Psychonomic Bulletin & Review (2019) 26:1596–1618
https://doi.org/10.3758/s13423-019-01645-2

THEORETICAL REVIEW

Addressing the theory crisis in psychology

Klaus Oberauer[1] · Stephan Lewandowsky[2,3]

Published online: 12 September 2019

Abstract

A worrying number of psychological findings are not replicable. Diagnoses of the causes of this "replication crisis," and recommendations to address it, have nearly exclusively focused on methods of data collection, analysis, and reporting. We argue that a further cause of poor replicability is the often weak logical link between theories and their empirical tests. We propose a distinction between discovery-oriented and theory-testing research. In discovery-oriented research, theories do not strongly imply hypotheses by which they can be tested, but rather define a search space for the discovery of effects that would support them. Failures to find these effects do not question the theory. This endeavor necessarily engenders a high risk of Type I errors—that is, publication of findings that will not replicate. Theory-testing research, by contrast, relies on theories that strongly imply hypotheses, such that disconfirmation of the hypothesis provides evidence against the theory. Theory-testing research engenders a smaller risk of Type I errors. A strong link between theories and hypotheses is best achieved by formalizing theories as computational models. We critically revisit recommendations for addressing the "replication crisis," including the proposal to distinguish exploratory from confirmatory research, and the preregistration of hypotheses and analysis plans.

Keywords Replication · Scientific inference · Hypothesis testing · Computational modeling · Preregistration

Psychology has a problem. Over the past decade it has become clear that many findings, among them some deemed well established, are not replicable (Marsman et al., 2017; Open Science Collaboration, 2015). Numerous recommendations have been made for how to address this "replication crisis" (Asendorpf et al., 2013; Munafò et al., 2017). Virtually all these recommendations pertain to our methods of data collection, analysis, and publication. Here, we argue that, in addition to poor methods, the replication crisis is also due to the prevalence of theories that have only a weak logical relation to the hypotheses through which they are evaluated empirically. We suggest that the replication crisis is best resolved by focusing attention on the role of theorizing, and we do not believe that current recommendations that focus entirely on data generation are sufficient to overcome the crisis. To help clarify our argument, we summarize the intended meaning of some key terms in Table 1.

Scientific reasoning relies on inferences on two levels (see Fig. 1). On the first, the *empirical level*, we link hypotheses (e.g., X, Y, Z) to data (e.g., "x," "y," "z"). Most of our elaborate tools of inferential statistics serve to formalize the inductive inference from data to hypotheses: To the extent that an effect observed in a sample is significant (in classic null-hypothesis testing) or supported by a strong Bayes factor (in Bayesian statistical inference), we gain confidence that the effect is real—that is, it holds in the population from which we drew the sample. Credible hypotheses are empirical generalizations: Systematic relationships between variables that we believe to hold in the population—for instance, the Stroop effect (MacLeod, 1991), or the correlation between working-memory capacity and fluid intelligence (Conway, Kane, & Engle, 2003). The inductive inference from data to empirical generalizations is mirrored by a deductive inference: If an empirical generalization holds, we can predict that we will observe it whenever we test it with an appropriate study design. In other words, if an effect is real, we expect that we will be able to replicate it.

✉ Klaus Oberauer
k.oberauer@psychologie.uzh.ch

[1] Department of Psychology–Cognitive Psychology, University of Zurich, Binzmühlestrasse 14/22, 8050 Zürich, Switzerland

[2] University of Bristol, Bristol, UK

[3] University of Western Australia, Crawley, Australia

Table 1 Terminology

Term	Notation	Definition
Theory	T	An integrated set of propositions about latent (not directly observable) mechanisms, processes, and variables, and their causal relations to each other and to manifest (observable) variables
(Formal) model	T	A theory that is formalized, so that hypotheses can be derived from it through automatic derivation (e.g., logical proof, mathematical proof, computer simulation)
Inferential link	(none)	The relation between a set of premises (e.g., a theory) and a conclusion (e.g., a hypothesis). Inferential links vary in strength. Strength of an inferential link can be defined by how many auxiliary assumptions must be added as further premises to render the link deductive, and how credible these auxiliary assumptions are.
Empirical generalization	X, Y, Z	A statement describing a phenomenon, that is, an empirical regularity that holds generally, that is, for all members of a defined population, across a set of (not always well defined) situations, and across time
Hypothesis	X, Y, Z	The assumption that an empirical generalization holds
Prediction	X, Y, Z	An expectation of an empirical generalization not yet established by data (i.e., an expectation formulated a priori)
Explanation	(none)	An empirical generalization is explained by a theory to the degree that there exists a strong inferential link from the theory to the hypothesis that the empirical generalization must hold.
Observation, data	"x", "y", "z"	The result of a single study supporting X, Y, and Z, respectively.
True/false positive	(none)	A result that provides evidence for a true/false empirical generalization
Confirmatory diagnosticity	P(X|T)/P(X|¬T)	Diagnosticity of the confirmation of a hypothesis X for theory T (i.e., increase in credibility of theory T)
Disconfirmatory diagnosticity	P(¬X|T)/P(¬X|¬T)	Diagnosticity of the disconfirmation of a hypothesis X for theory T (i.e., decrease in credibility of theory T)

On a second level of inference, the *theoretical level*, we use theories (T) to derive hypotheses—often referred to as predictions[1]—which claim that some empirical generalizations X, Y, or Z are real. Established empirical generalizations, in turn, license inferences about theories, supporting them if they match the hypotheses derived from them, and questioning them if not.

As most of the method development in psychology focuses on formalizing the inductive inference on the empirical level, it is perhaps not surprising that much of the discussion on the causes of the replication crisis, and how to fix it, also concentrates on that level: Underpowered studies (Button et al., 2013), deficiencies of null-hypothesis significance testing (Wagenmakers, 2007), *p*-hacking (Simmons, Nelson, & Simonsohn, 2011), and publication bias (Ferguson & Heene, 2012) have been cited as reasons for the limited credibility of empirical generalizations we infer from data. There is much value in those critical reflections on this branch of the scientific reasoning cycle. At the same time, we argue that too little attention has been paid to weaknesses of our inferences on the theory level (for two commendable exceptions, see Fiedler, 2017; Muthukrishna & Henrich, 2019). We argue that those weaknesses at the theory level also contribute to the replication crisis.

[1] The term *prediction* has the connotation of foreseeing future events. Therefore, we use the term *hypothesis* for a statement of an empirical generalization inferred from a theory regardless of whether that statement is formulated before or after the empirical generalization has been established; we reserve the term *prediction* for the more limited case of hypotheses about empirical generalizations not yet known.

Discovery-oriented research and theory-testing research

To illuminate the problem, we outline the steps that are often involved in generating results that turn out to be nonreplicable. Step 1: Start from a theory that implies that a certain class of phenomena can be observed under some circumstances. For instance, take a theory of *embodiment priming* that we put

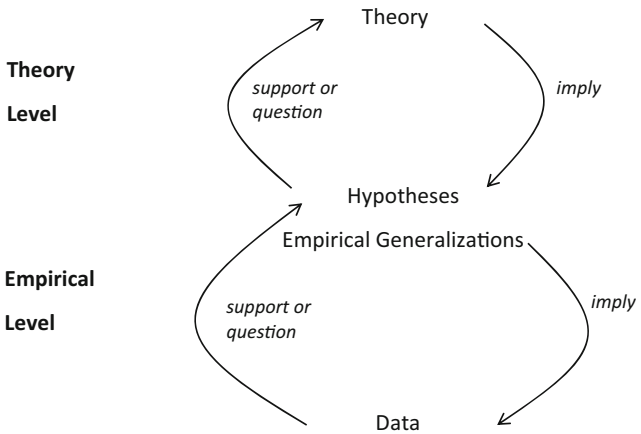

Fig. 1 Two levels of scientific inference. On the theory level, theories imply hypotheses. Hypotheses may be confirmed as empirical generalizations. Empirical generalizations support or question theories, depending on whether they match or mismatch the hypotheses derived from them. On the empirical level, empirical generalizations imply expectations for data from individual studies. Data support or question hypotheses. Well-supported hypotheses become empirical generalizations

forward to illustrate our case. We imbue the theory with the following core assumptions: (1) All abstract concepts are grounded in bodily states, sensations, or movements. (2) Experimentally inducing the bodily state, sensation, or movement in which a given concept is grounded activates (primes) that concept. (3) The activated concept influences behavior related to that concept.[2] This theory entails the hypothesis that inducing a state, sensation, or movement *can* lead to biases in judgments and decisions that depend on the concepts primed by this state, sensation, or movement.

Empirically testing this hypothesis involves search in a very large search space: There are many abstract concepts; for each of them, there are many ways in which it could be grounded in bodily states, sensations, and movements, and for each such presumed embodiment there are numerous ways in which it can be experimentally induced. Moreover, for each abstract concept there are judgments or decisions potentially influenced by it that one could investigate for the predicted bias. The combination of all these possibilities constitutes the space of possible tests of the embodiment priming theory (EPT). For instance, researchers could test the hypothesis that having people turn kitchen-paper rolls clockwise (as opposed to counterclockwise) activates their orientation toward the future, thereby priming the concept of novelty, so that they score higher on the personality scale "openness to experience" (Topolinski & Sparenberg, 2012). Or researchers could test whether people who are asked to briefly hold a cup of hot coffee subsequently rate another person as more "warm" compared with people who held a cup of cold coffee (Williams & Bargh, 2008).

The theory does not imply that the predicted bias will occur for each possible test in that space—it merely predicts that it occurs in *some*, arguably small, subset of possible tests. As a consequence, each individual test that confirms the predicted bias counts as evidence for the theory, but each individual test failing to show the predicted bias does not count as evidence against it—it merely shows that this particular combination of a concept, its assumed grounding, the chosen manipulation, and the chosen judgment is not an informative test of the theory. In that case, instead of revising the theory, researchers need to ask "what went wrong" with their study: They might have chosen the wrong judgment or decision as the dependent variable, they might have made the wrong assumptions about how exactly the concept in question is embodied, or they might have failed to induce the relevant bodily state. In any case, it is reasonable for the researcher to dismiss such a failure as uninformative and move on to another spot in the search space.

[2] Whereas our embodiment priming theory is inspired by several sources (e.g., Jostmann, Lakens, & Schubert, 2009; Körner, Topolinski, & Strack, 2015), we do not ascribe it to any particular author because it is meant to be a rational reconstruction of contemporary ideas rather than a faithful expression of one specific author's writing.

It might be tempting to dismiss this kind of research as flawed—especially in light of the fact that the examples cited above did not hold up in replication attempts (Lynott et al., 2014; Wagenmakers et al., 2015)—but that would be unjustified. Many respectable and useful research programs follow the same rationale, such as the search for exoplanets, the search for new drugs, or, closer to home, the search for neural correlates of a psychological phenomenon. We can call this endeavor *discovery-oriented* research.[3] What makes this kind of research generate nonreplicable results is the second step in the sequence: Conduct an extensive search through the vast space of possible tests; carry out each test only once; and evaluate the evidence from each test by the conventional standards of statistical inference (e.g., a p value <.05; Pashler & Harris, 2012). Those inferential standards were, however, designed for a different kind of endeavor, which we will call *theory-testing* research.

To explain the difference, we characterize the two kinds of research formally. We will use T for the theory in question, X for an empirical generalization that can be stated as a testable hypothesis (i.e., the proposition that a certain experimental effect or correlation exists in the population), and "x" as evidence from an individual study supporting hypothesis X (e.g., an experiment yielding a significant effect as expected from X). Table 2 presents all equations and a numerical example. Table 3 presents the corresponding equations for the case where a study yields disconfirming evidence "¬x" that speaks against hypothesis X.

Discovery-oriented research

In discovery-oriented research, the theory motivating it implies the existence of phenomena in a broad class Ω, of which X is an instance. For example, the EPT implies that the phenomenon of priming of abstract concepts through bodily states, sensations, or movements exists. The theory does not specify under which conditions this phenomenon will be observed—it merely motivates the search for it. Each test of that general expectation tests a specific hypothesis X—for instance, that people's scores on an "openness for experience" questionnaire will increase after they have rotated kitchen rolls in a clockwise direction. However, the theory implies no more than that a small subset of possible hypotheses X out of Ω are actually true (i.e., they describe real effects). In other words, for any given X, P(X|T) is small—to give a numerical example, let's say P(X|T) = 0.1. Confirming X as an empirical generalization is still diagnostic, supporting theory T, as long as the probability of such an effect being real, assuming the theory is false, is much lower, say P(X|¬T) = 0.02. These are the values we assume in the numerical example in Table 2. Note that in this example we assume that the confirmatory diagnosticity of X for T (as defined in Table 1; it can be computed as the ratio of rows 2 and 3 in Table 2) is equal in

[3] We borrow the term *discovery oriented* from Ioannidis (2005)

Table 2 Equations for inferences, with examples for discovery-oriented and theory-testing research

Term	Equation	Discovery oriented	Theory testing
Theory level			
Prior of theory being true / being false	$P(T)$, $P(\neg T) = 1 - P(T)$	.5	.5
Likelihood of hypothesis if theory is true	$P(X \mid T)$	.1	1
Likelihood of hypothesis if theory is false	$P(X \mid \neg T)$	.02	.2
Prior of hypothesis	$P(X) = P(X \mid T)P(T) + P(X \mid \neg T)P(\neg T)$	.06	.6
Posterior of theory, given hypothesis is true	$P(T \mid X) = \frac{P(X \mid T)P(T)}{P(X \mid T)P(T)+P(X \mid \neg T)P(\neg T)}$	$\frac{.1\times.5}{(.1\times.5+.02\times.5)} = .83$	$\frac{1\times.5}{(1\times.5+.2\times.5)} = .83$
Posterior of theory, given hypothesis is false	$P(T \mid \neg X) = \frac{P(\neg X \mid T)P(T)}{P(\neg X \mid T)P(T)+P(\neg X \mid \neg T)P(\neg T)}$	$\frac{.9\times.5}{(.9\times.5+.98\times.5)} = .48$	$\frac{0}{(0+.8\times.5)} = 0$
Empirical level			
Prior of hypothesis being true / being false	$P(X)$, $P(\neg X) = 1 - P(X)$	.06, .94	.6, .4
Likelihood of empirical support if hypothesis is true	$P(\text{"}x\text{"} \mid X) = 1 - \beta$	.8	.8
Likelihood of empirical support if hypothesis is false	$P(\text{"}x\text{"} \mid \neg X) = \alpha$	.05	.05
Posterior of hypothesis, given empirical support	$P(X \mid \text{“}x\text{”}) = \frac{P(\text{“}x\text{”} \mid X)P(X)}{P(\text{“}x\text{”} \mid X)P(X)+P(\text{“}x\text{”} \mid \neg X)P(\neg X)}$	$\frac{.8\times.06}{(.8\times.06+.05\times.94)} = .51$	$\frac{.8\times.6}{(.8\times.6+.05\times.4)} = .96$
Combining both levels			
Posterior of theory, given empirical support for hypothesis	$P(T \mid \text{"}x\text{"}) = P(T \mid X)P(X \mid \text{"}x\text{"}) + P(T \mid \neg X)P(\neg X \mid \text{"}x\text{"})$	.83 × .51 + .48 × .49 = .66	.83 × .96 + 0 = .80

the discovery-oriented and the theory-testing cases because we want to compare them free of confounds with unjustified differences. We do the same for other parameters, such as P(T), the prior probability that the theory is true.

Taken together, in discovery-oriented research the probability that X is a true hypothesis is low, regardless of whether T is true or not:

$$P(X) = P(X|T)P(T) + P(X|\neg T)P(\neg T).$$

Assuming equal priors for our theory to be true or not, P(T) = P(¬T) = 0.5, then P(X) = 0.06 (see Table 2). In other words, phenomena of the kind predicted by our theory have a low *base rate* of occurrence. Therefore, success in the search for these phenomena is sometimes hailed as demonstrating surprising, "counterintuitive" effects, with the added benefit of attracting attention and headlines.

If X is a true hypothesis, we can expect that a well-designed study provides evidence for it—which we call "x"—with

Table 3 Equations for inferences from disconfirming evidence

Term	Equation	Discovery oriented	Theory testing
Empirical level			
Prior of hypothesis being true / being false	$P(X)$, $P(\neg X) = 1 - P(X)$	.06, .94	.6, .4
Likelihood of disconfirming evidence if hypothesis is true	$P(\text{"}\neg x\text{"} \mid X) = \beta$	.2	.2
Likelihood of disconfirming evidence if hypothesis is false	$P(\text{"}\neg x\text{"} \mid \neg X) = 1 - \alpha$	.95	.95
Posterior of hypothesis, given disconfirming evidence	$P(X \mid \text{“}\neg x\text{”}) = \frac{P(\text{“}\neg x\text{”} \mid X)P(X)}{P(\text{“}\neg x\text{”} \mid X)P(X)+P(\text{“}\neg x\text{”} \mid \neg X)P(\neg X)}$	$\frac{.2\times.06}{(.2\times.06+.95\times.94)} = .01$	$\frac{.2\times.6}{(.2\times.6+.95\times.4)} = .24$
Combining both levels			
Posterior of theory, given evidence disconfirming the hypothesis	$P(T \mid \text{"}\neg x\text{"}) = P(T \mid X)P(X \mid \text{"}\neg x\text{"}) + P(T \mid \neg X)P(\neg X \mid \text{"}\neg x\text{"})$	.83 × .01 + .48 × .99 = .48	.83 × .24 + 0 = .20

Note. Equations for the theory level, and numerical values used for the examples, are the same as in Table 2

reasonably high probability—for instance P("x"|X) = 0.8. If X is false, there is still a chance that evidence supporting it is obtained due to sampling error and measurement error, say P("x"|¬X) = 0.05. The values we assigned to these two conditional probabilities are for illustrative purposes only, but they are not chosen arbitrarily. In the framework of null-hypothesis significance testing, P("x"|X) is the statistical power of a study, $1 - \beta$, and P("x"|¬X) is the criterion for significance, α, which sets the expected rate of false positives. The values 0.8 and 0.05 are representative of the power that is often recommended, and the conventional Type I error rate, respectively.

We can now use Bayes' rule to calculate the probability that a study yielding evidence "x" reflects a real effect X:

$$P(X|x) = \frac{P(\text{“x”}|X)P(X)}{P(\text{“x”}|X)P(X)P(\text{“x”}|\neg X) + P(\neg X)}.$$

The implications of Bayes' rule are easy to grasp when working through an example in frequency formats (Gigerenzer & Hoffrage, 1995): With the probabilities given above, out of 100 tests of our theory T—each testing a different hypothesis X out of the theory's search space—we can expect six to test a true effect, because P(X) = 0.06. Assuming P("x"|X) = 0.8, we can expect around five of them to provide evidence "x" for the effect. Of the remaining 94 tests aimed at nonexisting effects, we must expect approximately five to yield evidence "x" due to the false-positive rate P("x"|¬X) = 0.05 (we report rounded results in the text to facilitate exposition, whereas Table 2 contains results expressed to the second decimal digit). Hence, the proportion of true effects out of those tests that yielded evidence for an effect is around 5/10. In general, the lower the base rate P(X), the higher the posterior probability that an observation "x" speaking in favor of an effect is a false positive (Fiedler, 2017; Ioannidis, 2005; Miller, 2009).

Theory-testing research

Now, consider theory-testing research. This kind of research starts from a theory that provides strong inferential links for deriving hypotheses. Whereas a theory that is guiding discovery implies that X in Ω *can* be the case, a theory suitable for theory-testing research implies that, under conditions specified in the theory, X *must* be the case. Take, for example, a temporal-context theory of episodic memory such as SIMPLE (Brown, Neath, & Chater, 2007). This theory implies that extending the (filled or unfilled) delay between encoding and retrieval reduces the temporal distinctiveness of events (such as words in a memory list), which necessarily reduces the chance of accurate retrieval.[4] This hypothesis follows deductively from the core assumptions of temporal-context theories; in the case of SIMPLE, which formalizes these assumptions by a set of equations, they can be derived mathematically. This tight logical link between theory and hypothesis implies that establishing X as an empirical generalization speaks in favor of theory T, and conversely, empirically establishing that X is not true counts as evidence against T (e.g., see Lewandowsky, Duncan, & Brown, 2004, for evidence against the prediction from SIMPLE mentioned above). This is why we call this kind of research *theory testing*: It offers a chance to obtain strong evidence both in favor and against a theory.

Ideally, the hypothesis follows deductively from the theory, such that P(X|T) = 1. For confirmation of X to be diagnostic (see Table 1), we hope that P(X|¬T) << 1 (in the example in Table 2 we set this value to 0.2, keeping confirmatory diagnosticity the same as for discovery-oriented research). Now the prior probability of X is at a minimum equal to the prior of the theory. As long as the theory has a reasonably high prior—meaning that it is not highly implausible to begin with—this implies a fairly high base rate of hypothesis X. Assuming P(T) = 0.5 (the same value as for our discovery-oriented example), with P(X|T) = 1 and P(X|¬T) = 0.2, we obtain P(X) = 0.6 (see Table 2). This means we can expect that out of 100 tests of hypotheses that share the assumed characteristics of X, 60 are true. Of these, given our presumed power of 0.8, we can expect 48 to be supported by evidence "x." Of the remaining 40, we expect two to yield false positives. Hence, the posterior probability that an observation "x" reflects a true effect is 48/50, or 0.96.

Obviously, real cases are rarely that ideal, and the inferential link between theory T and hypothesis X is often less than perfect, so that P(X|T) < 1. We should therefore think of the distinction between discovery-oriented and theory-testing research as a continuum that varies with the strength of the inferential link from T to X; here, we focus on its extremes to clarify the distinction, and we show how the probabilities of interest vary as we vary P(X|T) continuously in Fig. 2. The left panel of Fig. 2 shows how the prior probability of a hypothesis, P(X), increases with the strength of the inferential link, P(X|T). The figure also shows the implications of that prior for the posterior probability of the hypothesis when supported by data, P(X|"x"), and in light of contradictory evidence, P(X|"¬x"), based on the equations and assumed parameters in Tables 2 and 3, respectively. Discovery-oriented research corresponds to values on the left of the *x*-axis, near zero, whereas theory-testing research corresponds to values nearer 1. The right panel of Fig. 2 shows how the two possible outcomes of a study ("x" or "not x") translate into the posterior probability of the theory being true, P(T|"x") and P(T|"¬x"), respectively, as a function of P(X|T). This plot shows how, as the logical link between theory and hypothesis becomes stronger, we learn more from both possible outcomes: If the data support the hypothesis, they provide stronger evidence for the theory, so that P(T|"x") rises more above the theory's prior, P(T) = 0.5 (indicated by the horizontal red line). Perhaps more striking is

[4] Assuming the words are still presented at the same rate.

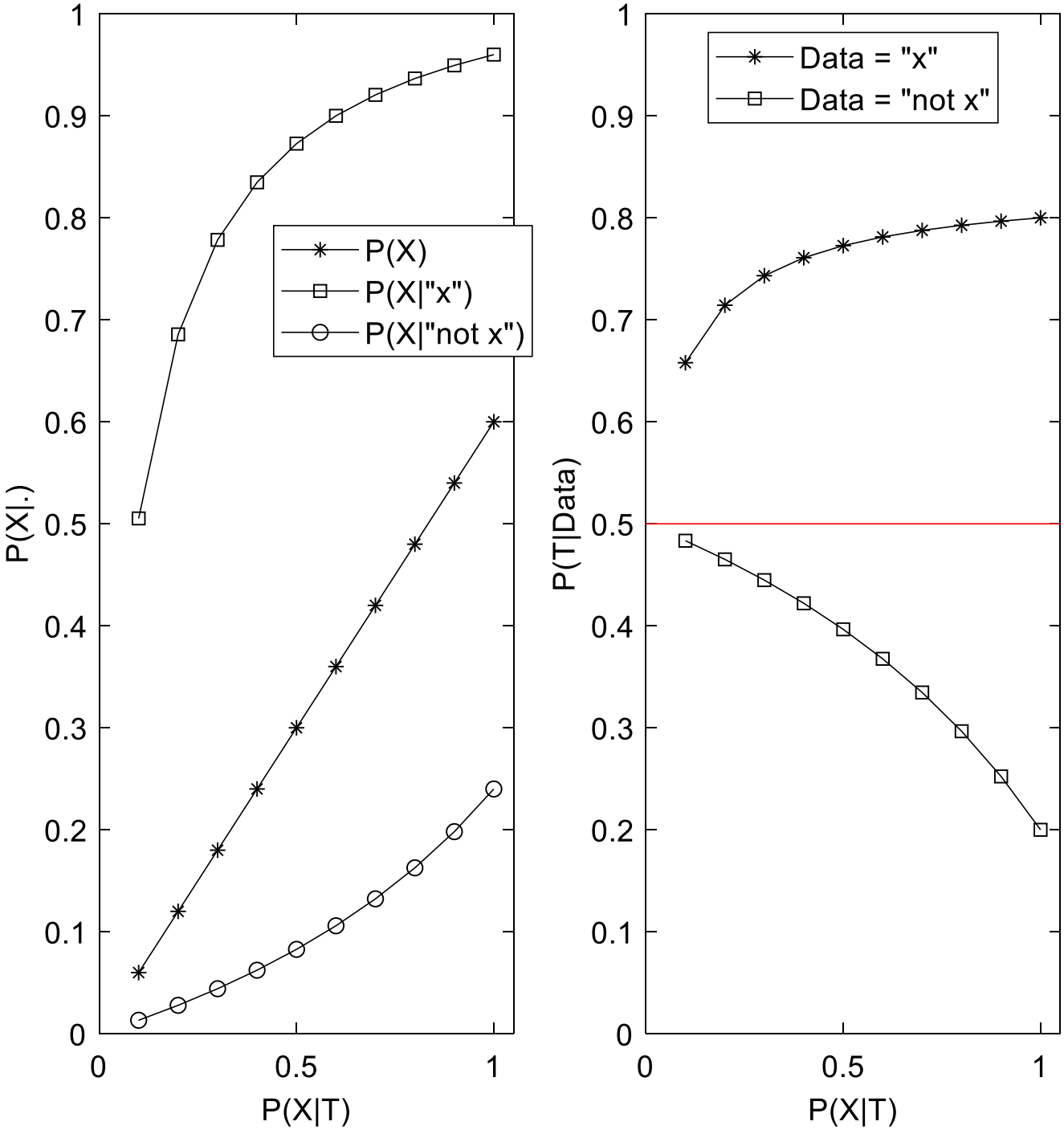

Fig. 2 Effects of continuously varying the strength of the inferential link from theory to hypothesis, P(X|T). Left: Prior probability of the hypothesis, P(X), and posterior probability of the hypothesis in light of confirming ("x") or disconfirming ("not x") data. Right: Posterior probability of theory, given confirming or disconfirming data. The red horizontal line demarcates the prior of the theory. Probabilities are calculated with the equations and parameter values given in Table 2. (Color figure online)

the pattern of P(T|"¬x"), showing that evidence against the hypothesis becomes increasingly informative (i.e., moving the posterior of the theory away from its prior) as P(X|T) becomes larger. We can use the disconfirmatory diagnosticity—that is, the informativeness of failed hypothesis tests (see bottom row of Table 1)—as the key criterion for placing a research endeavor on the continuum between discovery-oriented and theory-testing research by asking, To what extent does disconfirmation of a hypothesis derived from the theory count as evidence against the theory?

To conclude, applying conventional criteria of evidence for an effect, such as $\alpha = 0.05$ and $1 - \beta = 0.8$, results in an

arguably tolerable false-positive rate in the context of theory-testing research, but in a clearly unacceptable false-positive rate in the context of discovery-oriented research. Obviously, false positives are typically not replicable. One reason for the replication crisis in psychology, we suspect, is that a large part of psychological research is discovery oriented, but uses evidentiary criteria that are suited for theory-testing research, but are far too lax for discovery-oriented research.

For these reasons, the focus of the contemporary discussion on how to address the "replication crisis" by reducing the chance of Type I errors, or "false positives," is entirely justified for discovery-oriented research. Researchers engaging in theory-testing research, by contrast, should be more concerned with fully exploiting the evidence in the data. In particular, they should be concerned as much with establishing that a hypothesis is false as with establishing that it is true. The specific strength of theory-testing research is that we can leverage the disconfirmation of a hypothesis as evidence against any theory that entails it (right panel in Fig. 2). To make full use of this high disconfirmatory diagnosticity, we need methods to establish that a hypothesis is false. Null-hypothesis significance testing does not provide the tools for that purpose—we can at best fail to provide evidence against the null hypothesis. Bayesian model comparison, by contrast, enables researchers to gauge the evidence both for the alternative hypothesis—as with conventional frequentist statistics—and also—unlike frequentist statistics—for the null hypothesis (Wagenmakers, Marsman, et al., 2018b).

We now discuss some of the proposed remedies for the replication crisis within the framework of the two levels of inference. We explore the limitations of those remedies before turning to sketching a way forward that emphasizes the development of stronger theories over improvements to data collection and analysis.

Proposed remedies for the replication crisis

Many remedies have been proposed to address the replication crisis, among them: (1) More stringent statistical standards for inferring that an observed effect is real (Benjamin et al., 2018). (2) High-powered direct replications using the exact same materials and protocol as the original study, preferentially distributed across many labs to ensure generalization (e.g., Wagenmakers et al., 2016). (3) Open data, open materials, and open analysis algorithms. (4) A clear distinction between hypotheses formulated a priori (before looking at the data) and those formulated a posteriori—sometimes referred to as *HARKing* (Hypothesizing After Results Are Known; Kerr, 1998)—and, relatedly, a clear distinction between *exploratory* and *confirmatory* research (Wagenmakers, Dutilh, & Srafoglou, 2018a; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). (5) Preregistration of the hypotheses, the data collection plan (in particular, stopping rules for data collection), and the analysis plan to limit the "researcher degrees of freedom" (Simmons et al., 2011) in hypothesis formulation and in making analytical choices that invite HARKing and *p*-hacking, respectively.

Many of these proposals are helpful ideas for raising the standards of good research practice, primarily ensuring more trustworthy inferences on the empirical level of scientific inference—the level connecting observations to empirical generalizations (see Fig. 1). Our concern is that shoring up the strength of inferences on the empirical level does not by itself address deficits on the theory level—the level connecting empirical generalizations to theories. As we show next, the recommendations presently discussed are either irrelevant to the theory level or even misleading about it.

More stringent statistical standards and direct replication

An obvious remedy to reduce the rate of false positives is to raise the bar for declaring a discovery. An impressive lineup of scholars recently proposed to redefine statistical significance as a p value <.005 (Benjamin et al., 2018). Adopting a stricter criterion reduces false positives but also true positives, unless the loss of power is compensated for by an increase in sample size. Is this worth the price? Returning to our numerical examples: When a discovery-oriented research endeavor starts with a low prior probability of an effect, P(X) = 0.06, then setting α to 0.005 reduces the false-positive rate from 5 out of 94 to 0.5 out of 94, and assuming we invest the necessary resources to maintain power at 0.8 (i.e., through increasing the typical sample size by 70%; see Benjamin et al., 2018), we can still expect about 5 out of 100 true positive results. This means that the proportion of true positives among all positives becomes a reasonably acceptable 5 out of 5.5 (or 91%), a large improvement over the 5/10 (50%) we had with α = 0.05. In the case of theory-testing research, starting from P(X) = 0.6, we reduce the false positives from 2 to 0.2 out of 40. Again keeping power at 0.8, we can expect to improve the proportion of true positives among all positives from 48/50 (96%) to 48/48.2 (99%). Unless false positives have very high practical costs, this is arguably a negligible improvement, hardly worth the extra cost (see also Fiedler, Kutzner, & Krueger, 2012). Figure 3 summarizes the effects of reducing α on the posteriors of the hypothesis (left panel) and the posterior of the theory that entails it (right panel) after obtaining a significant result supporting the hypothesis. The figure shows that discovery-oriented research stands much to gain from reducing α—theory-testing research, not so much.

An analogous argument holds for direct replications. A single reasonably well-powered study conducted as part of discovery-oriented research tests a hypothesis with a low prior probability, and therefore empirical support for the hypothesis returns only a modest posterior probability for it—in our numerical example above (and in Table 2), that posterior would be P(X|"x") = 5/10. In that case, direct replication is indispensable

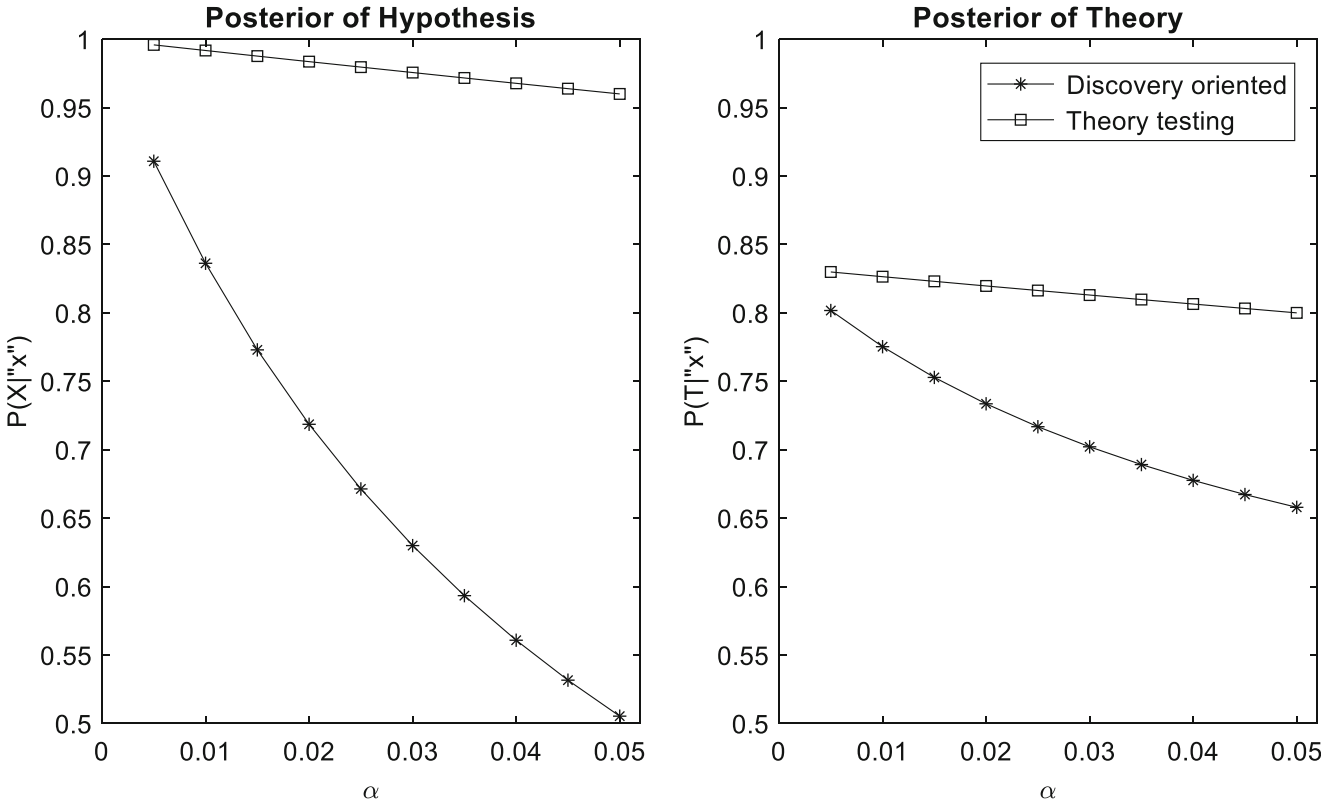

Fig. 3 Effects of reducing the alpha level for obtaining evidence "x" in favor of a hypothesis X by rejecting the null hypothesis (Simulations based on values in Table 2, varying alpha). With more stringent alpha levels, the posterior of both the hypothesis (left panel) and the theory that entails it (right panel) after obtaining a significant result "x" increase. For discovery-oriented research, this increase is steeper because the hypothesis starts from a lower prior

to gain a reasonable level of confidence that X is a real effect. In contrast, when the same study is carried out as part of a theory-testing endeavor, we start with a higher prior, and therefore obtain a higher posterior—in our numerical example, P(X|"x") = 48/50. A successful direct replication would still increase the posterior further, but not by much. The resources for a second study might therefore be better invested for testing a second hypothesis derived from the same theory.

Table 4 presents the equations for calculating the posterior probability of T, given "x_1" followed by a successful replication, "x_2," in comparison to the posterior of T, given "x_1" followed by observation "y" that supports a further hypothesis Y derived from the same theory. The likelihoods of this new prediction, P(Y|T) and P(Y|¬T), are assumed to be the same as those for X. For the numerical example illustrating theory-testing research, the posterior after replication, P(T|"x_1" + "x_2") = .83, whereas the posterior after successful test of a second prediction, P(T|"x_1" + "y") = .94. Note that for the example illustrating discovery-oriented research, by contrast, the direct replication yields not only a substantially increased posterior of X but also a slightly larger posterior of T than the test of a new prediction (.81 vs. .79). Figure 4 shows that this is a general result that holds regardless of the prior of the theory.

We explored the issue further with Fig. 5, which plots the advantage (or disadvantage, if negative) of direct replication compared with testing a second hypothesis, defined as the difference in the posterior of the theory achieved after two experiments. Here, we varied P(X|T) continuously, rather than focusing on the two extreme ends—discovery-oriented versus theory-testing research—and in addition varied the confirmatory diagnosticity of the hypothesis—that is, the ratio of P(X|T) to P(X|¬T). Figure 5 shows that, with the majority of constellations explored here, a successful test of a second hypothesis yields more evidence for the theory than a successful direct replication (i.e., most data points fall below the zero line, indicating that a direct replication was less beneficial than testing a second hypothesis).

Specifically, whenever the links between theory and hypotheses, P(X|T) and P(Y|T), are at least moderately strong, or the confirmatory diagnosticity of the hypotheses is medium to low, then the theory stands to gain more support from a successful test of a second hypothesis derived from it than by a successful direct replication of the first test. In Fig. 5, this refers to values below the red lines, which cluster toward the left (lower diagnosticity) and involve stronger links between theory and hypotheses—higher values of P(X|T) (see legend). The exception is the test of highly

Table 4 Posterior probabilities of theories after a direct replication and after a test of a second hypothesis

Term	Equation	Discovery oriented	Theory testing
Knowledge after first experiment (from Table 2)			
Prior of theory after first finding	$P(T \mid \text{"}x_1\text{"})$	.66	.80
Prior of hypothesis X after first finding	$P(X \mid \text{"}x_1\text{"})$	.51	.96
Posterior of theory, given hypothesis is true	$P(T \mid X)$, $P(T \mid Y)$	.83	.83
Posterior of theory, given hypothesis is false	$P(T \mid \neg X)$, $P(T \mid \neg Y)$	.48	0
Likelihood of empirical support if hypothesis is true	$P(\text{"}x\text{"} \mid X) = P(\text{"}y\text{"} \mid Y) = 1 - \beta$	.8	.8
Likelihood of empirical support if hypothesis is false	$P(\text{"}x\text{"} \mid \neg X) = P(\text{"}y\text{"} \mid \neg Y) = \alpha$	.05	.05
Direct replication			
Posterior of hypothesis, given successful replication	$P(X \mid \text{“}x_1\text{”} + \text{“}x_2\text{”}) = \frac{P(\text{“}x_2\text{”} \mid X)P(X \mid \text{“}x_1\text{”})}{P(\text{“}x_2\text{”} \mid X)P(X \mid \text{“}x_1\text{”}) + P(\text{“}x_2\text{”} \mid \neg X)P(\neg X \mid \text{“}x_1\text{”})}$	$\frac{.8\times.51}{(.8\times.51+.05\times.49)} = .94$	$\frac{.8\times.96}{(.8\times.96+.05\times.04)} \approx 1.0$
Posterior of theory, given all outcomes of all empirical tests	$P(T \mid \text{"}x_1\text{"} + \text{"}x_2\text{"}) = P(T \mid X)P(X \mid \text{"}x_1\text{"} + \text{"}x_2\text{"}) + P(T \mid \neg X)P(\neg X \mid \text{"}x_1\text{"} + \text{"}x_2\text{"})$	$.83 \times .94 + .48 \times .06 = .81$	$.83 \times 1.0 + 0 = .83$
Test of new hypothesis			
Likelihood of new hypothesis if theory is true	$P(Y \mid T)$	.1	1
Likelihood of new hypothesis if theory is false	$P(Y \mid \neg T)$	.02	.2
Prior of new hypothesis, given current prior of theory	$P(Y) = P(Y \mid T)P(T \mid \text{"}x\text{"}) + P(Y \mid \neg T)P(\neg T \mid \text{"}x\text{"})$	$.1 \times .66 + .02 \times .34 = .07$	$1 \times .80 + .2 \times .20 = .84$
Posterior of new hypothesis, given empirical support for new hypothesis	$P(Y \mid \text{“}y\text{”}) = \frac{P(\text{“}y\text{”} \mid Y)P(Y)}{P(\text{“}y\text{”} \mid Y)P(Y) + P(\text{“}y\text{”} \mid \neg Y)P(\neg Y)}$	$\frac{.8\times.07}{(.8\times.07+.05\times.94)} = .56$	$\frac{.8\times.84}{(.8\times.84+.05\times.4)} = .99$
Posterior of theory, given empirical support for old hypothesis, and new hypothesis is true	$P(T \mid \text{“}x_1\text{”} + Y) = \frac{P(Y \mid T)P(T \mid \text{“}x_1\text{”})}{P(Y \mid T)P(T \mid \text{“}x_1\text{”}) + P(Y \mid \neg T)P(\neg T \mid \text{“}x_1\text{”})}$	$\frac{.1\times.66}{(.1\times.66+.02\times.34)} = .91$	$\frac{1\times.80}{(1\times.80+.2\times.20)} = .95$
Posterior of theory, given support for first hypothesis, and new hypothesis is false	$P(T \mid \text{“}x_1\text{”} + \neg Y) = \frac{P(\neg Y \mid T)P(T \mid \text{“}x_1\text{”})}{P(\neg Y \mid T)P(T \mid \text{“}x_1\text{”}) + P(\neg Y \mid \neg T)P(\neg T \mid \text{“}x_1\text{”})}$	$\frac{.9\times.66}{(.9\times.66+.98\times.34)} = .64$	$\frac{0\times.80}{(0\times.80+.8\times.20)} = 0$
Posterior of theory, given support for first and second hypothesis	$P(T \mid \text{"}x_1\text{"} + \text{"}y\text{"}) = P(T \mid \text{"}x_1\text{"} + Y)P(Y \mid \text{"}y\text{"}) + P(T \mid \text{"}x_1\text{"} + \neg Y)P(\neg Y \mid \text{"}y\text{"})$	$.91 \times .56 + .64 \times .44 = .79$	$.95 \times .99 + 0 = .94$

“counterintuitive” hypotheses typical for discovery-oriented research: This kind of hypothesis is not derived from the theory but merely motivated by it, so P(X|T) is low, and at the same time the hypothesis is highly diagnostic, because if the theory were false, its prior probability would be close to zero. As a result, its overall prior is very low, so it has much to gain from direct replication. This corresponds to values above the red lines in Fig. 5, which arise with high diagnosticity and low values of P(X|T) only.

All this is not to say that direct replications are useless. Obviously, when our goal is to establish whether or not an empirical generalization for which we have initial evidence is real, then direct replication of the initial study is the only way to achieve that. As long as our focus is on establishing reliable facts at the empirical level, direct replication remains the gold standard. When our primary interest is with establishing which theories are credible, however, we find that often testing new hypotheses to evaluate a theory—provided the

theory strongly implies the hypotheses—is a better investment of our resources than a direct replication.

"Exploratory" and "confirmatory" research revisited

A further common recommendation to address the replication crisis is to clearly distinguish between *exploratory* and *confirmatory* research, with an understanding that only the latter can provide strong evidence for a hypothesis (Wagenmakers, et al., 2018a; Wagenmakers et al., 2012). This recommendation is usually coupled with the appeal to *preregister* hypotheses and analysis plans. The distinction between exploratory and confirmatory research is perhaps reminiscent of our distinction between discovery-oriented and theory-testing research introduced above, but the exploratory-confirmatory contrast is usually defined in a different and, we argue, unhelpful way. Research is regarded as confirmatory if and only if hypotheses and data analysis plan are fixed before looking at the data, whereas hypotheses and analysis decisions that are chosen after looking at the data, and perhaps in response to characteristics of the data, count as exploratory. Exploratory research is criticized for being vulnerable to intentional or unintentional confirmation bias: When the data inform which hypothesis to test, or which combination of data transformation and statistical analysis procedure to choose, then researchers keen on finding reportable effects are tempted to choose hypotheses and analysis plans likely to yield confirmation of an expected effect. A related argument is the critique of HARKing, the practice of presenting a post hoc hypothesis that was formulated after looking at the data as if it were an a priori hypothesis that has been formulated before looking at the data (Kerr, 1998).

This critique is certainly valid. What we find unhelpful about it is its emphasis on the temporal order in which a researcher specifies their hypothesis and analysis plan on the one hand, and interrogates the data on the other (see also Rubin, 2017b, for a similar critique). Fixing hypotheses and analysis plans *before* analyzing the data is regarded as good practice; reversing the order is considered bad practice. Preregistration is argued to address the problem because it enforces the right temporal order of these actions. This way of framing the problem and the solution remains superficial because it uses a distinction that does not matter—temporal order—as a proxy for one that does matter—the distinction between justified and arbitrary choices of hypotheses and analysis procedures.

The question of whether temporal order matters in establishing the validity of hypotheses or theories is known as the *paradox of predictivism* in philosophy of science (Barnes, 2008). The paradox arises from a contradiction between two strong intuitions. One is that a theory receives stronger confirmation from the prediction of a novel finding—not known to the theorist at the time of formulating the theory—than from the explanation of an already known finding. The other intuition is that the evidential value of a finding for a theory should not depend

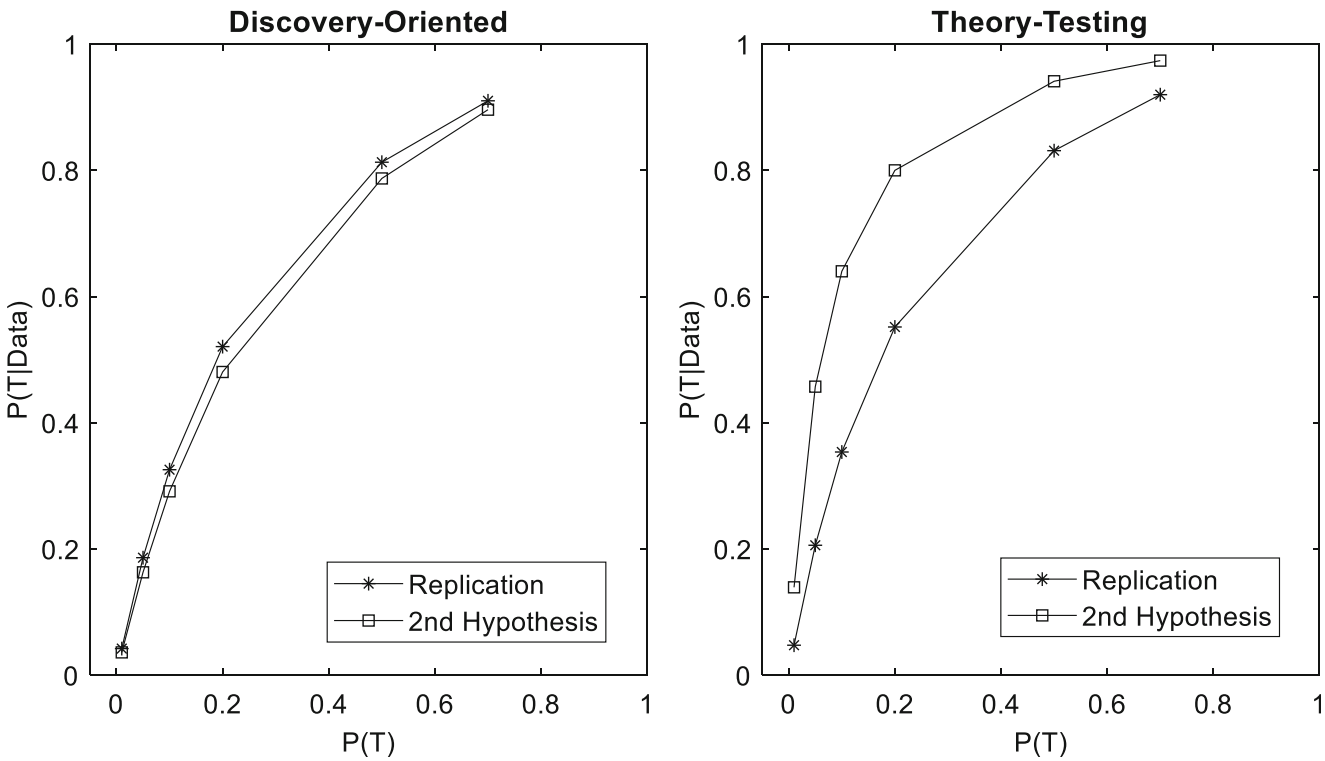

Fig. 4 Posterior of theory T, given data from an initial finding "x_1" confirming hypothesis X, and either a replication of that finding, "x_2", or data "y" confirming a second hypothesis, Y

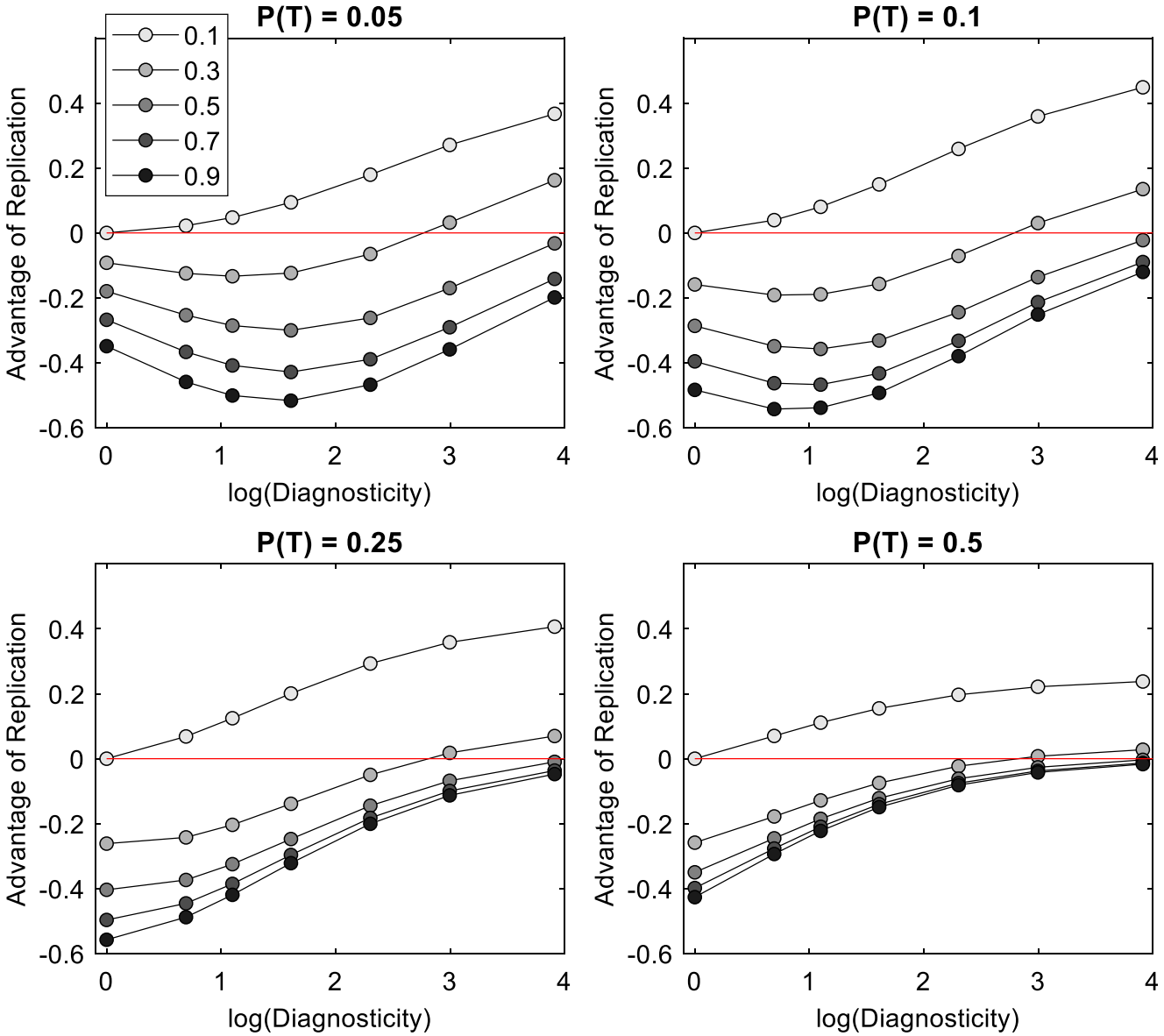

Fig. 5 Advantage of replication, defined as P(T|"x1" & x2") – P(T|"x1" & "y"). Each line represents one level of P(X|T), and the *x*-axis represents the log of the confirmatory diagnosticity of the tested hypothesis. Confirmatory diagnosticity, defined as P(X|T)/P(X|¬T), was varied over values [1, 2, 3, 5, 10, 20 50]. We set the diagnosticity equal for both X and Y, so the *x*-axis represents log(P(X|T) / P(X|¬T)) = log(P(Y|T) / P(Y|¬T))

on historical accidents such as when a theorist first learned about an empirical finding relative to when she first thought of a theory that predicts or explains that finding. The notion that the history of the researcher's state of mind should determine to what degree some piece of empirical evidence supports a given theory is generally regarded as unacceptable in the philosophy of science. To appreciate why, consider the following scenarios:

Scenario A: Researcher A designs the following test of the embodiment priming theory (EPT) sketched above: A sample of participants is divided randomly into two groups. Upon entering the laboratory, members of the experimental group are asked to leap over a gap in the floor; members of the control group are tested in a second lab that is identical to the first, except that there is no gap to cross. Researcher A reasons that the activity of the experimental group primes the concept of a "leap of faith," and therefore this group should, on average, score higher on a subsequent religiosity questionnaire than the control group. Researcher A preregisters this prediction, together with a straightforward *t* test as analysis plan, then runs the experiment and finds a significant effect in the expected direction.

Scenario B: Researcher B runs an experiment on paired-associates memory: Participants encode lists of four to 12 word pairs; after each list, their memory is tested by presenting one word from each pair, and they are asked to reproduce the other element. Upon exploring the data, Researcher B notices an unanticipated pattern: When participants confuse the

correct word with a word from another pair in the memory list, it is more likely to be a pair close to the correct one in the presentation order than one further away. After some reading and discussion with colleagues, Researcher B finds out that this regularity is predicted by a class of episodic-memory models in which the temporal context of events acts as an important retrieval cue for the events, such as SIMPLE (Brown et al., 2007) and TCM (Sederberg, Howard, & Kahana, 2008).

Researcher A did everything by the books to claim that she did confirmatory research. In contrast, Researcher B made his discovery through exploratory research, and could be accused of HARKing when claiming to have obtained empirical support for temporal-context theories of memory. Yet we argue that the evidence supports the theory in question more strongly in Scenario B than in Scenario A.

We designed Scenario A to fit the blueprint of discovery-oriented research: The inferential link between the theory and the expected effect is weak; it requires a number of questionable auxiliary assumption to be turned into a deductive link: "If EPT is true, and if the concept of a 'leap of faith' is embodied in the physical action of jumping over a gap, and if jumping over a gap once primes that concept for at least a few minutes, and if priming that concept shifts people's responses on a questionnaire toward higher religiosity scores, then the experimental group will show higher religiosity scores (X_a)." Each of the auxiliary assumptions could fail, so that failure to observe "x_a" (i.e., obtaining a nonsignificant group difference in religiosity scores, or even a Bayes factor strongly in favor of the null hypothesis) does not count as evidence against EPT, whereas observing "x_a" counts as evidence not only for EPT but also for all auxiliary assumptions. For the same reason, P(X|EPT) is arguably low, and as we have shown above, that implies that P(X_a|"x_a") is not very high, either.

In contrast, we regard Scenario B as a typical case of theory-testing research: The class of temporal-context theories of episodic memory (TCTs) logically entails the hypothesis that events closer together in time are more likely to be confused with each other than events more separated in time (X_b). No auxiliary assumptions are needed to arrive at that implication. Therefore, P(X_b|TCT) = 1. Moreover, TCTs have received substantial empirical support so far, so that their current prior, P(TCT), is reasonably high. Together, these two probabilities imply that the posterior probability P(X_b|"x_b") is high, too: The finding "x_b" is unlikely to be a false positive. Therefore, observing "x_b" provides strong evidence for X_b, and by implication for TCTs. Conversely, failure to observe "x_b" (assuming strong statistical power, or a convincing Bayes factor in favor of the null) counts as evidence against TCT. This is so whether or not Researcher B carried out the experiment with the purpose of testing TCTs. The intention and insight of Researcher B does not matter—it could be that Researcher B is so dense that he never realizes the connection between his finding and TCTs, but a smarter colleague notices it, reanalyzes Researcher B's data, and publishes the result. The scientific community does not need to know about the history of mind of Researcher B and his colleague to appreciate that the data support TCTs.

Philosophers have proposed several solutions to the paradox of predictivism, which have in common that they assign no evidential value to temporal order per se. Rather, they regard temporal order as a proxy for something else that distinguishes impressive cases of successful a priori predictions of as yet unknown empirical generalizations from unimpressive successful post hoc explanations of known empirical generalizations (Barnes, 2008; Snyder, 1994). What is this "something else"? A post hoc explanation is vulnerable to the suspicion that the theorist designed the theory so that it explains the known phenomena, cobbling together just the right assumptions that are needed to arrive at the correct set of hypotheses. An a priori prediction does not attract that suspicion because the theorist, not knowing which prediction will turn out to be true, could not design the theory to fit the findings. Another way to put the difference—first proposed by Keynes (Barnes, 2008)—is this: When a theorist develops a theory T, and that theory successfully predicts a novel finding X, then the theorist must have good reasons to propose T that do not involve knowing X. These reasons could be other, already known empirical phenomena Y_1, Y_2, … Y_k that the theory explains, and/or theoretical reasons R_1, R_2, … R_m such as the plausibility of the theory's assumptions in light of what we know, or the theory's simplicity and elegance. Hence, when X is later confirmed by observation, theory T is supported not only by X but also by all Y and/or all R. In contrast, when a theorist designs a theory T that explains a known phenomenon X, then it is possible that knowledge of X is the only good reason for the theorist to propose T, and hence X is the only support for T.

These considerations narrow down the "something else" underlying our intuition that a priori predictions are preferable to post hoc explanations: When a theory T implies a hypothesis X, and evidence establishes X credibly, then this provides more support to T the more strongly T is justified independently of X. Independent justification means that there is a rational argument toward T that does not involve X as a premise—whether or not that argument accurately reflects the theorist's thought process (Ladyman, 2002). That is, even if a theorist had knowledge of X while constructing T, if that construction is based on good reasons without *using* knowledge of X, then X may strongly support T notwithstanding the reversed temporal contiguity.

The degree of independent justification arguably correlates with the distinction between a priori and post hoc, but, crucially, it is not necessarily tied to it. We can think of extreme cases of trivial post hoc explanations, in which a theorist designs a new theory such that it explains a single known

phenomenon X, and the theory has no other justification than its ability to explain X. But then, the history of science offers numerous counterexamples to the correlation: Cases in which a theory provides a first explanation for a long-known phenomenon, although the theory was not designed for that purpose at all.[5]

For the complementary counterexample to the correlation, we can think of cases in which a theory, designed arbitrarily without any justification, successfully predicts X a priori by sheer luck. This is arguably very unlikely—but a much more likely case is this: A theory motivates a large, perhaps infinite, number of predictions in the manner we delineated as characterizing discovery-oriented research (for instance, all the different ways of testing embodiment priming). Most of these predictions will fail, but a few of them might turn out to be true, and these successes would receive the credit of successful a priori predictions (perhaps formally established through preregistration). We argue that such credit would be unjustified, because the predictions receive only weak justification from the theory: For each of these predictions, the theory does not imply that they *will* be true, but only that they *can* be true, and P(X|T) is fairly low for each individual prediction. In other words, the rare prediction successes must be seen in the context of the many prediction failures, where the experiments "did not work." The a priori nature of a successful prediction is only a proxy for the prediction having a strong independent justification, or for some inherent quality that empowers the theory to make successful predictions. But if there is no strong justification, and if the predictive power of a theory is quite modest in light of its many predictive failures that accompany the few successes, then that proxy is invalid. Even preregistration does not make it more valid.[6]

Preregistration and "researcher degrees of freedom"

So what is the point of preregistration? Preregistration serves to reduce "researcher degrees of freedom" (Simmons et al., 2011)—that is, researchers' choices among large sets of equally defensible hypotheses to test, and analysis plans to test them. Within the classical framework of null-hypothesis testing—still the dominant statistical approach in psychology—uncontrolled researcher degrees of freedom have been argued to entail an uncontrolled inflation of Type I error rate due to multiple testing (de Groot, 1956/2014). Independent of the statistical framework chosen—classical null-hypothesis testing or Bayesian statistical inference—researcher degrees of freedom open the door to inadvertent biases when researchers choose hypotheses or analysis paths (e.g., choices about data preprocessing and statistical model) that lead to a desired result (Wagenmakers et al., 2012).

We have no doubts that preregistration does curtail researcher degrees of freedom, and as such it serves an important purpose in preventing a number of fallacies in scientific inference. At the same time, we think that preregistration, when applied mechanically—preregistering hypotheses and analysis plans chosen with little concern about their justification—remains a cure of the symptoms rather than a solution addressing the roots of the problem. Therefore, we want to ask: Where do the excessive researcher degrees of freedom come from, and can we do something to reduce them systematically rather than through an arbitrary decision that we privilege by uploading it on a preregistration repository?

Researcher degrees of freedom arise on both levels of scientific inference (see Fig. 1). On the empirical level they arise because we have a multitude of data transformations and data analytical tools at our disposal that are often equally justifiable. On the theory level they arise when theories can be used to motivate a large number of hypotheses—perhaps even contradictory hypotheses—equally well. We argue that on both levels there are more principled solutions than preregistration of an arbitrary choice, but the solutions for the two levels differ, so we will discuss them separately. We begin with a consideration of the problem of Type I error inflation through multiple comparisons in null-hypothesis testing. After that we will address the more general problem arising from researcher degrees of freedom—namely, the room they create for inadvertent bias. We will first discuss solutions to this problem on the empirical level, concerning the selection of data-analysis paths, and then on the theory level, concerning selection of hypotheses to test.

Preregistration and the problem of multiple comparisons

When researchers do not determine their hypotheses and their analysis plans before looking at the data, they will usually test multiple possible hypotheses (e.g., every possible pairwise correlation between n variables), and test each hypothesis through multiple possible analysis paths, thereby carrying out multiple tests (de Groot, 1956/2014). This is problematic in the context of null-hypothesis significance testing: When the Type I error for each test is set to α, the chance of committing at least one Type I error—erroneously rejecting a true null hypothesis—increases beyond α with multiple tests. As the number of tests researchers choose from is usually unknown, there is no way to correct for this error inflation.

It is useful to distinguish two cases of multiple testing (Rubin, 2017a), which correspond to our distinction between inferences on the empirical level and the theoretical level. Case 1 is the situation where a researcher tests the same hypothesis through multiple analysis paths and is willing to reject the null hypothesis if at least one analysis path results in a significant outcome. This case concerns the exploitation of

[5] A famous example is the explanation of a known anomaly in the orbit of Mercury by Einstein's general theory of relativity.

[6] If the preregistrations are automatically and obligatorily made public, then at least the research community gets to know about the context of prediction failures.

researcher degrees of freedom on the empirical level. It leads to an inflation of the Type I error rate for the hypothesis under investigation—this has become known as "*p*-hacking." Replacing null hypothesis testing by Bayesian statistics does not circumvent this problem: Running multiple analyses testing the same hypothesis, and selecting the one yielding the highest Bayes factor for one's preferred hypothesis, inevitably biases the conclusion. Under both statistical approaches, preregistration of analysis plans can avoid this bias because it reduces the number of tests to one.

Case 2 is the situation where a researcher tests multiple hypotheses, testing each of them with only one analytical approach (e.g., running a standard significance test on each of 100 correlation coefficients). This scenario does not lead to an inflated Type I error rate for each individual hypothesis. It does increase the chance of committing at least one Type I error among all hypotheses tested, and as such it increases the Type I error rate for the "joint null hypothesis" (de Groot, 1956/2014), which states that all individual hypotheses tested are false. But the joint null hypothesis—or its negation, the claim that "at least one of the *n* alternative hypotheses tested is true"—is rarely of scientific interest.

Testing a large, unconstrained number of hypotheses, as in Case 2, can still be problematic, but the problem does not arise from an inflated Type I error rate on the hypotheses tested. The problem arises if the researcher goes on a "fishing expedition," searching through a large hypothesis space with few constraints from theory or prior findings. This is the scenario arising from discovery-oriented research: In the absence of strong predictions from a theory, researchers have large degrees of freedom on the theory level, that is, freedom to choose from a large set of hypotheses through HARKing. The problem with HARKing is that the hypotheses usually have a low prior, P(X|T) (Dienes, 2011).

The distinction between the two cases is made transparent within a Bayesian approach to inference, which distinguishes between the prior of a hypothesis and the evidence in the data for that hypothesis (as expressed, for instance, in the Bayes factor). Case 1—the exploitation of researcher degrees of freedom in data analysis—compromises the assessment of the evidence. Case 2—exploiting degrees of freedom in hypothesis selection—implies low priors of the hypotheses. Classical null hypothesis testing has no place for priors and therefore tends to obscure the difference. A significant *p* value arising from HARKing in a discovery-oriented context can easily be misinterpreted as providing as much credibility (i.e., posterior probability) to a hypothesis as a significant test supporting a hypothesis derived from a theory in theory-testing research, when in fact the former should be far less convincing than the latter. To address that problem, classical statisticians often recommend treating Case 2 in the same way as Case 1 and correct for multiple testing (with paradoxical consequences pointed out by O'Keefe, 2003). Preregistration of hypotheses can then be seen as a way to limit the number of tests. From a Bayesian perspective, Case 2 should be addressed by assigning low priors to hypotheses.

To bring the difference between the two approaches to Case 2 into focus, consider Scenario C: Researcher C_1 runs an EEG study to find out how the brains of researchers react differently to significant versus not-significant *p* values. She determines that the effect of *p*-value significance on 10 dependent variables obtained from the EEG data is of interest (e.g., the N200, the P300, the posterior alpha power) and preregisters these 10 hypotheses. Therefore, she needs to correct her alpha level for multiple testing. Researcher C_2 is interested in the same hypotheses, but takes a more wasteful (though smarter) approach: He runs the same experiment 10 times, each time preregistering only one of the 10 hypotheses, and thereby evades the correction for multiple testing. The same data that lead to a significant result in favor of, say, an effect of *p*-value significance on the P300 for C_2 could fall short of being significant for C_1, thereby influencing how credible we find this hypothesis after considering the data. A Bayesian approach would treat both researchers' experiments in the same way, considering the priors for each hypothesis. These priors should depend on how strongly each hypothesis follows from a theory, and not on how many hypotheses a researcher plans to test in the same data set (Dienes, 2011). The role of preregistration in the Bayesian approach is to make researchers think about their priors without being biased by the data, but the act of preregistering a hypothesis does not increase its prior, and therefore has no impact on its posterior.

We next consider alternatives to preregistration as tools to manage researcher degrees of freedom, first for the empirical level, and subsequently for the theory level.

From data to empirical generalizations: Degrees of freedom in analysis plans There are plenty of sources of researcher degrees of freedom on the empirical level of scientific inference—from data to empirical generalizations—and much has been written about them. The "garden of forking paths" (Gelman & Loken, 2014) includes decisions about data selection (e.g., which variables, or which observational units, to include in the analysis), data preprocessing (e.g., whether or not to transform response times logarithmically), and about which statistical analysis procedure to use. Because our methodological toolbox for most research problems is already large, with a tendency to grow fast, researchers often face a situation where several alternatives at each forking point are equally justifiable (for an example, see Silberzahn et al., 2018). Choosing among these options depending on their outcomes invites biases in favor of choosing an analysis path that yields the researchers' preferred conclusion, or an outcome likely to be publishable.

If there are multiple analysis paths that are equally justifiable for achieving a goal of statistical inference (e.g., testing a given hypothesis), then the optimal solution is to run all equally justifiable analyses and record to what extent they converge on the same results. When the number of options is too large to run them all, one can still run a sample of different analysis plans (similar to a sensitivity analysis; Thabane et al., 2013). The degree of convergence provides information about the robustness of the results against variation of analytic choices that should not matter. This approach is nicely illustrated by the "multiverse" analysis of Steegen, Tuerlinckx, Gelman, and Vanpaemel (2016), who investigated the robustness of inferences from a data set across a multitude of data preprocessing decisions. Our proposal is to generalize this approach to other decisions in the analysis pipeline (e.g., concerning outlier treatment, statistical model to be tested, inclusion of independent variables and covariates). This multiverse analysis circumvents the problem of Type I error inflation because the researcher no longer draws conclusions from a single test chosen according to its outcome, but from all tests run.

Compare that gold standard to preregistering one analysis plan, and regarding that plan only as providing strong, "confirmatory" evidence (Nosek, Ebersole, DeHaven, & Mellor, 2018). If there truly are, before looking at the data, several equally justifiable analysis plans, then choosing one of them for preregistration is arbitrary. It engenders the risk that this one way of analyzing the data happens to miss an interesting pattern that would be revealed by, say, 90% of all other equally justifiable analysis plans. By sticking to the preregistered plan, we will never find out. Departing from that plan, however, would mean that we officially enter "exploratory" territory, and the outcomes would carry less weight in the minds of researchers who focus on preregistration as a sign of quality—in fact, within the framework of null-hypothesis testing, the exploratory analyses have no evidential value concerning the hypothesis under investigation at all (de Groot, 1956/2014).

To illustrate, consider Scenario D in which Researcher D anticipates that for her experiment there are 50 a priori equally defensible data analysis paths, and she decides to preregister one of them, chosen by a random draw. After analyzing the data according to the preregistered procedure, she continues to run the remaining 49 analysis plans as "exploratory." Assume that the preregistered analysis yields evidence in favor of X, and of the 49 exploratory analyses, a percentage ε yields evidence against X. What percentage ε would convince you to trust the exploratory analyses more than the "confirmatory" preregistered one? Should the preregistered analysis really carry more weight in the balance of evidence than each of the other 49? Readers who think *yes* might consider the following variant of this scenario: A team of 10 researchers D_1 to D_{10} plan a large study together. They each independently preregister an analysis plan that they choose at random from the 50 equally justifiable options. When the data are in, each of them runs all 50 analyses on the same data, and—necessarily—arrives at the same mixture of results: Some analyses support X, others support the null hypothesis. Because for each team member a different analysis plan counts as confirmatory, and thereby receives higher weight, they are likely to come to different conclusions about what the data imply for hypothesis X. The only reason for their discrepant conclusions is a series of coin tosses.

To conclude, preregistering analysis plans curtails researcher degrees of freedom, and this is an important safeguard against bias. However, preregistration of one analysis plan chosen arbitrarily from a set of equally justifiable alternatives remains an arbitrary choice, and does not receive a privileged status of evidential force through preregistration per se. Preregistration is a public record of the mental history of the researcher, and the history of the researcher's reasoning process has no bearing on the rationality of their decisions (see our earlier discussion of the "paradox of predictivism"). If a researcher chooses an analysis method that is inappropriate, or biased in favor or against one hypothesis, then preregistering this method does nothing to mitigate these deficiencies. Therefore, giving the results of preregistered analyses more weight in evaluating a hypothesis than the results of equally reasonable but not preregistered analyses would be a mistake.

We believe that the following alternative is more defensible: When faced with a garden of equally justifiable forking paths, walk down a sample of them that covers the garden with reasonable breadth, and report how consistent the findings are. In addition, make the raw data publicly available so that other researchers who doubt your conclusions can run any alternative analysis plan to check the robustness further. One could, of course, preregister the entire intended multiverse of analyses, but if the multiverse spans the space of reasonable options—as it should—then preregistration adds nothing. In summary, preregistration of one analysis path negates the advantage of a multiverse analysis by privileging one arbitrarily chosen approach; preregistering the entire multiverse adds nothing because, by design, a multiverse analysis is dealing with researcher degrees of freedom already.

We say this not to discourage our colleagues from preregistering analysis plans. To the contrary, we believe that preregistration is useful as an explicit stage in the research process, dedicated to considering which data preprocessing and analysis paths are reasonable and justifiable. It is also a very useful exercise to keep one's own hindsight bias under control, because what appears reasonable and justifiable to us after analyzing the data tends to be biased by the results. Preregistration thus serves as a tool to overcome our psychological shortcomings, and we should embrace this protective function. Following a single preregistered analysis plan is clearly more reasonable and rigorous than running an analysis in multiple ways and cherry-picking the one that works in

one's favor for publication. But like all useful tools, preregistration is not without limitations, and it is crucial that researchers are aware of those limitations. An important limitation to recognize is that preregistration of an analysis path does not change the logical status of the results arising from that analysis path. Limiting ourselves to one single way of looking at the data as the one that counts for making inferences, and downgrading all other analyses as "exploratory," risks replacing a self-serving bias by a blind bias. Giving equally justifiable analysis approaches equal weight, as in a multiverse analysis, overcomes that limitation.

From theories to hypotheses: Degrees of freedom in hypothesis selection Large degrees of freedom for HARKing arise naturally from discovery-oriented research. Much of psychology is characterized by theories that are so vague that they do not strongly imply any hypothesis without a host of auxiliary assumptions. These auxiliary assumptions are not constrained by the theory and can therefore be chosen in any way that suits the researcher. When hypothesizing after the results are known, researchers can choose the auxiliary assumptions in such a way that the hypothesis matches the result. To illustrate, we noted earlier in our example involving embodiment priming that the prediction for an experiment involving religiosity might involve a deductive chain such as "If EPT is true, and if the concept of a 'leap of faith' is embodied in the physical action of jumping over a gap, and if jumping over a gap once primes that concept for at least a few minutes, and if priming that concept shifts people's responses on a questionnaire towards higher religiosity scores, then the experimental group will show higher religiosity scores (X_a)." Upon finding no significant effect on religiosity, the researcher could freely hypothesize that any of the auxiliary assumptions might be in need of revision: For example, the gap may have primed the idea of a "leap of faith," as postulated, but that type of faith may be unrelated to the arguably more fundamental faith in a supreme being, and rather pertain to the more mundane faith in other people. The researcher could then proceed to test the new hypothesis that participants in the experimental group had more faith in the trustworthiness of the experimenter, and therefore were more likely to sign a postexperimental consent form agreeing that a video recording of their behavior could be used for further research.

These degrees of freedom are much curtailed in theory-testing research. Take, for instance, theories of recognition memory. There is a long-running debate between proponents of two families of theories: On the one side there are theories assuming that recognition decisions are made by evaluating whether a signal from memory that varies on a continuum of strength exceeds a criterion, as formalized in signal-detection theory (Wixted, 2007). On the other side are theories assuming that recognition decisions arise from two or three discrete mental states: A detect state (remembering that the probe has been experienced as part of the relevant memory set) resulting in an "old" response, a guessing state (not remembering anything about the probe) resulting in an uninformed guess of "old" or "new," and (in some theories) a second detect state (remembering that the probe was not in the memory set) resulting in a "new" response (Bröder & Schütz, 2009). Recently, Kellen and Klauer (2014, 2015) have derived hypotheses from the two classes of theories by which they can be distinguished, and these hypotheses follow from the core assumptions of the theories alone, without auxiliary assumptions. In this way, Kellen and Klauer reduced the degrees of freedom for hypothesizing to zero.

To conclude, researcher degrees of freedom in formulating hypotheses can be more or less constrained by the theories from which they are derived. On one end of the continuum, which we described as theory-testing research, hypotheses are strongly implied by theories with little, if any, flexibility arising from varying auxiliary assumptions or parameter values. Preregistering these hypotheses makes their a-priori character explicit but does not add anything substantive, because the hypotheses follow from the theories no matter when, or whether, a researcher thinks of them. In other cases, which we described as discovery-oriented, hypotheses are merely motivated—rather than strongly entailed—by theories, and therefore researchers have many degrees of freedom to vary them. An even more extreme case at that end of the continuum is a search for empirical effects not guided by any theory, motivated perhaps by practical questions (e.g., asking whether students learn better when lecturers make jokes), or just as a fishing expedition (e.g., asking whether any of 37 personality scales that happen to be available for a large sample predicts a person's sexual orientation, or asking which of >100 cortical areas' BOLD signal correlates with people's report of subjective awareness of a stimulus). These are the cases that give researchers huge leeway for HARKing, and that have given HARKing its bad reputation: When the results are known, inventive researchers can always come up with a plausible story explaining why the results had to come out exactly as they did.[7] Preregistration of hypotheses can curtail that practice—but then, which hypotheses could a researcher preregister when no hypothesis is strongly implied by any theory? A researcher engaged in discovery-oriented research or a fishing expedition would have to place a blind bet on a prediction to preregister. If that prediction turns out to be true, it would still be nothing but a lucky guess—it looks like strong support for the researcher's theoretical claim, but it is not. There is no reason to expect future predictive success from a theory supported by lucky guesses—just as there is no reason to expect that a stock broker who was lucky on the stock

[7] Of course jokes improve learning: They increase alertness. Of course, jokes impair learning: They distract from the material to be learned. Of course jokes have no effect: Students don't listen to instructors, period.

market five times in a row is a financial genius who can outsmart the market.

So what is the value of preregistering hypotheses? Like with preregistration of analysis plans, we see its main benefit in controlling our biases. It motivates researchers to think about what they predict for a study, and about how, and how strongly, their prediction is actually justified by the theory that motivates it. When the argument from theory to prediction is thought through before seeing the data it cannot be biased by the data. Preregistration of hypotheses therefore reassures us and our colleagues that our reasoning is unbiased—it does not, however, replace such reasoning. Preregistering a hypothesis without providing a reason for it is pointless, and does nothing to increase the credibility of the hypothesis even if it happens to be supported by a study's finding.

Toward stronger theories: Formalization and computational modeling

We have argued throughout this article that the development of strong theories – with a tight logical link between theoretical assumptions and hypotheses derived from them – goes a long way towards addressing the replication crisis. We now consider in more detail what it means to develop strong theories.

As the example of recognition theories mentioned above (Kellen & Klauer, 2014, 2015) illustrates, deductive derivation of a hypothesis is greatly facilitated when theories are formalized as a set of equations or propositions. We can then use mathematical analysis or formal logic to determine what hypothesis does or does not follow from the theory, and where that is too difficult, we can use simulation to derive hypotheses unambiguously. Sometimes, formalizing a theory and investigating it through simulation can help uncover that a hypothesis a theorist had thought to derive from their theory actually does not follow from the assumptions of the theory; occasionally even the negation of the original hypothesis follows (for examples, see Lewandowsky & Oberauer, 2015; Oberauer & Lewandowsky, 2014). This possibility gives rise to a further illuminating scenario, Scenario E: Researcher E develops a well-spelled-out theory T and argues that the theory entails a novel prediction X. After preregistering that prediction (and the analysis plan), Researcher E carries out an empirical study that provides strong evidence for X. Subsequently, another researcher formalizes the assumptions of T, and demonstrates through simulation that T cannot generate the effect pattern X. The conclusion must be that the successful prediction of X (whether or not it was preregistered) counts as evidence against T.

Free parameters and arbitrary assumptions Formal modeling helps to determine what hypotheses are entailed by a theory, but it does not by itself solve the problem of degrees of freedom about hypotheses. Many formal models have considerable flexibility in what data patterns they generate. This flexibility has two sources. One source is the flexibility inherent in any formal model, arising from the model's free parameters in conjunction with its functional form. This source of flexibility is being intensely studied in the field of statistics concerned with model selection (Pitt, Myung, & Zhang, 2002), and increasingly sophisticated methods are being developed to take model flexibility into account when determining which model gives a better account of some data (Shiffrin, Lee, Kim, & Wagenmakers, 2008).

One way to keep this first source of model flexibility in check is to constrain the values of free parameters, either by making theoretical assumptions about plausible parameter values or by drawing on prior empirical knowledge. In classical statistical methods of model fitting (e.g., maximum-likelihood methods), such constraints can only be set in a hard way by placing upper and lower bounds on parameter values, or fixing parameters to a single value. For instance, modelers often require some parameters to remain invariant across applications of the model to different data sets (so-called "universal free parameters" according to Wills & Pothos, 2012). Within a Bayesian framework, constraints on parameter values can be incorporated in informative priors on parameters (Lee & Vanpaemel, 2018). Informative prior distributions can implement soft constraints by concentrating probability mass on the most plausible parameter values while still assigning some prior probability to a broad range of less plausible values. A principled empirical way of determining priors is to use the posteriors of parameter values from one data set as the priors on these parameters for the next data set (Kary, Taylor, & Donkin, 2016). Doing this successively is likely to progressively narrow the priors, thereby reducing the parameters' freedom to vary. In this way, empirical certainty is accumulated over a series of studies, but they don't need to be direct replications of each other, they just need to reuse (in part) the same free parameters. We can think of the parameter estimates carrying over from previous model applications as a form of preregistration of parameter values—but one that constrains parameters in a principled way informed by data.

A second source of flexibility lies with the decisions that researchers make when building a formal model. Once these decisions are made, they are hard-wired into the model—we can think of them as preregistered in the model equations (Muthukrishna & Henrich, 2019). Yet they are degrees of freedom at the model-development stage (Farrell & Lewandowsky, 2018, Chapter 2). Assumptions built into a formal model vary in the degree to which they are justified by theoretical considerations independent of the empirical generalizations the model is built to explain.

Take, for example, resource models of visual working memory. Several models in this class are built on the basic

idea of a sample-size model (Palmer, 1990). The core assumption is that visual working memory has a limited number of units for coding visual features (e.g., colors, line orientations); each unit codes a visual feature with a limited degree of precision, and when multiple units redundantly code the same feature, the information from these units is averaged when that feature needs to be retrieved. These assumptions mathematically imply that the precision of the retrieved feature (expressed as the standard deviation of report, σ, or sensitivity, d', in recognition tests) increases with the square root of the number of units redundantly coding it—in the same way as the standard error of an estimated mean decreases with the square root of the sample size of a study. With the additional assumption that, when multiple visual objects are held in working memory, they share the available feature-coding units, this class of resource models makes a precise prediction for how memory precision σ declines as memory set size n increases (Bays & Husain, 2008; Sewell, Lilburn, & Smith, 2014; Smith, Lilburn, Corbett, Sewell, & Kyllingsbæk, 2016). This prediction is expressed by a power function:

$$\sigma_n = \frac{\sigma_1}{n^{0.5}}.$$

At this point, some models in this class include the assumption that the exponent can differ from 0.5, and therefore the constant is replaced by a free parameter (Bays & Husain, 2008; van den Berg, Shin, Chou, George, & Ma, 2012). This free parameter gives the model extra flexibility and enables a better account of the data. However, it turns the power function from something that follows from the theory's core assumptions into a convenient choice of a mathematical function for describing the effect of set size on memory precision (Oberauer & Lin, 2017).[8]

Our argument is this: Researchers are free to make decisions when building a formal model, but not all such decisions are equally justifiable. If we carefully scrutinize how well model assumptions are justified—by the theoretical ideas they intend to formalize, by their degree of coherence and integration with the rest of the model, by their convergence with assumptions in other empirically successful models, or by empirical knowledge outside the set of phenomena that the model is built to explain (e.g., knowledge about how individual neurons work constraining neural-network models of behavior; see O'Reilly & Munakata, 2000)—we can reduce the researcher degrees of freedom for building models.

[8] Recently, resource theorists have addressed this conundrum in two ways. Smith, Corbett, Lilburn, and Kyllingsbæk (2018) have shown that the freely estimated exponent increases with more attention-demanding stimuli, and that a resource model assigning resources unequally to stimuli predicts exponents >0.5. Van den Berg and Ma (2018) abandon the notion of a constant resource and instead propose that the resource amount assigned to memory representations follows a regime of rational cost–benefit analysis.

Does formal modelling exaggerate the replication problem?

Whereas we believe that formal models of psychological processes are part of the solution to curtail researcher degrees of freedom in hypothesizing, Fiedler (2017) argues that they are part of the problem. Fiedler points out two weaknesses of formal models. First, models consist of a number of very specific assumptions about psychological mechanisms and processes. This degree of specificity—precisely the characteristic that licenses strong inferences to hypotheses—entails that the prior probability of a model being true is low: For the model to be true, each of its assumptions needs to be true, and the more specific each assumption is formulated, the more alternatives it excludes, which reduces its prior probability of being true. If the prior of a model is low, then the prior of every hypothesis deduced from it is low, too. Therefore, Fiedler argues, formal process models are ill suited to lead the way to highly replicable findings. The second weakness Fiedler points out is that the confirmation of hypotheses derived from models is often not diagnostic—other models incorporating entirely different assumptions may entail the same hypotheses.

Fiedler's second argument partially neutralizes his first: If a hypothesis X is necessarily implied not only by a model T_1 but also by at least one other model T_2, then the prior of the hypothesis, P(X), is already larger than the prior of that model, $P(T_1)$. For instance, consider two mutually exclusive models T_1 and T_2, each of which imply the same hypothesis X. In this case, P(X) is the sum of $P(X|T_1) \times P(T_1)$ and $P(X|T_2) \times P(T_2)$; if the hypothesis follows deductively from each theory, the two conditional probabilities are both one, and P(X) is the sum of the two model priors, $P(T_1) + P(T_2)$. More generally, the hypotheses implied by a model are often not unique to that model—they are usually also implied by other, similar models, and often also by other models starting from entirely different assumptions. Hence, even if the prior of each individual model is, on average, very small, the priors of the hypotheses entailed by them are not necessarily small. They are small if the hypothesis is unique to the model under investigation, but larger to the extent that other models also imply the same hypothesis. When other models imply the same hypothesis, confirming the hypothesis as an empirical generalization does not uniquely single out one model as the winner, but it is still informative as it reduces the set of viable models.

The situation is different from discovery-oriented research, where the link between theory and hypothesis is weak, meaning that P(X|T) is low for each testable hypothesis X. If X is to have confirmatory diagnosticity for T (i.e., confirming X is to be counted as support for the theory), then $P(X|\neg T)$ has to be substantially smaller still—in other words, the hypothesis has to be bold, and evidence for it must be surprising. This is why the prior of a hypothesis X is necessarily small in discovery-oriented research. By contrast, in theory-testing research, and in particular when testing formal models, P(X|T) is very high—in the ideal case, when the hypothesis follows

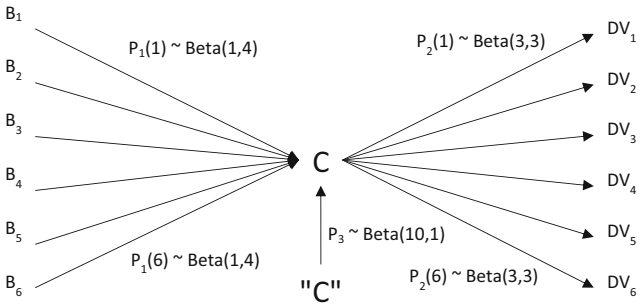

Fig. 6 Core causal structure of embodiment priming theory. A set of bodily states or movements B_1 to B_n is considered as plausible manifestations of the embodied representation of concept C, such that experimentally inducing B_x activates (primes) C. Each B_x has a probability $P_1(x)$ of actually priming C. A set of observable behaviors is considered as dependent variables DV_1 to DV_n. If C is primed, it influences each DV_x with probability $P_2(x)$. The linguistic expression "C" activates C with probability P_3. The figure includes plausible priors for the probabilities P_1, P_2, and P_3; the priors of P_1 and P_2 were chosen so that the expected value of $P(X|EPT) = P_1 \times P_2 = 0.1$, in agreement with the numerical example for discovery-oriented research in Table 2 and the text

deductively from the theory or model, it is unity—and therefore $P(X|\neg T)$ can be fairly high, too, without confirmation of X losing its diagnostic value for T. In our numerical examples we set $P(X|\neg T)$ to 0.2. In the context of testing competing computational models against each other, we can interpret this value as saying that, among those (known and not yet known) models that are alternatives to T (our current model of interest), there are some that also imply X, and the sum of their priors is 0.2. This implies that the prior of the hypothesis, P(X), is larger than 0.2. To conclude, Fiedler's claim that the prior probability of a hypothesis derived from a formal model is necessarily low is disconfirmed by his own recognition that multiple models may make the same prediction.

A second point, neglected in Fiedler's critique, is that empirical tests of formal models—and theory-testing empirical research in general—afford symmetric evidence: Data confirming a hypothesis X provide evidence in favor of any theory or model that implies hypothesis X, but at the same time, data disconfirming X provide evidence *against* these theories or models (see Fig. 2), and provides more credibility to those competing theories or models that imply the negation of X. In other words, each test of X is also a test of ¬X, and therefore we should care as much about the prior of ¬X as about the prior of X. As noted above, to the extent that P(X) is large, $P(\neg X) = 1 - P(X)$ is small. Therefore, Fiedler's (2017) recommendation to maximize P(X)—for instance, by deriving hypotheses from necessary statistical truths—although doubtlessly addressing the problem of nonreplicable findings, is not the best way forward for gaining knowledge on the level of theories. When taken to an extreme, it leads to tests of hypotheses that are trivially true, and confirming them is largely uninformative. Arguably, the ideal hypothesis is one that is entailed by about half of the credible models in an area (i.e., those models with a nonnegligible prior), such that the hypothesis' prior is around 0.5. In this way, whichever outcome of an empirical test is obtained—confirming or disconfirming the hypothesis—the result cuts the set of remaining credible models in half.

Conclusion: Remedies for the theory crisis in psychology

We need to reduce researcher degrees of freedom on both levels of scientific inference. On the empirical level, we propose that researchers check the robustness of their inferences against variations of analysis decisions that are equally justifiable, and that they make their raw data publicly available whenever possible (see Lewandowsky & Bishop, 2016, for boundary conditions) so that others can continue checking their robustness. On the theory level—our primary concern in this article—there are two paths toward reducing researcher degrees of freedom. One is to do discovery-oriented research, but do it right. Researchers on this path accept that the current state of theorizing in their field does not license strong inferences to hypotheses, and that as a consequence, their hypotheses usually have a lower prior probability. The implication is that large sample sizes and/or direct replications are needed to establish a new empirical generalization with a satisfactory level of credibility. The other path is to do theory-testing research. Researchers on this path make an effort to formulate their theories as precisely as possible, thereby strengthening the inferential link from the theory to the hypotheses derived from it. Expressing the theory formally is likely to help in this endeavor. This path deemphasizes the need for direct replication and favors successive tests of *different* hypotheses.

We acknowledge that some subdisciplines of psychology have a longer tradition of formal modeling to build on than others, and therefore theory-testing research might appear to be out of reach for many psychological scientists. We do not want to downplay the difficulties of formulating a precise theory that enables strong inferential links to hypotheses. At the same time, we argue that researchers can always make steps toward formulating their theoretical ideas more precisely, and even formally. A formal theory does not need to spell out mechanism and processes in much detail—formal models exist on various levels of abstraction. A formal model could simply consist of a path diagram making explicit the monotonic causal links that are assumed between two or more continuous variables, or a Bayesian network (Glymour, 2003) making explicit the probabilistic dependencies between discrete variables. Developing such a model would involve identifying and explicitly incorporating assumed moderator variables, boundary conditions, and other auxiliary assumptions. Often theorists hesitate making these additional assumptions explicit because there is so much uncertainty about them that

fixing them would come down to an arbitrary guess. The way forward in these cases would be to incorporate uncertainty into the model. The Bayesian modeling framework is ideally suited for that purpose: Uncertainty is incorporated through priors. Usually, priors are placed on quantitatively varying free parameters, expressing our degree of uncertainty about the quantity in question. However, we can use priors also to express uncertainty about qualitatively different choices in building a model—for instance, the choice between different functional forms for the relation between two variables (i.e., the relation could be linear, or exponential, or a power function). The prior would then be a probability distribution over a set of discrete options in the model. Uncertainty about model assumptions can be expressed explicitly and formally, implying that uncertainty is not the same as vagueness about model assumptions. The former is no excuse for the latter.

We end by revisiting the embodiment priming theory that we used as an example for the kind of theory that often motivates discovery-oriented research, and ask what it would take to transform this theory into one that guides a theory-testing research program. Figure 6 presents a blueprint for the core causal model of EPT. At its center is the assumed embodied representation of a selected concept C. This concept is assumed to be activated by an unknown subset of possible bodily states or movements. That subset is unknown because the theory does not specify how C is embodied; there is arguably a large set of possible embodiments of a concept, of which only one is true for each person, so for each possible experimentally induced bodily state or movement there is an unknown, probably small, probability that it activates C. The concept C, when activated, is expected to influence a subset of possible judgments, decisions, and actions that are semantically related to C. Again, there is a vast set of such behaviors that could be chosen as dependent variables, and the theory does not specify which of them will be affected by activating C, so the probability for each of them being activated is unknown, and perhaps small. Finally, the concept can be expressed linguistically. Assuming that researchers identify C linguistically, and that they come from the same linguistic community as the population they investigate, the linguistic expression is highly likely to be understood by the study participants as referring to C, so we can assume a high probability that the linguistic expression "C" activates C.

With this admittedly highly simplified formalization of EPT in place, we can consider two research programs for testing it. The first implements discovery-oriented research, but is heeding our recommendations for this kind of pursuit. For a given concept of interest (e.g., "faith"), the researchers would select one experimentally induced bodily state or motion (e.g., leaping over a gap in the floor) and one dependent variable that could be influenced by priming the concept (e.g., the score on a religiosity questionnaire) in an unprincipled manner (e.g., by discussing in a lab meeting how an experiment on the embodiment of a "leap of faith" could be done in an inexpensive way). They run a first experiment with acceptable power ($1 - \beta = 0.8$). If that experiment yields evidence for the predicted effect, they follow it up by a direct replication, perhaps with even more power. If things go well, and if the experiments conform to the empirical recommendations made earlier (e.g., multiverse analyses), this endeavor can establish an empirical generalization such as "Asking people to jump over a gap in the floor increases their scores on a religiosity test (taken within a certain time window)."

The second research program implements theory-testing research. The researchers would recognize that EPT does not offer a strong inferential link to a prediction for any individual combination of an experimentally induced bodily state or movement with a dependent variable. However, if the formalization of EPT includes a commitment to at least moderately informative priors about the probabilities P_1 and P_2, the theory does license a strong prediction for a representative sample of the population of bodily states and movements that could embody the concept, combined with a representative sample of potentially affected outcome variables: The prediction is that about $P_1 \times P_2$ such combinations will produce a true effect. Testing this prediction presents several challenges, the least of which is the large number of necessary experiments (a collaborative effort across many labs could overcome that hurdle). Researchers have to first clarify the population of possible bodily states and movements that could, with some plausibility, embody the concept in question, and assign it a distribution of prior probability. The same would have to be done for the population of behaviors plausibly affected by priming the concept, which could be used as dependent variables. Then a representative sample from both populations needs to be drawn to arrive at a set of experiments. By representative we mean representative for the population of experimental situations and behaviors we want to generalize the theory to. Assuming that that is the population of situations and behaviors occurring in people's everyday lives, our proposal converges with what is known as Brunswikian experimental design (Freund & Isaacowitz, 2013; Gigerenzer, Hoffrage, & Kleinbölting, 1991). The experiments need to be analyzed jointly—the relevant statistical inference question is not whether any individual experiment shows a true effect, but whether the sample of measured effects supports a model assuming a proportion of $P_1 \times P_2$ true effects in the population over a model with zero true effects. Such a model comparison could turn out either way, and therefore, this empirical test has a chance to provide strong support in favor, but also against the theory—as we demonstrated above, this symmetry of possible conclusions is a hallmark of theory-testing research that distinguishes it from discovery-oriented research.

The theory-testing research program sketched above obviously far exceeds the discovery-oriented research program in the amount of conceptual effort and data-collection resources

it requires. But then, we get much more in return: If supported by the evidence, EPT would be supported not merely by one randomly chosen—albeit firmly established—empirical generalization, but by findings in a representative sample of possible experimental tests of the theory, which allows generalizing the theory to the population of situations and behaviors it is meant to apply to. The positive findings in that sample could be considered *conceptual replications* of each other with respect to EPT. Because the result of our hypothetical megastudy consists of the joint outcome of all experiments in the sample, there is no leeway for selectively publishing the subset of studies that "worked." None of the positive findings in the sample of experiments would be confirmed through direct replication, so we will not know whether any individual effect is true. Does it matter? That depends on what we ultimately want to know: Whether leaping over a gap in the floor raises scores on a religiosity test—or whether EPT captures something true about how the mind works? In other words: Is the goal of our science to establish empirical generalizations, or to work towards better theories?

References

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

Barnes, E. C. (2008). The paradox of predictivism. Cambridge: Cambrdige University Press.

Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science, 321,* 851–854.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour, 2*, 6–10. https://doi.org/10.1038/s41562-017-0189-z

Bröder, A., & Schütz, J. (2009). Recognition ROCs are curvilinear—Or are they? On premature arguments against the two-high-threshold model of recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 587–606.

Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review, 114*, 539–576.

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. https://doi.org/10.1038/nrn3475

Conway, A. R. A., Kane, M. J., & Engle, R. W. (2003). Working memory capacity and its relation to general intelligence. *Trends in Cognitive Sciences, 7*, 547–552.

de Groot, A. D. (2014). The meaning of "significance" for different types of research (E.-J. Wagenmakers, D. Borsboom, J. Verhagen, R. Kievit, M. Bakker, A. Cramer, D. Matzke, D. Mellenbergh, & H. L. J. van der Maas, Trans. and annotated). *Acta Psychologica, 148*, 188–194. https://doi.org/10.1016/j.actpsy.2014.02.001 (Original work published 1956)

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science, 6*, 274–290.

Farrell, S., & Lewandowsky, S. (2018). Computational modeling of cognition and behavior. Cambridge: Cambridge University Press.

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. https://doi.org/10.1177/1745691612459059

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

Fiedler, K., Kutzner, F., & Krueger, J. I. (2012). The long way from α-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

Freund, A. M., & Isaacowitz, D. M. (2013). Beyond age comparisons: A plea for the use of a modified Brunswikian approach to experimental designs in the study of adult development and aging. *Human Development, 56*, 351–371.

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102*. Retrieved from http://www.americanscientist.org/issues/feature/2014/6/the-statistical-crisis-in-science

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*, 684–704.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98*, 506–528.

Glymour, C. (2003). Learning, prediction and causal Bayes nets. *Trends in Cognitive Sciences, 7*, 43–48.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine, 2*, 696–701.

Jostmann, N. B., Lakens, D., & Schubert, T. W. (2009). Weight as an embodiment of importance. *Psychological Science, 20*, 1169–1174. https://doi.org/10.1111/j.1467-9280.2009.02426.x

Kary, A., Taylor, R., & Donkin, C. (2016). Using Bayes factors to test the predictions of models: A case study in visual working memory. *Journal of Mathematical Psychology, 72*, 210–219. https://doi.org/10.1016/j.jmp.2015.07.002

Kellen, D., & Klauer, K. C. (2014). Discrete-state and continuous models of recognition memory: Testing core properties under minimal assumptions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 1795–1804. https://doi.org/10.1037/xlm0000016

Kellen, D., & Klauer, K. C. (2015). Signal detection and threshold modeling of confidence-rating ROCs: A critical test with minimal assumptions. *Psychological Review, 122*, 542–557. https://doi.org/10.1037/a0039251

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*, 196–217.

Körner, A., Topolinski, S., & Strack, F. (2015). Routes to embodiment. *Frontiers in Psychology, 6*, 940. https://doi.org/10.3389/fpsyg.2015.00940

Ladyman, J. (2002). Understanding philosophy of science. Oxon: Routledge.

Lee, M. D., & Vanpaemel, W. (2018). Determining informative priors for cognitive models. *Psychonomic Bulletin & Review, 25*, 114–127. https://doi.org/10.3758/s13423-017-1238-3

Lewandowsky, S., & Bishop, D. (2016). Don't let transparency damage science. *Nature, 529*, 459–461.

Lewandowsky, S., Duncan, M., & Brown, G. D. A. (2004). Time does not cause forgetting in short-term serial recall. *Psychonomic Bulletin & Review, 11*, 771–790.

Lewandowsky, S., & Oberauer, K. (2015). Rehearsal in serial recall: An unworkable solution to the non-existent problem of decay. *Psychological Review, 122*, 674–699. https://doi.org/10.1037/a0039684

Lynott, D., Corker, K. S., Wortman, J., Connell, L., Donnellan, M. B., Lucas, R. E., & O'Brien, K. (2014). Replication of "Experiencing Physical Warmth Promotes Interpersonal Warmth". *Social Psychology, 45*, 216–222.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin, 109*, 163–203.

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*. Retrieved from https://doi.org/10.1098/rsos.160426

Miller, J. (2009). What is the probability of replicating a statistically significant effect? *Psychonomic Bulletin & Review, 16*, 617–640. https://doi.org/10.3758/PBR.16.4.617

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., … Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behavior, 1*. https://www.nature.com/articles/s41562-016-0021

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

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

O'Keefe, D. J. (2003). Colloquy: Should familywise alpha be adjusted? *Human Communication Research, 29*, 431–447.

O'Reilly, R. C., & Munakata, Y. (2000). Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain. Cambridge: MIT Press.

Oberauer, K., & Lewandowsky, S. (2014). Further evidence against decay in working memory. *Journal of Memory and Language, 73*, 15–30. https://doi.org/10.1016/j.jml.2014.02.003

Oberauer, K., & Lin, H.-Y. (2017). An interference model of visual working memory. *Psychological Review, 124*, 21–59.

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

Palmer, J. (1990). Attentional limits on the perception and memory of visual information. *Journal of Experimental Psychology: Human Perception and Performance, 16*, 332–350.

Pashler, H., & Harris, C. R. (2012). Is the replication crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7*, 531–536. https://doi.org/10.1177/1745691612463401

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109*, 472–491.

Rubin, M. (2017a). Do *p* values lose their meaning in exploratory analyses? It depends how you define the familywise error rate. *Review of General Psychology, 21*, 269–275. https://doi.org/10.1037/gpr0000123

Rubin, M. (2017b). When does HARKing hurt? Identifying when different types of undisclosed post hoc hypothesizing harm scientific progress. *Review of General Psychology, 21*, 308–320. https://doi.org/10.1037/gpr0000128

Sederberg, P. B., Howard, M. C., & Kahana, M. J. (2008). A context-based theory of recency and contiguity in free recall. *Psychological Review, 115*, 893–912.

Sewell, D. K., Lilburn, S. D., & Smith, P. L. (2014). An information capacity limitation of visual short-term memory. *Journal of Experimental Psychology: Human Perception and Performance, 40*, 2214–2242. https://doi.org/10.1037/a0037744

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science, 32*, 1248–1284. https://doi.org/10.1080/03640210802414826

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., … Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic 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.

Smith, P. L., Corbett, E. A., Lilburn, S. D., & Kyllingsbæk, S. (2018). The power law of visual working memory characterizes attention engagement. *Psychological Review, 125*, 435–451. https://doi.org/10.1037/rev0000098

Smith, P. L., Lilburn, S. D., Corbett, E. A., Sewell, D. K., & Kyllingsbæk, S. (2016). The attention-weighted sample-size model of visual short-term memory: Attention capture predicts resource allocation and memory load. *Cognitive Psychology, 89*, 71–105. https://doi.org/10.1016/j.cogpsych.2016.07.002

Snyder, L. J. (1994). Is evidence historical? In P. Achinstein & L. H. Snyder (Eds.), Scientific methods: Conceptual and historical problems (pp. 95–117). Malabar: Krieger.

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

Thabane, L., Mbuagbaw, L., Zhang, S., Samaan, Z., Marcucci, M., Ye, C., … Goldsmith, C. H. (2013). A tutorial on sensitivity analyses in clinical trials: The what, why, when and how. *BMC Medical Research Methodology, 13*, 92. https://doi.org/10.1186/1471-2288-13-92

Topolinski, S., & Sparenberg, P. (2012). Turning the hands of time: Clockwise movements increase preference for novelty. *Social Psychology and Personality Science, 3*, 208–214. https://doi.org/10.1177/1948550611419266

van den Berg, R., & Ma, W. J. (2018). A resource-rational theory of set size effects in human visual working memory. *eLIFE, 7*, e34963. https://doi.org/10.7554/eLife.34963

van den Berg, R., Shin, H., Chou, W.-C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences, 109*, 8780–8785.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review, 14*, 779–804.

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

Wagenmakers, E.-J., Beek, T. F., Rotteveel, M., Gierholz, A., Matzke, D., Steingroever, H., … Pinto, Y. (2015). Turning the hands of time again: A purely confirmatory replication study and a Bayesian analysis. *Frontiers in Psychology, 6*, 494. https://doi.org/10.3389/fpsyg.2015.00494

Wagenmakers, E.-J., Dutilh, G., & Srafoglou, A. (2018a). The creativity-verification cycle in psychological science: New methods to combat old idols. *Perspectives on Cognitive Science, 13*, 418–427. https://doi.org/10.1177/1745691618771357

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., … Morey, R. D. (2018b). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review, 25*, 35–57. https://doi.org/10.3758/s13423-017-1343-3

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*, 632–638. https://doi.org/10.1177/1745691612463078

Williams, L. E., & Bargh, J. A. (2008). Experiencing physical warmth promotes interpersonal warmth. *Science, 322*, 606–607. https://doi.org/10.1126/science.1162548

Wills, A. J., & Pothos, E. M. (2012). On the adequacy of current empirical evaluations of formal models of categorization. *Psychological Bulletin, 138*, 102–125. https://doi.org/10.1037/a0025715

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review, 114*, 152–176.