authors store their fully specified theories in a joint and versioned theory database. Such a database would (1) allow authors to efficiently and instantly revise their theory (e.g., add boundary conditions), (2) allow readers (and students) to immediately retrieve the current status of the theory (instead of, for instance, an outdated textbook version), and (3) allow researchers interested in replication to test the relevant versions of a theory. Post-hoc debates between the replicators and the original authors concerning boundary conditions and "hidden moderators" could thus be avoided, and theory development would become a more collaborative enterprise. To date, there is no empirical evidence that "cumulative theory-building" works, how it works best, and whether it is actually likely to have its desired effects (and no undesired side effects).

Taken together, projects included in "META-REP" will

1. **evaluate different measures, strategies, and tools** that have been proposed and implemented to increase replication rates with regard to (a) the theoretical plausibility and empirical validity of their underlying assumptions, (b) their acceptability and implementability in different research areas and disciplines, (c) their effectiveness with regard to the primary outcome variable (i.e., replication rates), and (d) their unintended side effects;

2. **refine existing suggestions** and propose new suggestions to increase replication rates while keeping a holistic, multi-level approach in mind (i.e., addressing both individual-level and system-level conditions as well as their interactions; see also Section 2.1);
3. **assess and evaluate instruments** to monitor changes in norms, incentive structures, and scientific practices in different disciplines (e.g., quality assurance measures in scientific journals, funding agencies, and scientific organizations).

3 Merits of the Proposal Taking into Account the Objectives of the Program

3.1 Originality of Research Questions in Terms of Topic and/or Methodology

We are in the midst of a lively and constructive discussion about the meanings and causes of low replication rates in the behavioral, social, and cognitive sciences and about potential improvements for the current situation. Now, the time is ripe to investigate these perspectives systematically and in a collaborative and interdisciplinary fashion. In other words, what is currently needed is a solid, evidence-based, and widely applicable (meta-)science of replicability (or "replication science"). The presently proposed "META-REP" program will contribute to this emerging field by (1) **describing and defining** "replication" (including "successful" vs. "failed" replications) across different scientific disciplines (the "what" question; see Section 2.1), (2) **explaining** why replication rates vary across and within different disciplines (the "why" question; see Section 2.2), and (3) **evaluating** measures that have been proposed and implemented to increase and maintain replication rates (the "how" question; see Section 2.3). The underlying assumption is that replicability is a research object that needs to be treated with the same scrutiny and scientific rigor as any other research object.

To achieve this goal, "META-REP" will focus on replicability in a selected subset of empirical sciences (i.e., the behavioral, social, and cognitive sciences), including psychology (basic and applied), cognitive neuroscience, psycholinguistics, sociology, behavioral economics, and communication science. Of course, replicability is also an issue in other sciences (such as the "hard" natural sciences or the life sciences; see Prinz et al., 2011). Yet, to ensure a certain sense of mutual understanding among researchers involved in "META-REP" and to facilitate collaboration between them, projects joining "META-REP" should focus on issues of replicability specifically in the behavioral, social, and cognitive sciences. We believe that such a focus in necessary to keep "META-REP" as coherent as possible while, at the same time, taking different data structures (such as experimental and observational, low and high dimensional data) and research practices in different disciplines into account. The disciplines mentioned here were selected because they have started treating "(non)replicability" as a phenomenon that warrants a meta-scientific explanation. In addition, these disciplines have already implemented measures to increase replication rates, which now require a systematic evaluation. Contributions from neighboring disciplines, such as philosophy of science, history of science, scientometrics, or the life sciences, are welcome if they explicitly adopt a meta-scientific perspective on replicability in the behavioral/social/cognitive sciences.

3.2 Delimitation of Scope Taking into Account the Duration of a Priority Program

The proposed "META-REP" program comprises two funding phases with three years each, that is, six years in total. In the first funding phase, we expect projects to focus more on the "what" question (describing and defining replication, Section 2.1) and the "why" question (explaining replication, Section 2.2), because a common definition of (non)replicability and a comprehensive understanding of its causes help develop appropriate measures to improve the status quo. Individual projects in the

second funding phase will focus more on the "how" question, that is, the evaluation of potential improvements and actions as described in Section 2.3.

3.3 Coherence of Planned Research Activities

First, all individual projects will have to adopt a meta-scientific focus on the (non)replicability of findings in the behavioral, social, and cognitive sciences. Project proposals will have to clearly indicate whether they intend to focus on the "what" question (see Section 2.1), the "why" question (see Section 2.2), or the "how" question (see Section 2.3). The list of potential applicants and tentative project titles in Table 2 shows that this is possible and feasible.

Second, to facilitate mutual cooperation and exchange between individual projects in this Priority Program, a collaborative groupware environment will be developed and operated by the Leibniz Center for Psychological Research (ZPID) at the University of Trier. This scientific groupware platform will allow all research groups to easily create, share, synchronize, and discuss all kinds of research-related digital objects (e.g., raw data, experimental control- and statistics scripts, software and documentation) between the partners, tightly integrated with collaboration and communication tools. Following an accepted publication, the research outcomes (e.g. data, scripts, supplements) and the documentation of all processes yielding these outcomes (e.g., communication history, protocol development steps) can be made available long-term as downloadable, annotated files in PsychArchives, ZPID's certified and GDPR-compliant repository. Sensitive research data, usually raw data, will also be stored in the same repository and made available through a secure access procedure for scientists (requiring in-person appointments to on-site data centers). Furthermore, the workspace and accompanying support infrastructures (developer located at ZPID) will facilitate and coordinate distributed lab projects to harness synergies between projects.

Third, all participating projects will be encouraged to pre-register their hypotheses, methods, and analysis plans (unless their research question is explicitly exploratory, or non-empirical) and to document and share their research data and materials according to the DFG guidelines. The goal of this tool is to render all digital research objects generated in "META-REP" projects reproducible, findable, accessible, interoperable, and reusable. The diverse data collection from this program, to be provided long-term for re-usability via PsychArchives operated by ZPID, will be a rich source for secondary data analysis that might answer substantive as well as meta-scientific research questions that go beyond the primary project goals. The usage of registered reports as publication format will be encouraged where appropriate to alleviate publication bias.

3.4 Strategies for Collaborating/Networking across Disciplines and Locations

In addition to the infrastructural aspects mentioned in the previous section, we will encourage (and even expect) collaborative efforts among individual projects and offer the possibility for

- ***Distributed research initiatives***, where individual projects may initiate "deep replication" attempts of a particular effect in a wide network of labs. Certain costs for such initiatives (e.g., for translating the materials into other languages, for participant reimbursement etc.) may – at least in part – be covered by the coordination project.
- ***Adversarial collaborations***, where dissenting teams/labs (that, for instance, disagree on whether low replication rates are due to inflated false-positive rates or to context-dependency; see Section 2.2) work together to test their respective hypothesis in a most rigid fashion (see Mellers, Hertwig, & Kahneman, 2001; or Kerr et al., 2018, for examples).

- ***Cumulative theory-building***, where scholars store a (fully specified) theory that they developed in a "theory database" (Glöckner et al., 2018; see also Section 2.3), allowing researchers from other projects to add boundary conditions or other specifications, and to test specific aspects of that theory in their own project.
- ***Data sharing and data pooling***, where all researchers involved in "META-REP" are encouraged to share and use the data generated in the program to the extent that it is informative and useful for their own project. This will be facilitated by a repository, respective tools, and data curation services provided by the requested professional curator to be located at ZPID in Trier.

Project proposals should specify which of the aforementioned forms of collaboration is feasible for them, and − roughly − what such a collaboration could look like. During the selection process of individual projects into the program, the program committee will keep potential opportunities for collaboration among projects in mind. In addition, a kick-off meeting for all PIs will be organized within the first month of the program, that is, right at the start. In this meeting, collaborations between projects will be explicitly elaborated by the PIs.

In addition to these specific measures, networking strategies that have been effective in previous Priority Programs will be implemented in the presently proposed program:

- ***Regular meetings*** (once a year, with a kick-off meeting at the beginning of each funding period) among all PIs and scientific personnel employed in the respective projects, to take place at LMU Munich (2 days each, 3 meetings in total during the first funding phase)
- ***Regular retreats*** (in Years 1 and 2 of each funding period) for all PIs and scientific personnel employed in the projects (4 days each). During these retreats, pre- and post-doctoral researchers in the projects will have the opportunity to present the current status of their projects, to receive professional feedback from all PIs, and to "swap papers" with their peers. In addition, external guests (e.g., international collaboration partners, journal editors, representatives from scientific organizations, etc.) will be invited to give talks and participate in discussions.
- ***Program-specific workshops*** (approx. 2 per year, i.e., 6 in total during the first funding period) specifically designed for the current program. Three workshops will deal with meta-scientific methodological or philosophical aspects or with similar "replicability debates" in other disciplines, for instance, the natural sciences or the life sciences. These "interface workshops", to which researchers from other disciplines will be specifically invited, aim to build bridges between the sciences. Based on the insights from these "interface workshops", we will also discuss whether "META-REP" should be thematically broadened in the second funding phase. In addition to these "interface workshops", other workshops will be specifically designed for pre- and post-doctoral researchers in the program. One workshop per funding period will be on "good scientific practice" (following the curriculum developed by the "*Ombudsman für die Wissenschaft*").
- ***International Conferences*** on Replication Science (at the end of the first and the second funding period, i.e., Years 3 and 6). At this conference, projects will present their findings to a wider, and more interdisciplinary audience. Researchers, representatives from scientific organizations (such as the APA or the APS) and from funding agencies (such as the European Commission), science journalists, and other interested individuals will be invited to discuss the state of replication research and the situation of behavioral, social and cognitive science as a whole in a broader context. Thus, these conferences aim to disseminate the findings obtained in the Priority Program to the general public. They will be organized in collaboration with the Center for Advanced Studies (CAS) at LMU Munich.

Finally, the coordination project will encourage and facilitate the development of interdisciplinary products coming out of this Priority Program, that is, edited books, special issues in high-impact scientific journals, symposia at international conferences, etc. This will attract the attention of an international audience and increase the program's visibility in the global scientific community.

3.5 Early Career Support, Promotion of Female Researchers, Family-friendly Policies

To support pre- and post-doc researchers involved in the program as scientific personnel, the following measures will be implemented and coordinated by the coordination project:

- ***Supervisory agreements*** are encouraged and, if necessary, provided for pre-doctoral researchers. Supervisory agreements are nowadays standard elements of PhD procedures. The coordination project can offer help and services to develop these agreements and to make them as comparable to each other (across projects) as possible.
- ***"Treasure box" funding***: Pre- and post-doctoral researchers in the projects can apply for extra funding to conduct special collaborative mini-projects for which no funds were applied in the project proposals. These mini-projects could be, for instance, the re-analysis of a dataset to test a specific hypothesis or to apply a special analytic method, the organization of a mini-conference with a specific thematic focus, the invitation of an international expert for a workshop, etc. Funds for these mini-projects are capped at 2,000 Euro each; these projects must be open to all researchers involved in the program, and at least two labs/teams must collaborate on these projects. Up to five of these mini-projects can be funded per funding period.
- ***Secondary supervision*** of doctoral theses by other PIs involved in the program is encouraged and facilitated by the coordination project. For post-doctoral researchers, secondary mentorships by other PIs in the program are encouraged. Researchers who want to spend some time (2-3 months) at their secondary supervisor's/mentor's lab can apply for special funding that covers the travel costs for such a "lab visit."
- ***Start-up funding***: Post-doctoral researchers and pre-doctoral researchers in their final year can apply for a stipend extending their project contract by 12 months. This stipend includes funds for travel and other expenses; during that time, we expect recipients to write their own grant proposal to the DFG ("*Antrag auf eigene Stelle*"). This allows them to develop their own independent research program, and to pursue and strengthen their academic career. In total, six start-up stipends will be awarded (three in each funding period).

To particularly support female researchers in their academic career, the following measures will be implemented and coordinated by the coordination project:

- Pre- and post-doctoral researchers can apply for funding for external mentorship (by a mentor of their choice) and for specific training programs focusing on gender-specific challenges, skill enhancement, networking, and leadership (open for all genders).
- Successful nationally and internationally visible female researchers will be invited to give talks and workshops.

To support researchers with children and to create a family-friendly environment in the proposed program, researchers with children can specifically apply for:

- special childcare funding (e.g. babysitters) if this childcare is necessary and takes place outside the regular opening hours of childcare facilities (according to DFG regulations),
- student research assistants who help researchers with young children better coordinate academic work and childcare,
- funds for professional childcare during the central events of the program, that is, annual meetings, annual retreats, workshops, and conferences.

These funds will be centrally administered by the coordination project. Thus, researchers can directly apply for these funds at the coordination project of the Priority Program. In addition, meetings, retreats, etc. will take place on family-friendly dates and times.

3.6 Networking of Planned Research Activities within the International Research System

A central website for the program will be set up and hosted at the ZPID in Trier. This website will inform readers about the aims and scope of "META-REP," the participating researchers, the individual projects, program-specific events, and scientific products that result from the program (i.e., publications, talks, tools, etc.). The content will be presented both in German and in English in order to attract and address not only a scientific community, but also a lay audience.

In addition, "META-REP" will make contact with stakeholders who may be interested in or profit from the research carried out in the program. This includes journal editors, representatives from scientific organizations as well as from funding agencies, science journalists, and other individuals inside and outside academia. These groups will be invited to the international conferences that will take place near the end of each funding period. In addition, individuals can be invited at other occasions, for instance, to give talks, workshops, or to participate in round-table discussions with the researchers involved in the program.

Finally, "META-REP" will establish relationships with scientific organizations that are specifically involved in the "replicability debate" in the behavioral, social and cognitive sciences, including the German Research Foundation (*Deutsche Forschungsgemeinschaft*, DFG), the Society for the Improvement of Psychological Science (SIPS), the American Psychological Society (APS), the American Academy for the Advancement of Science (AAAS), or the U.S. National Academy of Science (NAS). Besides submitting special symposia at conferences organized by these societies and organizations, the coordination project of "META-REP" will regularly inform representatives from these societies about new publications or findings and upcoming events. The goal is to contribute to and sustainably enrich international discussions about the replicability of behavioral, social and cognitive science research on the basis of the findings obtained in "META-REP."

4 References

Abele-Brehm, A., Gollwitzer, M., Steinberg, U., & Schönbrodt, F. (2019). Attitudes towards Open Science and public data sharing: A survey among members of the German Psychological Society. *Social Psychology, 50*, 252-260. http://doi.org/10.1027/1864-9335/a000384

Anderson, C. J., Bahník, S., Barnett-Cowan, M., Bosco, F. A., Chandler, J., ..., & Zuni, K. (2016). Response to comment on "Estimating the reproducibility of psychological science". *Science, 351*, 1037c. https://doi.org/10.1126/science.aad9163

Anvari, F., & Lakens, D. (2019). The replicability crisis and public trust in psychological science [Online First Publication]. *Comprehensive Results in Social Psychology*. https://doi.org/10.31234/osf.io/vtmpc

Asendorpf, J., Conner, M., De Fruyt, F., ..., Fiedler, S., …, & Wicherts, J. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27*, 108-119. https://doi.org/10.1002/per.1919

Auspurg, K., & Hinz, T. (2011). What fuels publication bias? Theoretical and empirical analyses of risk factors using the caliper test. *Journal of Economics and Statistics, 235*, 630-660. https://doi.org/10.1515/jbnst-2011-5-607

Auspurg, K., & Hinz, T. (2017). Social dilemmas in science: Detecting misconduct and finding institutional solutions. In B. Jann & W. Przepiorka (Eds.), *Social dilemmas, institutions, and the evolution of cooperation* (pp. 189-214). Berlin: De Gruyter Oldenbourg.

Auspurg, K., Hinz, T., & Schneck, A. (2014). Ausmaß und Risikofaktoren des Publication Bias in der deutschen Soziologie [Prevalence and Risk-Factors of Publication Bias in German Sociology.] *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 66*, 549-573. https://doi.org/10.1007/s11577-014-0284-3

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

Barba, L. A. (2018). Terminologies for reproducible research [Preprint]. arXiv:1802.03311. Retrieved from http://arxiv.org/abs/1802.03311

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature, 483*, 531-533. https://doi.org/10.1038/483531a

Benjamin, D. J., Berger, J. O., Johannesson, M., ..., Schönbrodt, F. D., ..., & Johnson, V. E. (2017). Redefine statistical significance. *Nature Human Behaviour, 2*, 6-10. https://doi.org/10.1038/s41562-017-0189-z

Bird, A. (2018). Understanding the replication crisis as a base rate fallacy. *The British Journal for the Philosophy of Science* [Preprint]. axy051. https://doi.org/10.1093/bjps/axy051

Brainerd, C. J., & Reyna, V. F. (2018). Replication, registration, and scientific creativity. *Perspectives on Psychological Science, 13*, 428-432. https://doi.org/10.1177/1745691617739421

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., ..., & Van't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology, 50*, 217-224. https://doi.org/10.1016/j.jesp.2013.10.005

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*, 365–376. https://doi.org/10.1038/nrn3475

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., ..., & Heikensten, E. (2016). Evaluating replicability of laboratory experiments in economics. *Science, 351*, 1433-1436. https://doi.org/10.1126/science.aaf0918

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

Carp, J. (2012). On the plurality of (methodological) worlds: Estimating the analytic flexibility of fMRI experiments. *Frontiers in Neuroscience, 6*, e149. https://doi.org/10.3389/fnins.2012.00149

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science, 2*, 115-144. https://doi.org/10.1177/2515245919847196

Christensen, G., & Miguel, E. (2018). Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature, 56*, 920-980. https://doi.org/10.1257/jel.20171350

Christensen, G., Freese, J., & Miguel E. (2019). *Transparent and reproducible social science research.* Oakland, CA: University of California Press.

Cottrell, R. C. (2014). Scientific integrity and the market for lemons. *Research Ethics*, *10*, 17-28. https://doi.org/10.1177/1747016113494651

Crandall, C. S., & Sherman, J. W. (2016). On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology, 66*, 93-99. https://doi.org/10.1016/j.jesp.2015.10.002

Cremers, H. R., Wager, T. D., & Yarkoni, T. (2017). The relation between statistical power and inference in fMRI. *PloS One, 12*, e0184923. https://doi.org/10.1371/journal.pone.0184923

De Boeck, P., & Jeon, M. (2018). Perceived crisis and reforms: Issues, explanations, and remedies. *Psychological Bulletin, 144*, 757-777. http://doi.org/10.1037/bul0000154

Dettweiler, U., Hanfstingl, B., Schröter, H., & Burzan, N. (2019). Epistemological and ethical aspects of research in the social sciences. *Frontiers in Psychology Research Topic* https://www.frontiersin.org/research-topics/5839/epistemological-and-ethical-aspects-of-research-in-the-social-sciences

Dougherty, M. R., & Horne, Z. (2019). Citation counts and journal impact factors do not capture research quality in the behavioral and brain sciences [Preprint]. https://doi.org/10.31234/osf.io/9g5wk

Dubben, H.-H., & Beck-Bornholdt, H.-P. (2005). Systematic review of publication bias in studies on publication bias. *BMJ, 331*, e433. https://doi.org/10.1136/bmj.38478.497164.F7

Duvendack, M., Palmer-Jones, R., & Reed, W. R. (2017). What is meant by "replication" and why does it encounter resistance in economics? *American Economic Review*, *107*, 46-51. https://doi.org/10.1257/aer.p20171031

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

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

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences, 113, 7900-7905*. https://doi.org/10.1073/pnas.1602413113

Erdfelder, E., & Ulrich, R. (2018). Zur Methodologie von Replikationsstudien [On the methodology of replication studies]. *Psychologische Rundschau, 69*, 13-21. https://doi.org/10.1026/0033-3042/a000387

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

Feest, U. (2019). Why replication is overrated. *Philosophy of Science* [Online First Publication]. http://doi.org/10.1086/705451

Feigenbaum, S., & Levy, D. M. (1993). The market for (ir)reproducible econometrics. *Social Epistemology, 7*, 215-232. https://doi.org/10.1080/08989629308573828

Fiedler, K. (2017). What constitutes strong psychological science? The (neglected) role of diagnosticity and a priori theorizing. *Perspectives on Psychological Science, 12*, 46-61. https://doi.org/10.1177/1745691616654458

Fiedler, K. (2018a). Wo sind die wissenschaftlichen Standards für hochwertige Replikationsstudien? [Where are the scientific standards for high-impact replication studies?]. *Psychologische Rundschau, 69*, 45–56. https://doi.org/10.1026/0033-3042/a000388

Fiedler, K. (2018b). The creative cycle and the growth of psychological science. *Perspectives on Psychological Science, 13*, 433-438. https://doi.org/10.1177/1745691617745651

Fiedler, K., & Prager, J. (2018). The regression trap and other pitfalls of replication science—Illustrated by the report of the Open Science Collaboration. *Basic and Applied Social Psychology, 40*, 115-124. https://doi.org/10.1080/01973533.2017.1421953

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

Fiedler, K., Kutzner, F., & Krueger, J. I. (2012). The long way from alpha-error control to validity proper: Problems with a short-sighted false-positive debate. *Perspectives on Psychological Science, 7*, 661-669. https://doi.org/10.1177/1745691612462587

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science, 345*, 1502-1505. https://doi.org/10.1126/science.1255484

Freese, J., & Peterson, D. (2017). Replication in social science. *Annual Review of Sociology, 43*, 147-165. https://doi.org/10.1146/annurev-soc-060116-053450

Fricker Jr., R. D., Burke, K., Han, X., & Woodall, W. H. (2019). Assessing the statistical analyses used in Basic and Applied Social Psychology after their p-value ban. *The American Statistician, 73*, 374-384. http://doi.org/10.1080/00031305.2018.1537892

Fuchs, H. M., Jenny, M., & Fiedler, S. (2012). Psychologists are open to change, yet wary of rules. *Perspectives on Psychological Science, 7*, 639-642. https://doi.org/10.1177/1745691612459521

Gergen, K. J. (1976). Social psychology, science and history. *Personality and Social Psychology Bulletin, 2*, 373-383. https://doi.org/10.1177/014616727600200409

Gergen, K. J. (1996). Social psychology as social construction: The emerging vision. In C. McGarty & A. Haslam (Eds.), *The message of social psychology: Perspectives on mind in society* (pp. 113-128). Oxford, UK: Blackwell.

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

Glöckner, A., & Betsch, T. (2011). The empirical content of theories in judgment and decision making: Shortcomings and remedies. *Judgment and Decision Making, 6*, 711-721.

Glöckner, A., Fiedler, S., & Renkewitz, F. (2018). Belastbare und effiziente Wissenschaft: Strategische Ausrichtung von Forschungsprozessen als Weg aus der Replikationskrise [Sound and efficient science: a strategic alignment of research processes as way out of the replication crisis]. *Psychologische Rundschau, 69*, 22-36. https://doi.org/10.1026/0033-3042/a000384

Gollwitzer, M., & Schwabe, J. (2019). Psychological Science in Year 5 of the "Replication Crisis:" A Treated Client in Search of a Diagnosis. *Manuscript submitted for publication*.

Gollwitzer, M., Schönbrodt, F. D., Steinberg, U., & Abele-Brehm, A. E. (2018). Die Datenmanagement-Empfehlungen der DGPs: Ein Zwischenstand (Bericht) [The DGPs Data Management Recommendations: Intermediate results (Report)]. *Psychologische Rundschau, 69*, 366-373. https://doi.org/10.1026/0033-3042/a000415

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translational Medicine, 8*, 341ps12. https://doi.org/10.1126/scitranslmed.aaf5027

Grahe, J. (2018). Another step towards scientific transparency: Requiring research materials for publication. *The Journal of Social Psychology, 158*, 1-6. https://doi.org/10.1080/00224545.2018.1416272

Hamermesh, D. S. (2007). Viewpoint: Replication in economics. *Canadian Journal of Economics*, *40*, 715-733. https://doi.org/10.1111/j.1365-2966.2007.00428.x

Hardwicke, T. E., & Shanks, D. R. (2016). Reply to Walker and Stickgold: Proposed boundary conditions on memory reconsolidation will require empirical verification. *Proceedings of the National Academy of Sciences*, *113*, e3993-e3994. https://doi.org/10.1073/pnas.1608235113

Hendriks, F., Kienhues, D., & Bromme, R. (2016). Disclose your flaws! Admission positively affects the perceived trustworthiness of an expert science blogger. *Studies in Communication Sciences, 16*, 124-131. https://doi.org/10.1016/j.scoms.2016.10.003

Henrich, J., Heine, S., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*, 61-83. https://doi.org/10.1017/S0140525X0999152X

Hong, Y.-W., Yoo, Y., Han, J., Wager, T. D., & Woo, C.-W. (2019). False-positive neuroimaging: Undisclosed flexibility in testing spatial hypotheses allows presenting anything as a replicated finding. *NeuroImage, 195*, 384–395. https://doi.org/10.1016/j.neuroimage.2019.03.070

Hussey, I., & Hughes, S. (2018). Hidden invalidity among fifteen commonly used measures in social and personality psychology [Preprint]. Retrieved from https://doi.org/10.31234/osf.io/7rbfp

Inbar, Y. (2016). Association between contextual dependence and replicability in psychology may be spurious. *Proceedings of the National Academy of Sciences, 113*, e4933-e4934. https://doi.org/10.1073/pnas.1608676113

Ingre, M., & Nilsonne, G. (2018). Estimating statistical power, posterior probability and publication bias of psychological research using the observed replication rate. *Royal Society Open Science, 5*, 181190. https://doi.org/10.1098/rsos.181190

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

Ioannidis, J. P. A., Munafò, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: Detection, prevalence, and prevention. *Trends in Cognitive Sciences, 18*, 235–241. https://doi.org/10.1016/j.tics.2014.02.010

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

Jussim, L., Crawford, J. T., Anglin, S. M., Stevens, S. T., & Duarte, J. L. (2016). Interpretations and methods: Towards a more effectively self-correcting social psychology. *Journal of Experimental Social Psychology, 66*, 116-133. https://doi.org/10.1016/j.jesp.2015.10.003

Kaufman, J. C., & Glăveanu, V. P. (2018). The road to uncreative science is paved with good intentions: Ideas, implementations, and uneasy balances. *Perspectives on Psychological Science, 13*, 457-465. https://doi.org/10.1177/1745691617753947

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

Kerr, N. L., Ao, X., Hogg, M. A., & Zhang, J, (2018). Addressing replicability concerns via adversarial collaboration: Discovering hidden moderators of the minimal intergroup discrimination effect. *Journal of Experimental Social Psychology, 78*, 66-76. https://doi.org/10.1016/j.jesp.2018.05.001

Kidwell, M. C., Lazarević, L. B., Baranski, E., ..., Fiedler, S., & Nosek, B. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLoS Biology*, *14*, e1002456. https://doi.org/10.1371/journal.pbio.1002456

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., ..., & Cemalcilar, Z. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology, 45*, 142-152. https://doi.org/10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr., R. B., ..., & Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science, 1,* 443-490. https://doi.org/10.1177/2515245918810225

Lakatos, I. (1976). Falsification and the methodology of scientific research programmes. In S. G. Harding (Ed.), *Can theories be refuted? Essays on the Duhem-Quine thesis, Synthese Library* (pp. 205–259). Dordrecht: Springer Netherlands. http://doi.org/10.1007/978-94-010-1863-0_14

LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A unified framework to quantify the credibility of scientific findings. *Advances in Methods and Practices in Psychological Science*, *1*, 389-402. https://doi.org/10.1177/2515245918787489

Lieb, K., von der Osten-Sacken, J., Stoffers-Winterling, J., Reiss, N., & Barth, J. (2016). Conflicts of interest and spin in reviews of psychological therapies: a systematic review. *BMJ Open, 6*, e010606. http://doi.org/10.1136/bmjopen-2015-010606

Lonsdorf, T. B., Klingelhöfer-Jens, M., Andreatta, M., Beckers, T., Chalkia, A., ..., & Merz, C. J. (2019). *How to not get lost in the garden of forking paths: Lessons learned from human fear conditioning research regarding exclusion criteria* [Preprint]. PsyArXiv, https://doi.org/10.31234/osf.io/6m72g

Lonsdorf, T. B., Menz, M. M., Andreatta, M., Fullana, M. A., Golkar, A., ..., & Drexler, S. M. (2017). Don't fear 'fear conditioning': Methodological considerations for the design and analysis of studies on human fear acquisition, extinction, and return of fear. *Neuroscience & Biobehavioral Reviews, 77*, 247-285. https://doi.org/10.1016/j.neubiorev.2017.02.026

Lonsdorf, T. B., Merz, C. J., & Fullana, M. A. (2019). Fear extinction retention – is it what we think it is? *Biological Psychiatry, 85*, 1074-1082. https://doi.org/10.1016/j.biopsych.2019.02.011

Luck, S. J., & Gaspelin, N. (2017). How to get statistically significant effects in any ERP experiment (and why you shouldn't). *Psychophysiology, 54*, 146–157. https://doi.org/10.1111/psyp.12639

Magraw-Mickelson, Z., & Gollwitzer, M. (2018). Relational and group collective self responses to observed victimization across cultures. *Social Justice Research, 31*, 113-132. https://doi.org/10.1007/s11211-018-0304-y

Marsman, M., Schönbrodt, F. D., Morey, R. D., Yao, Y., Gelman, A., & Wagenmakers, E.-J. (2017). A Bayesian bird's eye view of 'Replications of important results in social psychology'. *Royal Society Open Science, 4*, 160426. https://doi.org/10.1098/rsos.160426

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

Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science, 12*, 269-275. https://doi.org/10.1111/1467-9280.00350

Merton, R. K. (1957). Priorities in scientific discovery: A chapter in the sociology of science. *American Sociological Review, 22*, 635-659. https://doi.org/10.2307/2089193

Miller, J. & Ulrich, R. (2016). Optimizing research payoff. *Perspectives on Psychological Science, 11*, 664-691. https://doi.org/10.1177/1745691616649170

Morey, R. D., Chambers, C. D., Etchells, R., ..., Schönbrodt, F. D., ..., & Zwaan, R. A. (2016). The Peer Reviewers' Openness Initiative: Incentivising open research practices through peer review. *Royal Society Open Science, 3*, e150547. https://doi.org/10.1098/rsos.150547

Moshontz, H., Campbell, L., Ebersole, C. R., ..., Fiedler, S., …, & Chartier, C. R. (2018). The psychological science accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science*, *1*, 501-515. https://doi.org/10.1177/2515245918797607

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., ..., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behavior, 1*, e0021. https://doi.org/10.1038/s41562-016-0021

Munoz, J., & Young, C. (2018). We ran 9 billion regressions: Eliminating false positives through computational model robustness. *Sociological Methodology, 48*, 1-33. https://doi.org/10.1177/0081175018777988

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behavior, 3*, 221-229. https://doi.org/10.1038/s41562-018-0522-1

National Academies of Sciences, Engineering, and Medicine (2019). *Reproducibility and Replicability in Science*. Washington, DC: The National Academies Press. https://doi.org/10.17226/25303.

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology*, *69*, 511-534. https://doi.org/10.1146/annurev-psych-122216-011836

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

Nosek, B. A., & Errington, T. M. (2019). What is replication? [Preprint]. https://doi.org/10.31222/osf.io/u4g6t

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115*, 2600-2606. https://doi.org/10.1073/pnas.1708274114

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

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985-2013). *Behavior Research Methods, 48*, 1205-1226. https://doi.org/10.3758/s13428-015-0664-2

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review* [Online First Publication]. https://doi.org/10.3758/s13423-019-01645-2

OECD (2015). *Making Open Science a Reality* [Online Document]. https://doi.org/10.1787/5jrs2f963zs1-en

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

Patil, P., Peng, R. D., & Leek, J. T. (2016). What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspectives on Psychological Science, 11*, 539-544. https://doi.org/10.1177/1745691616646366

Pettigrew, T. F. (2018). The emergence of contextual social psychology. *Personality and Social Psychology Bulletin, 44*, 963-971. https://doi.org/10.1177/0146167218756033

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., ..., & Yarkoni, T. (2017). Scanning the horizon: Towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience, 18*, 115-126. https://doi.org/10.1038/nrn.2016.167

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery, 10*, 712. https://doi.org/10.1038/nrd3439-c1

Renkewitz, F., Fuchs, H. M., & Fiedler, S. (2011). Is there evidence of publication biases in JDM research? *Judgment and Decision Making, 6*, 870-881. http://journal.sjdm.org/11/m27/m27.pdf

Rohrer, J. M., Egloff, B., & Schmukle, S. C. (2017). Probing birth-order effects on narrow traits using specification-curve analysis. *Psychological Science, 28*, 1821-1832. https://doi.org/10.1177/0956797617723726

Sakaluk, J. K., & Graham, C. A. (2018). Promoting transparent reporting of conflicts of interests and statistical analyses at The Journal of Sex Research. *The Journal of Sex Research, 55*, 1-6. https://doi.org/10.1080/00224499.2017.1395387

Schönbrodt, F. D., Maier, M., Heene, M., & Bühner, M. (2018). Forschungstransparenz als hohes wissenschaftliches Gut stärken: Konkrete Ansatzmöglichkeiten für Psychologische Institute [Fostering research transparency as a high scientific good: Specific approaches for psychological departments]. *Psychologische Rundschau, 69*, 37-44. https://doi.org/10.1026/0033-3042/a000386

Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. *BMJ, 340*, c332. https://doi.org/10.1136/bmj.c332

Silberzahn, R., Uhlmann, E. L., Martin, D. P., ..., Schönbrodt, F. D., ..., & Nosek, B. A. (2018). Many analysts, one dataset: Making transparent how variations in analytical choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*, 337-356. https://doi.org/10.1177/2515245917747646

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

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science, 26*, 559-569. https://doi.org/10.1177/0956797614567341

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Specification curve: Descriptive and inferential statistics on all reasonable specifications [Online Document]. http://doi.org/10.2139/ssrn.2694998

Sjouwerman, R., Niehaus, J., Kuhn, M., & Lonsdorf, T. B. (2016). Don't startle me – Interference of startle probe presentations and intermittent ratings with fear acquisition. *Psychophysiology, 53*, 1889-1899. https://doi.org/10.1111/psyp.12761

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science, 3*, e160384. https://doi.org/10.1098/rsos.160384

Spellman, B., Gilbert, E. A., & Corker, K. S. (2017). Open Science: What, why, and how [Preprint]. https://doi.org/10.31234/osf.io/ak6jr

Stanley, D. J., & Spence, J. R. (2014). Expectations for replications: Are yours realistic? *Perspectives on Psychological Science, 9*, 305–318. https://doi.org/10.1177/1745691614528518

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science, 11*, 702–712. https://doi.org/10.1177/1745691616658637

Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A non-obtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology, 54*, 768–777. http://dx.doi.org/10.1037/0022-3514.54.5.768

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology, 15*, e2000797. https://doi.org/10.1371/journal.pbio.2000797

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

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

Turner, B. O., Paul, E. J., Miller, M. B., & Barbey, A. K. (2018). Small sample sizes reduce the replicability of task-based fMRI studies. *Communications Biology, 1,* 62. https://doi.org/10.1038/s42003-018-0073-z

Ulrich, R., Erdfelder, E., Deutsch, R., Strauß, B., Brüggemann, A., Hannover, B., ... & Rief, W. (2016). Inflation von falsch-positiven Befunden in der psychologischen Forschung. *Psychologische Rundschau, 67*, 163-174. https://doi.org/10.1026/0033-3042/a000296.

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

Vazire, S. (2018). Implications of the credibility revolution for productivity, creativity, and progress. *Perspectives on Psychological Science, 13*, 411-417. https://doi.org/10.1177/1745691617751884

Verhagen, J., & Wagenmakers, E. J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General, 143*, 1457-1475. https://doi.org/10.1037/a0038326

Vilhuber, L. (2019). Report by the AEA Data Editor. *AEA Papers and Proceedings, 109*, 718-729. https://doi.org/10.1257/pandp.109.718

Wacker, J. (2017). Increasing the reproducibility of science through close cooperation and forking path analysis. *Frontiers in Psychology, 8*, e1332. https://doi.org/10.3389/fpsyg.2017.01332

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

Wai, J., & Halpern, D. F. (2018). The impact of changing norms on creativity in psychological science. *Perspectives on Psychological Science, 13*, 466-472. https://doi.org/10.1177/1745691618773326

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PloS ONE, 6*, e26828. https://doi.org/10.1371/journal.pone.0026828

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology, 7*, e1832. https://doi.org/10.3389/fpsyg.2016.01832

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., ..., & Bouwman, J. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3. https://doi.org/10.1038/sdata.2016.18

Wilson, B. M., & Wixted, J. T. (2018). The prior odds of testing a true effect in cognitive and social psychology. *Advances in Methods and Practices in Psychological Science*, *1*, 186-197. https://doi.org/10.1177/2515245918767122

Wingen, T., Berkessel, J., & Englich, B. (2019). No replication, no trust? How low replicability influences trust in psychology [Preprint]. https://doi.org/10.31219/osf.io/4ukq5

Wong, V. C. & Steiner, P. M. (2018). *Replication designs for causal inference*. Working Paper. Retrieved from https://curry.virginia.edu/sites/default/files/uploads/epw/62_Replication_Designs.pdf

Wuttke, A. (2019). Why too many political science findings cannot be trusted and what we can do about it: A review of meta-scientific research and a call for academic reform. *Politische Vierteljahresschrift, 60*, 1-19. https://doi.org/10.1007/s11615-018-0131-7

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences, 41*, e120. https://doi.org/10.1017/S0140525X17001972