ARTICLE TEMPLATE

Why does preregistration increase the persuasiveness of evidence? A Bayesian rationalization.

Aaron Peikert[a,b], Andreas M. Brandmaier[a,c,d]

[a]Center for Lifespan Psychology, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany; [b]Department of Psychology, Humboldt-Universität zu Berlin, Unter den Linden 6, 10117 Berlin, Germany; [c]MSB Medical School Berlin, Rüdesheimer Str. 50, 14197 Berlin, Germany; [d]Max Planck UCL Centre for Computational Psychiatry and Ageing Research Berlin, Germany and London, UK

ARTICLE HISTORY
Compiled February 16, 2023

ABSTRACT
The replication crisis has led many researchers to preregister their hypotheses and data analysis plans before collecting data. A widely held view is that preregistration is supposed to limit the extent to which data may influence the hypotheses to be tested. Only if data have no influence an analysis is considered confirmatory. Consequently, many researchers believe that preregistration is only applicable in confirmatory paradigms. In practice, researchers may struggle to preregister their hypotheses because of vague theories that necessitate data-dependent decisions (aka exploration). We argue that preregistration benefits any study on the continuum between confirmatory and exploratory research. To that end, we formalize a general objective of preregistration and demonstrate that exploratory studies also benefit from preregistration. Drawing on Bayesian philosophy of science, we argue that preregistration should primarily aim to reduce uncertainty about the inferential procedure used to derive results. This approach provides a principled justification of preregistration, separating the procedure from the goal of ensuring strictly confirmatory research. We acknowledge that knowing the extent to which a study is exploratory is central, but certainty about the inferential procedure is a prerequisite for persuasive evidence. Last, we discuss the implications of these insights for the practice of preregistration.

KEYWORDS
preregistration; confirmation; exploration; hypothesis testing; bayesian; open science

The scientific community has long pondered the vital distinction between exploration and confirmation, discovery and justification, hypothesis generation and hypothesis testing, or prediction and postdiction (Hoyningen-Huene 2006; Shmueli 2010; Nosek et al. 2018). Despite the different names, it is fundamentally the same dichotomy that is at stake here. There is a broad consensus that both approaches are necessary for science to progress; exploration for making new discoveries and confirmation for exposing these discoveries to potential falsification and assess in how far they have been corroborated. However, mistaking exploratory findings for empirically confirmed results is dangerous. It inflates the likelihood of believing that there is evidence sup-

CONTACT Aaron Peikert. Email: peikert@mpib-berlin.mpg.de, Andreas M. Brandmaier. Email: brandmaier@mpib-berlin.mpg.de

porting a given explanation even if it is false. A variety of problems, such as researchers' degrees of freedom together with researchers' hindsight bias or naive p-hacking have led to such mistakes becoming commonplace yet unnoticed for a long time. Recognizing them has led to a crisis of confidence in the empirical sciences (Ioannidis 2005), and psychology in particular (Open Science Collaboration 2015). As a response to the crisis, more and more researchers preregister their hypotheses and their data collection and analysis plans in advance of the study (Nosek et al. 2018). They do so to stress the predictive nature of their registered statistical analyses; often with the hopes of obtaining a label that marks the study as "confirmatory". Indeed, rigorous application of preregistration prevents researchers from reporting a set of results produced by an arduous process of trial and error as a simple confirmatory story (Wagenmakers et al. 2012) while keeping low false-positive rates. This promise of a clear distinction between confirmation and exploration has obvious appeal to many who already accepted the practice. Still, the majority of empirical researchers do not routinely preregister their studies. One reason may be that some do not find that the theoretical advantages outweigh the practical hurdles, such as specifying every aspect of a theory in advance. We believe that we can reach a greater acceptance of preregistration by explicating a more general objective of preregistration that benefits all kinds of studies, even those that allow data-dependent decisions.

One goal of preregistration that has received widespread attention, is to clearly distinguish confirmatory from exploratory research (Mellor and Nosek 2018; Nosek et al. 2018; Wagenmakers et al. 2012; Simmons, Nelson, and Simonsohn 2021; Bakker et al. 2020). In such a narrative, preregistration is justified by a confirmatory research agenda. However, two problems become apparent under closer inspection. First, many researcher do not subscribe to a purely confirmatory research agenda. Second, it is not necessary to strictly map the categories preregistered vs. non-preregistered onto the categories confirmatory vs. exploratory research.

Obviously, researchers can conduct confirmatory research without preregistration—though it might be difficult to convince other researchers of the confirmatory nature of their research, that is, that they were free of cognitive biases, made no data-dependent decisions whatsoever, and so forth. The opposite, that is, preregistered but not strictly confirmatory studies, are also becoming more commonplace (Dwan et al. 2008; Chan et al. 2004; Silagy, Middleton, and Hopewell 2002).

Researchers may apply two strategies to evade the self-imposed restrictions of preregistrations: writing a loose preregistration to begin with (Stefan and Schönbrodt 2022) or deviating from the preregistration afterwards. Both strategies may be used for sensible scientific reasons or with the self-serving intent of generating desirable results. Thus, insisting on equating preregistration and confirmation has led to the criticism that, all things considered, preregistration is actually harmful, and neither sufficient nor necessary for doing good science (Szollosi et al. 2020; Pham and Oh 2021).

We argue that such criticism is not directed against preregistration itself but against a justification through a confirmatory research agenda (Wagenmakers et al. 2012). When researchers criticize preregistration as being too inflexible to fit their research question, they often simply acknowledge that their research goals are not strictly confirmatory. Forcing researchers into adopting a strictly confirmatory research agenda does not only imply changing *how* they investigate a phenomenon but also *what* research questions they ask. However reasonable such a move is, changing core believes of a large community is much harder than convincing them that a method is well justified. We therefore attempt to disentangle the *methodological* goals of preregistration

from the *ideological* goals of confirmatory science. It is likely that psychology needs more confirmatory studies to progress as a science, but quite independently of this, preregistration can be useful for any kind of study on the continuum between strictly confirmatory and fully exploratory.

To form such an objective for preregistration, we first introduce some tools of Bayesian philosophy of science and map the exploration/confirmation distinction onto a dimensional quantity we call "theoretical risk" (a term borrowed from Meehl 1978, but formalized as the probability of proving a hypothesis wrong, if it does not hold), which is inversely related to the type-I error rate in statistical test theory.

Further, we outline two interpretations of how preregistration impacts theoretical risk. The first interpretation corresponds to the traditional application of preregistration to research paradigms that focus on confirmation by maximizing the theoretical risk or equivalently by limiting type-I error (when dichotomous decisions about theories are an inferential goal). The second interpretation is our main contribution and demonstrates the broad applicability of preregistration to both exploratory and confirmatory studies that are implemented as preregistered or have undergone changes after preregistration. We argue that the classic view on the utility of preregistration can be interpreted as maximization of theoretical risk, which is reduced by researchers' degrees of freedom, p-hacking, and such. Importantly, we argue that theoretical risk is not necessarily directly maximized by preregistration, but rather the uncertainty in judging the theoretical risk is minimized.

To arrive at this interpretation, we rely on three arguments. The first is that theoretical risk is vital for judging evidential support for theories. The second argument is that the theoretical risk for a given study is generally uncertain. The third and last argument is that this uncertainty is reduced by applying preregistration. We conclude that because preregistration decreases uncertainty about the theoretical risk, which in turn increases our expectation to gain evidence for a theory, preregistration is potentially useful for any kind of study, no matter how exploratory.

1. Epistemic value and the Bayesian rationale

Let us start by defining what we call expected epistemic value. If researchers plan to conduct a study, they usually hope that it will change their assessment of some theory's verisimilitude (Niiniluoto 1998). In other words, they hope to learn something from conducting the study. The amount of knowledge researchers gain from a particular study concerning the verisimilitude of a specific theory (which itself is an ontologocial concept) is what we call epistemic value. Researchers cannot know what exactly they will learn from a study before they have run it. However, they can develop an expectation that helps them decide about the specifics of a planned study. This expectation is what we term expected epistemic value. To make our three arguments, we must assume three things about this estimation process and how it relates to what studies (preregistered vs not preregistered) to conduct.

1. Researchers judge the evidence for or against a hypothesis rationally.
2. They expect other researchers to apply a similar rational process.
3. Researchers try to maximize the expected epistemic value for other researchers.

The assumption of rationality can be connected to Bayesian reasoning and leads to our adoption of the framework. Our rationale is as follows. Researchers who decide to conduct a certain study are actually choosing a study to bet on. They have to

"place the bet" by conducting the study, therefore, invest resources and stand to gain epistemic value with some probability. This conceptualization of choosing a study as a betting problem allows us to apply a "Dutch Book" argument (Christensen 1991). This argument states that any better must follow the axioms of probability to avoid being "irrational," i.e., accepting bets that lead to sure losses. Fully developing a Dutch book argument for this problem requires careful consideration of what kind of studies to include as possible bets, defining a conversion rate from the stakes to the reward, and modelling what liberties researchers have in what studies to conduct. Without deliberating these concepts further, we find it persuasive that researchers should not violate the axioms of probability if they have some expectation about what they stand to gain with some likelihood from conducting a study. The axioms of probability are sufficient to derive the Bayes formula, on which we will heavily rely for our further arguments. The argument is not sufficient, however, to warrant conceptualizing the kind of epistemic value we reason about in terms of posterior probability; that remains a leap of faith. However, the argument applies to any reward function that satisfies the "statistical relevancy condition" (Salmon 1970; Fetzer 1974). That is, evidence only increases epistemic value for a theory, if the evidence is more likely to observe under the theory than without it.

Please note that our decision to adopt this aspect of the Bayesian philosophy of science does not imply anything about the statistical methods researchers use. In fact, this conceptualization is intentionally reductionistic to be compatible with a wide range of philosophies of science and statistical methods researchers might subscribe to.

2. Epistemic value and theoretical risk

Our first argument is that theoretical risk is crucial for judging evidential support for theories. Put simply, risky predictions create persuasive evidence if they turn out to be correct. This point is crucial because we attribute much of the appeal of a confirmatory research agenda to this notion.

Let us make some simplifying assumptions and define notation. We restrict ourselves to evidence of a binary nature (either it was observed or not) since continuous evidence would lead to some quite involved derivations. We denote the probability of a hypothesis before observing evidence as $P(H)$ and its complement as $P(\neg H) = 1 - P(H)$. The probability of observing evidence under some hypothesis is $P(E|H)$. We can calculate the probability of the hypothesis after observing the evidence with the help of the Bayes formula:

$$P(H|E) = \frac{P(H)P(E|H)}{P(E)} \tag{1}$$

The posterior probability $P(H|E)$ is of great relevance since it is often used directly or indirectly as a measure of corroboration of a hypothesis. In the tradition of Carnap, in its direct use, it is called corroboration as firmness; in its relation to the a priori probability $P(H)$, it is called increase in firmness (Carnap 1950, preface to the 1962 edition). As noted before, we concentrate on posterior probability as a measure of epistemic value, since no measure shows universally better properties than others. However, it is reasonable that any measure of corroboration increases monotonically

with an increase in posterior probability $P(H|E)$, and our argument applies to those measures as well. We assume that any measure shows gain if, and only if, evidence is observed assuming $P(E|H) > P(E|\neg H)$. We referred to this assumption as "statistical relevance condition" earlier (Salmon 1970; Fetzer 1974).

In short, we want to increase posterior probability $P(H|E)$. Increases in posterior probability $P(H|E)$ are associated with increased epistemic value, for which we want to maximize the expectation. So how can we increase posterior probability? The Bayes formula yields three components that influence corroboration, namely $P(H)$, $P(E|H)$ and $P(E)$. The first option leads us to the unsurprising conclusion that higher a priori probability $P(H)$ leads to higher posterior probability $P(H|E)$. If a hypothesis is more probable to begin with, observing evidence in favor will result a more corroborated hypothesis, all else equal. However, the prior probability of a hypothesis is nothing our study design can change. The second option is similarly commonsensical; that is, an increase in $P(E|H)$ leads to higher posterior probability $P(H|E)$. $P(E|H)$ is the probability of obtaining evidence for a theory, when the theory holds. We call this probability of detecting evidence given that the theory holds "detectability." Consequently, researchers should ensure that their study design allows them to find evidence for their hypothesis, in case it is true. When applied strictly within the bounds of null hypothesis testing, detectability is equivalent to power (or the inverse of type-II error rate). However, while detectability is of great importance for study design, it is not directly relevant to the objective of preregistration. Thus, $P(E)$ remains to be considered. Since $P(E)$ is the denominator, increasing it will decrease the posterior probability: The more unlikely it is to observe evidence, the more it increases the probability of the hypothesis if we do observe it. In other words, high risk, high reward.

If we equate riskiness with a low probability of obtaining evidence, the Bayesian rationale perfectly aligns with the observation that risky predictions lead to persuasive evidence. This tension between high risk leading to high reward is central to our consideration of preregistration. A high-risk, high-reward strategy is bound to result in many losses that are eventually absorbed by the high gains. Sustaining many "failed" studies is not exactly aligned with the incentive structure under which many, if not most, researchers operate. Consequently, researchers have an incentive to appear to take more risks than they actually do, which misleads their readers to give their claims more credence than they deserve. It is at this juncture that the practice and mispractice of preregistration comes into play. We argue that the main function of preregistration is to enable proper judgment of the riskiness of a study.

To better understand how preregistrations can achieve that, let us take a closer look at what factors contribute to $P(E)$. Using the law of total probability, we can split $P(E)$ into two terms:

$$P(E) = P(H)P(E|H) + P(\neg H)P(E|\neg H) \tag{2}$$

We already have noted that there is not much to do about prior probability ($P(H)$, and hence its counter probability $P(\neg H)$), and that it is common sense to increase detectability $P(E|H)$. The real lever to pull is, therefore, $P(E|\neg H)$. This probability tells us how likely it is that we find evidence in favor of the theory when in fact, the theory is not true. Its counter probability $P(\neg E|\neg H) = 1 - P(E|\neg H)$ is what we call "theoretical risk", because it is the risk a theory takes on in predicting the occurrence of particular evidence in its favour. We "borrow" the term from Meehl (1978), though

he has not assigned it to the probability $P(\neg E|\neg H)$. Kukla (1990) argued that the core arguments in Meehl (1990) can be reconstructed in purely Bayesian framework. However, while he did not name $P(\neg E|\neg H)$ but did suggest that Meehl (1978) used the term "very strange coincidence" for a small $P(E|\neg H)$ which would imply that $P(\neg E|\neg H)$ can be related to or even equated to theoretical risk.

Let us note some interesting properties of theoretical risk $P(\neg E|\neg H)$. First, increasing theoretical risk leads to higher posterior probability $P(H|E)$, our objective. Second, if the theoretical risk is smaller than detectability $P(E|H)$ it follows that the posterior probability must decrease when observing the evidence. If detectability exceeds theoretical risk, the evidence is less likely under the theory than it is when the theory does not hold. Third, if the theoretical risk equals zero, then posterior probability is at best equal to prior probability but only if detectability is perfect ($P(H|E) = 1$). In other words, observing a sure fact does not lend credence to a hypothesis.

The last statement sounds like a truism but is directly related to Popper's seminal criterion of demarcation. He stated that if it is impossible to prove that a hypothesis is false ($P(\neg E|\neg H) = 0$, theoretical risk is zero), it cannot be considered a scientific hypothesis (Popper 2002, p. 18). We note these relations to underline that the Bayesian rational we apply here is able to reconstruct many commonly held views on riskiness and epistemic value.

Both theoretical risk $P(\neg E|\neg H)$ and detectability $P(E|H)$ aggregate uncountable influences, otherwise they could not model the process of evidential support for theories. To illustrate the concepts we introduced up to here, consider the following example of a single theory and three experiments that may test it. The experiments were created to illustrate how they may differ in their theoretical risk and detectability. Suppose the primary theory is about the cognitive phenomenon of "insight." For the purpose of illustration, we define it, with quite some hand-waving, as a cognitive abstraction that allows agents to consistently solve a well-defined class of problems. We present the hypothesis that the following problem belongs to such a class of insight problems:

> Use five matches (IIIII) to form the number eight.

We propose three experiments that differ in theoretical risk and detectability. All experiments take a sample of ten psychology students. We present the students with the problem for a brief span of time. After that, the three experiments differ as follows:

1. The experimenter gives a hint that the problem is easy to solve when using Roman numerals; if all students come up with the solution, she records it as evidence for the hypothesis.
2. The experimenter shows the solution "IX" and explains it; if all students come up with the solution, she records it as evidence for the hypothesis.
3. The experimenter does nothing; if all students come up with the solution, she records it as evidence for the hypothesis.

We argue that experiment 1 has high theoretical risk $P(\neg E_1|\neg H)$ and high detectability $P(E_1|H)$. If "insight" has nothing to do with solving the problem ($\neg H$), then presenting the insight that Roman numerals could be used, should not lead to all students solving the problem ($\neg E_1$); the experiment therefore has high theoretical risk $P(\neg E_1|\neg H)$. Conversely, if insight is required to solve the problem (H), then it is likely to help all students to solve the problem (E_1), the experiment therefore has high detectability $P(E_1|H)$. The second experiment, on the other hand, has low theoretical risk $P(\neg E_2|\neg H)$. Even if "insight" has nothing to do with solving the problem ($\neg H$), there are other plausible reasons for observing the evidence (E_2), because the

students could simply copy the solution, without having any insight. With regard to detectability, experiments 1 and 2 differ in no obvious way. Experiment 3, however, also has low detectability. It is unlikely that all students come up with the correct solution in a short time (E_3), even if insight is required (H) experiment 3 therefore has low detectability $P(E_3|H)$. The theoretical risk, however, is also low in absolute terms, but high compared to the detectability (statistical relevancy condition is satisfied). In the unlikely event that all 10 students place their matches to form the Roman numeral VIII (E_3), it is probably due to insight (H) and not by chance $P(\neg E_3|\neg H)$). Of course, in practice, we would allow the evidence to be probabilistic, e.g., relax the requirement of "all students" to nine out of ten students, more than eight, and so forth.

As argued earlier, the remainder of the paper will focus on binary, non-probabilisitic evidence to keep the mathematical notation as simple as possible. We discuss the relation between statistical methods and theoretical risk in the Statistical Methods section.

3. Preregistration as a means to increase theoretical risk?

Having discussed that increasing the theoretical risk will increase the epistemic value, it is intuitive to task preregistration with maximizing theoretical risk, i.e., a confirmatory research agenda. Indeed, limiting the type-I error rate is commonly stated as *the* central goal of preregistration (Nosek et al. 2018; Oberauer 2019; Rubin 2020). We argue that while such a conclusion is plausible, we must first consider at least two constraints that place an upper bound on the theoretical risk.

First, the theory itself limits theoretical risk: Some theories simply do not make risky predictions, and preregistration will not change that. Consider the case of a researcher contemplating the relation between two sets of variables. Suppose each set is separately well studied, and strong theories tell the researcher how the variables within the set relate. However, our imaginary researcher now considers the relation between these two sets. For lack of a better theory, they assume that some relation between any variables of the two sets exists. This is not a risky prediction to make in psychology, even without statistical issues like alpha inflation (Orben and Lakens 2020). However, we would consider it a success if the researcher would use the evidence to develop a more precise (and therefore risky) theory, e.g., by specifying which variables from one set relate to which variables from the other set, to what extent, in which direction, with which functional shape, etc. We will later show that preregistration increases the degree of belief in the further specified theory, though it remains low till being substantiated by testing it again. The point, however, is that we want to show that preregistration increases the expected epistemic value regardless of the theory being tested.

Second, available resources limit theoretical risk. Increasing theoretical risk $P(\neg E|\neg H)$ will usually decrease detectability $P(E|H)$ unless more resources are invested. In other words, one cannot increase power while maintaining the same type-I error rate without increasing the invested resources. Tasking preregistration with an increase in theoretical risk makes it difficult to balance this trade-off. Mindlessly maximizing theoretical risk would either never produce evidence or require huge amounts of resources.

4. Uncertainty about theoretical risk

We have established that higher theoretical risk leads to more persuasive evidence. In other words, we have reconstructed the interpretation that preregistrations supposedly work by restricting the researchers, which in turn increases the theoretical risk (or equivalently limits the type-I error rate) and thereby creates more compelling evidence. Nevertheless, there are trade-offs for increasing theoretical risk. Employing a mathematical framework allows us to navigate the trade-offs more effectively and move towards a second, more favorable interpretation. To that end, we incorporate uncertainty into our framework.

4.1. *Statistical methods*

One factor that is known without much uncertainty is the contribution of statistical methods to theoretical risk. Theoretical risk is deeply connected with statistical methods, because it is the inverse of $P(E|\neg H)$. $P(E|\neg H)$ is equivalent to the type-I error rate in statistical hypothesis testing, if you consider the overly simplistic case where the research hypothesis is equal to the statistical alternative hypothesis, because then the nill hypothesis is $\neg H$. Because many researchers are familiar with the type-I error rate, it can be helpful to remember this connection to theoretical risk. Researchers who choose a smaller type-I error rate can be more sure of their results, if significant, because the theoretical risk is higher. However, the research hypothesis seldomly equals the statistical null hypothesis, and therefore, the relation between statistical type-I error rate and theoretical risk should not be overinterpreted. We argue that theoretical risk (and hence its inverse, $P(E|\neg H)$) also encompasses factors outside the statistical realm, most notably the study design and broader analytical strategies.

Statistical methods stand out among these factors because we have a large and well-understood toolbox for assessing and controlling their contribution to theoretical risk. Examples of our ability to exert this control are the choice of type-I error rate, adjustments for multiple testing, the use of corrected fit measures (i.e., adjusted R^2), information criteria, or cross-validation in machine learning. These tools help us account for biases in statistical methods that increase the likelihood of signifying results even when there are none $P(E|\neg H)$.

The point is that the contribution of statistical methods to theoretical risk can be formally assessed. For many statistical models it can be analytically computed under some assumptions. For those models or assumptions where this is impossible, one can employ Monte Carlo simulation to estimate the contribution to theoretical risk. The precision with which statisticians can discuss contributions to theoretical risk has lured the community concerned with research methods into ignoring other factors that are much more uncertain. We cannot hope to resolve this uncertainty; but we have to be aware of its implications. These are presented in the following.

4.2. *Sources of Uncertainty*

As we noted, it is possible to quantify how statistical models affect the theoretical risk based on mathematical considerations and simulation. However, other factors in the broader context of the study are much harder to quantify. If one chooses to focus only on the contribution of statistical methods to theoretical risk, one is bound to overestimate it. Take, for example, a t-test of mean differences in two samples.

Under ideal circumstances (assumption of independence, normality of residuals, equal variance), it stays true to its type-I error rate. However, researchers might do many very reasonable things in the broader context of the study that affect theoretical risk: They might exclude outliers, choose to drop an item before computing a sum score, broaden, enlarge their definition of the population to be sampled, translate their questionnaires into a different language, impute missing values, switch between different estimators of the pooled variance, or any number of other things. All of these decisions carry a small risk that they increase the likelihood of obtaining evidence despite the underlying research hypothesis being false. Even if the t-test itself perfectly maintains its type I error rate, these factors influence $P(E|\neg H)$. While, in theory, these factors may leave $P(E|\neg H)$ unaffected or even decrease it, we argue that this is not the case in practice. Whether researchers want to or not, they continuously process information about how the study is going, except under strict blinding. While one can hope that processing this information does not affect their decision making either way, this cannot be secured. We, therefore, conclude that statistical properties only guarantee a lower bound for theoretical risk. The only thing we can conclude with some certainty is that theoretical risk is not higher than what the statistical model guarantees without knowledge about the other factors at play.

4.3. *The effects of uncertainty*

Before we ask how preregistration influences this uncertainty, we must consider the implications of being uncertain about the theoretical risk. Within the Bayesian framework, this is both straightforward and insightful. To get an expectation, we express uncertainty as a probability distribution and then integrate over it:

$$\mathbb{E}(p(H|E)) = \int \frac{p(H)p(E|H)}{p(H)p(E|H) + p(\neg H)p(E|\neg H)} \, \mathrm{d}\, \mathbb{P}(p(E|\neg H)) \tag{3}$$

To illustrate the effect of uncertainty, let $p(E|H) = 0.8$ (e.g., power of 80%) and $p(H) = 0.1$ and assume a uniform distribution for theoretical risk $p(\neg E|\neg H)$ of the form:

$$f(x) = \begin{cases} \frac{1}{\tau} & \text{for } 0 \leq x \leq \tau, \\ 0 & \text{for } x < 0 \text{ or } x > \tau \end{cases} \tag{4}$$

Where τ is the upper bound of theoretical risk (e.g., 0.95 for a statistical model with a nominal type-I error rate of 5%). We chose this uniform distribution to capture our statement that a statistical model only guarantees an upper bound to theoretical risk (and since it is a probability the lower bound is 0) as it is the maximum entropy distribution under this assumption and conforms therefore to our Bayesian framework (Giffin and Caticha 2007).

Figure 1 shows exemplary the effect of theoretical risk (x-axis) on the posterior probability (y-axis) being fully certain (solid line) or fully uncertain (dashed line) about the theoretical risk of a study. Neither absolute certainty nor uncertainty are realistic scenarios but represent the boundary conditions into which all realistic conditions fall. Depending on how uncertain we are about the theoretical risk a study took on,

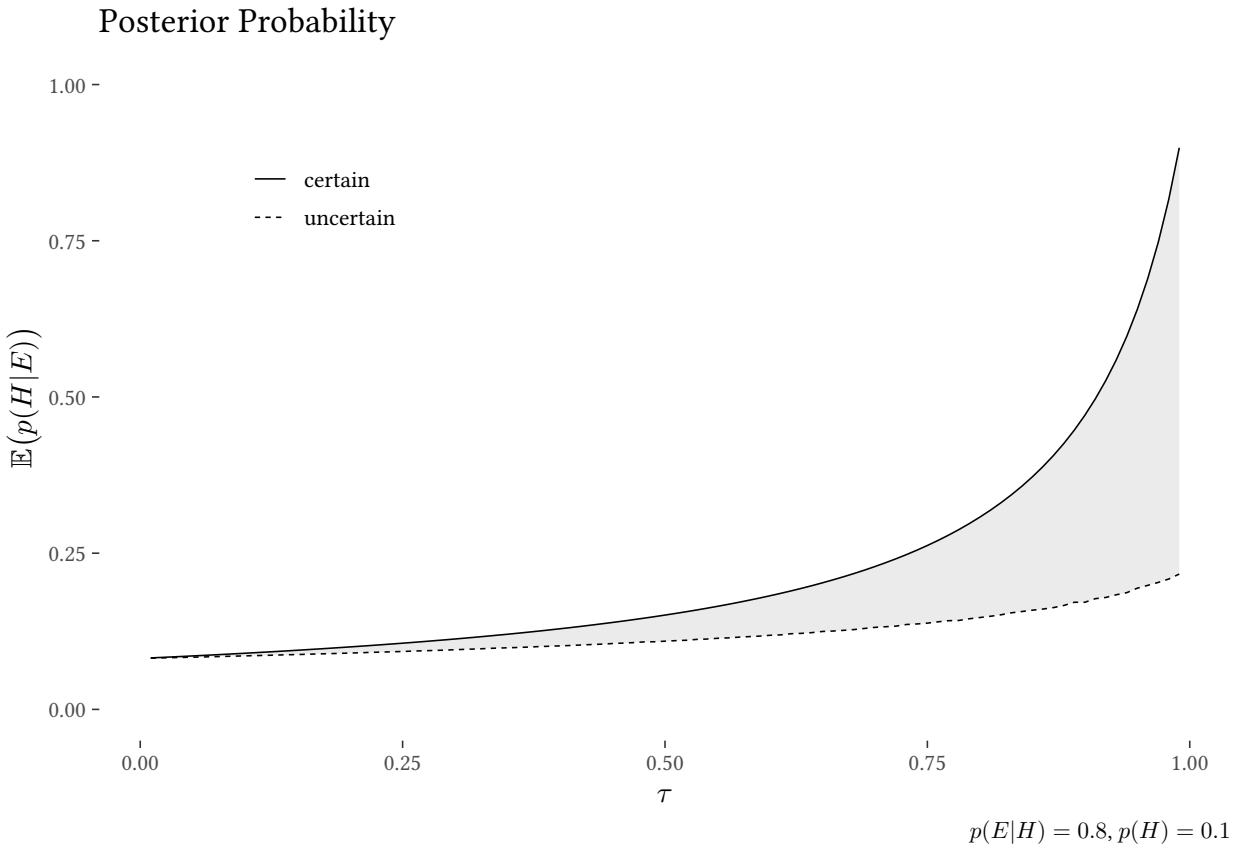

Figure 1. Posterior Probability (corroboration as firmness) as a function of theoretical risk τ, where τ is either certain (solid line) or maximally uncertain (dotted line).

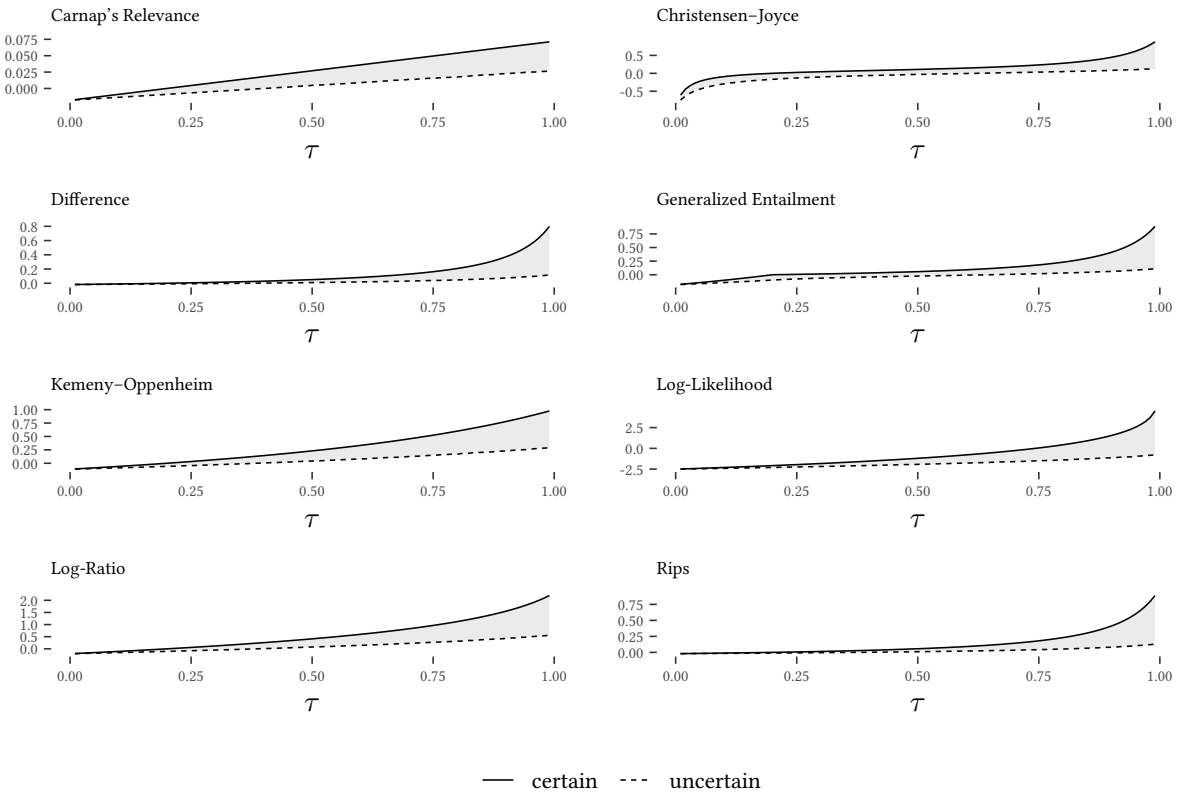

Figure 2. Several measures for corroboration as increase in firmness as a function of τ, where τ is either certain (solid line) or maximally uncertain (dotted line).

our expectation the gained epistemic value varies considerably. Mathematically, uncertainty about theoretical risk is expressed through the variance (or rather entropy) of the distribution. Generally, we expect that increases in uncertainty (expressed as more entropic distributions) lead to a decreased expected epistemic value.

The argument for a confirmatory research agenda is that by increasing theoretical risk we increase expected epistemic value, i.e., moving to the right on the x-axis in Figure 1 increases posterior probability (on the y-axis). However, if a hypothesis in a certain study has low theoretical risk, there is not much researchers can do about it. However, studies do not only differ by how high the theoretical risk is but also how certain the recipient is about the theoretical risk. A study that has a very high theoretical risk, e.g., 1% chance that if the hypothesis is wrong, evidence in its favor will be observed, i.e., alpha error, but has also maximum uncertainty will result in a posterior probability of 22%, while the same study with maximum certainty will result in 90% posterior probability. The other factors (detectability, prior beliefs, measure of epistemic value) and, therefore, the extend of the benefit very of course with the specifics of the study. Crucially, even studies with some exploratory aspects benefit from preregistration, e.g., in this scenario with a $\tau = 0.8$, so false positive rate of 0.2, moving from uncertain to certain increases the posterior by a factor of 2.09.

5. Preregistration as a means to decrease uncertainty about the theoretical risk

We hope to have persuaded the reader to accept two arguments: First, the theoretical risk is important for judging evidential support for theories. Second, the theoretical risk is inherently uncertain and the degree of uncertainty diminishes the persuasiveness of the gathered evidence. The third and last argument is that preregistrations reduce this uncertainty. Following the last argument, a preregistered study is represented by the solid line (certainty about theoretical risk) and a study that was not preregistered is more similar to the dashed line (maximally uncertain about theoretical risk) in Figure 1 and Figure 2.

Let us recall our three assumptions:

1. Researchers judge the evidence for or against a hypothesis rationally.
2. They expect other researchers to apply the same rational process.
3. All else being equal, researchers try to increase the expected epistemic value for other researchers.

The point we make with these assumptions is that researchers aim to persuade other researchers, for example, the readers of their articles. Not only the original authors are concerned with the process of weighing evidence for or against a theory but really all people the authors of a study hope to persuade. Unfortunately, readers of a scientific article (or, more generally, any consumer of a research product), is likely to lack insight into the various factors that influence theoretical risk. While the authors themselves may have a clear picture of what they did and how it might have influenced the theoretical risk they took, their readers have much greater uncertainty about these factors. In particular, they never know which relevant factors the authors of a given article failed to disclose, be it intentionally or not. From the perspective of the ultimate sceptic, they may claim maximum uncertainty.

Communicating clearly how authors of a scientific report gathered their data and consequently analyzed it to arrive at the evidence they present is crucial for judging the theoretical risk they took. Preregistrations are ideal to communicate just that, because any description after the fact is suspect to be incomplete. For instance, the authors could have opted for selective reporting, that is, they decided to exclude a number of analytic strategies they tried out. That is not to say that every study that was not-preregistered was subjected to practices of questionable research practices. The point is, that we cannot exclude it with certainty. This uncertainty is drastically reduced, if the researchers have described what they intended to do beforehand, and then report that they did exactly that. In that case, the readers can be certain, that they have received a complete account of the situation. They still might be uncertain about the actual theoretical risk the authors took, but to a much smaller extend than if the study would not have been preregistered. Remaining sources of uncertainty might be unfamiliarity with statistical methods or experimental paradigms used, the probability of an implementation error in the statistical analyses, a bug in the software used for analyses, etc. In any case a well written preregistration should aim to reduce the uncertainty about the theoretical risk and hence increase the persuasiveness of evidence. Therefore, a study that perfectly adhered to its preregistration will resemble the solid line in Figure 1/2. Crucially, perfect means here that the theoretical risk can be judged with low uncertainty, not that the theoretical risk is necessarily high.

6. Discussion

To summarize, we showed that both higher theoretical risk and lower uncertainty about theoretical risk lead to higher expected epistemic value across a variety of measures. The former result, that increasing theoretical risk leads to higher expected epistemic value, reconstructs the appeal and central goal of preregistration ofr confirmatory research agendas. However, theoretical risk is something researcher have only limited control over, for example, theories are often vague and ill-defined, resources are limited, and increasing theoretical risk usually decreases detectability of a hypothesized effect (a special instance of this trade-off is the well known tension between type-I error and statistical power). While we believe that preregistration is always beneficial, it might be counterproductive to pursue high theoretical risk if the research context is inappropriate for strictly confirmatory research. Specifically, appropriateness here entails the development of precise theories and the availability of necessary resources (often, large enough sample size, but also see Brandmaier et al. (2015)) to adequately balance detectability against theoretical risk.

In terms of preparing the conditions for confirmatory research, preregistration may at most help to invest some time into developing more specific, hence riskier, implications of a theory. But for a confirmatory science, it will not be enough to preregister all studies. This undertaking requires action from the whole research community (Lishner 2015). Incentive structures must be created to evaluate not the outcomes of a study but the rigor with which it was conducted (Cagan 2013; Schönbrodt et al. 2022). Journal editors could encourage theoretical developments that allow for precise predictions that will be tested by other researchers and be willing to accept registered reports (Fried 2020a,b; van Rooij and Baggio 2020, 2021). Funding agencies should demand an explicit statement about theoretical risk in relation to detectability and must be willing to provide the necessary resources to reach adequate levels of both (Koole and Lakens 2012).

Our latter result, on the importance of preregistration for minimizing uncertainty, has two important implications. The first is, that even if all imaginable actions regarding promoting higher theoretical risk are taken, confirmatory research should be preregistered. Otherwise the uncertainty about the theoretical risk will diminish the advantage of confirmatory research. Second, even under less than ideal circumstances for confirmatory research, preregistration is beneficial. Preregistering exploratory studies increases the expected epistemic value by virtue of reducing uncertainty about theorethical risk. Nevertheless, exploratory studies will have a lower expected epistemic value than a more confirmatory study, if both are preregistered and have equal detectability.

Focusing on uncertainty reduction also explains two common practices of preregistration that do not align with a confirmatory research agenda. First, researchers seldomly predict precise numerical outcomes, instead they use preregistrations to describe the process that generates the results. Precise predictions would have very high theoretical risk (they are likely incorrect if the theory is wrong). A statistical procedure, may have high or low theoretical risk depending on the specifics of the model used. Specifying the process, therefore, is in line with the rational we propose here, but is incompatible with strictly confirmatory research.

Second, researchers often have to deviate from the preregistration and make data-dependent decisions after the preregistration. If the only goal of preregistration is to ensure confirmatory research, such changes are not justifiable. However, under our rational some changes may be justified. Any change increases the uncertainty about the

theoretical risk and may even decrease the theorethical risk. The changes still may be worthwhile, if the negative outcomes may be offset by an increase in detectability due to the change. Consider a preregistration that failed to specify how to handle missing values, and researchers subsequently encountering missing values. In such case, detectability becomes zero because the data cannot be analyzed without a post-hoc decision about how to handle the missing data. Any such decision would constitute a deviation from the preregistration, which is perfectly waranted under our general objective. Note that a reader cannot rule out that the researchers leveraged the decision to decrease theoretical risk, i.e., picking among all options the one that delivers the most beneficial results for the theory (in the previous example, chosing between various options of handling missing values). Whatever decision they make, increased uncertainty about the theoretical risk is inevitable and the expected epistemic value is decreased compared to a world where they anticipated the need to deal with missing data. However, it is still justified to deviate. After all they have not anticipated the case and are left with a detectablilty of zero. Any decision will increase detectability to a non-zero value, likely offsetting the increase in uncertainty. The researchers also may do their best to argue that the deviation was not motivated by increasing theoretical risk, thereby, decreasing the uncertainty. Ideally, there is a default decision that fits well with the theory or with the study design. Or, if there is no obvious candidate, the researchers could conduct a multiverse analysis of the available options to deal with missings to show the influence of the decision (Steegen et al. 2016).

As explained above, reduction in uncertainty as the objective for preregistration does not only explain some existing behavior, that does not align with confirmation as a goal, it also allows to form recommendations to improve the practice of preregistration. Importantly, we now have a theoretical measure to gauge the functionality of preregistrations, which can only help increase its utility. In particular, a preregistration should be specific about the procedure that is intended to generate evidence for a theory. Such procedure may accommodate a wide range of possible data, i.e., it may be exploratory. The theoretical risk, however low, must be communicated clearly. Parts of the process left unspecified, imply uncertainty, which is what a preregistration should reduce. However, specifying procedures that can be expected to not work, will lead to deviation, and subsequently to larger uncertainty.

We have proposed a workflow for preregistration called preregistration as code (PAC) elsewhere (Peikert, van Lissa, and Brandmaier 2021). In a PAC, researchers use computer code for the planned analysis as well as a verbal description of theory and methods for the preregistration. This combination is facilitated by dynamic document generation, where the results of the code, such as numbers, figures, and tables, are inserted automatically into the document. The idea is that the preregistration already contains "mock results" based on simulated or pilot data, which are replaced after the actual study data becomes available. Such approach dissolves the distinction between the preregistration document and the final scientific report. Instead of fundamentally separate documents, preregistration, and final report are different versions of the same underlying dynamic document. Deviations from the preregistration can therefore be clearly (and if necessary, automatically) isolated, highlighted, and inspected using version control. Crucially, because the preregistration contains code, it may accommodate many different data patterns, i.e., it may be exploratory. However, while a PAC is not the extend of exploration, it is very specific about exactly how probable it is to generate evidence even when the theory does not hold (theoretical risk). Please note that while PAC is ideally suited to reduce uncertainty about theoretical risk, other more traditional forms of preregistration are also able to advance this goal.

Contrary to what is widely assumed about preregistration, a preregistration is not necessarily a seal of confirmatory research. Confirmatory research would almost always be less persuasive without preregistration, but in our view, preregistration primarily communicates the extent of confirmation, i.e., theoretical risk, of a study. Clearly communicating theoretical risk is important because it reduces the uncertainty and hence increases expected epistemic value.

Acknowledgement

We thank Maximilian S. Ernst, Caspar van Lissa, Felix Schönbrodt, the discussants at the DGPS2022 conference and Open Science Center Munich, and many more for the insightful discussions about disentangling preregistration and confirmation.

Notes

The materials for this article are available on GitHub (**?**). This version was created from git commit `a8a8958`. The manuscript is available as preprint (**?**) and will be submitted to Philosophical Psychology.

References

Bakker, Marjan, Coosje L. S. Veldkamp, Marcel A. L. M. van Assen, Elise A. V. Crompvoets, How Hwee Ong, Brian A. Nosek, Courtney K. Soderberg, David Mellor, and Jelte M. Wicherts. 2020. "Ensuring the Quality and Specificity of Preregistrations." *PLOS Biology* 18 (12): e3000937.

Brandmaier, Andreas M, Timo von Oertzen, Paolo Ghisletta, Christopher Hertzog, and Ulman Lindenberger. 2015. "LIFESPAN: A tool for the computer-aided design of longitudinal studies." *Frontiers in Psychology* 6: 272.

Cagan, Ross. 2013. "San Francisco Declaration on Research Assessment." *Disease Models & Mechanisms* dmm.012955.

Carnap, Rudolf. 1950. *Logical Foundations of Probability*. Chicago, IL, USA: Chicago University of Chicago Press.

Chan, An-Wen, Asbjørn Hróbjartsson, Mette T. Haahr, Peter C. Gøtzsche, and Douglas G. Altman. 2004. "Empirical Evidence for Selective Reporting of Outcomes in Randomized TrialsComparison of Protocols to Published Articles." *JAMA* 291 (20): 2457–2465.

Christensen, David. 1991. "Clever Bookies and Coherent Beliefs." *The Philosophical Review* 100 (2): 229–247.

Dwan, Kerry, Douglas G. Altman, Juan A. Arnaiz, Jill Bloom, An-Wen Chan, Eugenia Cronin, Evelyne Decullier, et al. 2008. "Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias." *PLOS ONE* 3 (8): e3081.

Fetzer, James H. 1974. "Statistical Explanations." In *PSA 1972: Proceedings of the 1972 Biennial Meeting of the Philosophy of Science Association*, edited by Kenneth F. Schaffner and Robert S. Cohen, Boston Studies in the Philosophy of Science, 337–347. Dordrecht: Springer Netherlands.

Fried, Eiko I. 2020a. "Lack of Theory Building and Testing Impedes Progress in The Factor and Network Literature." *Psychological Inquiry* 31 (4): 271–288.

Fried, Eiko I. 2020b. "Theories and Models: What They Are, What They Are for, and What They Are About." *Psychological Inquiry* 31 (4): 336–344.

Giffin, Adom, and Ariel Caticha. 2007. "Updating Probabilities with Data and Moments." In *AIP Conference Proceedings*, Vol. 954, 74–84.

Hoyningen-Huene, Paul. 2006. "Context of Discovery Versus Context of Justification and Thomas Kuhn." In *Revisiting Discovery and Justification: Historical and Philosophical Perspectives on the Context Distinction*, edited by Jutta Schickore and Friedrich Steinle, Archimedes, 119–131. Dordrecht: Springer Netherlands.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *PLOS Medicine* 2 (8): e124.

Koole, Sander L., and Daniël Lakens. 2012. "Rewarding Replications: A Sure and Simple Way to Improve Psychological Science." *Perspectives on Psychological Science* 7 (6): 608–614.

Kukla, Andre. 1990. "Clinical Versus Statistical Theory Appraisal." *Psychological Inquiry* 1 (2): 160–161.

Lishner, David A. 2015. "A Concise Set of Core Recommendations to Improve the Dependability of Psychological Research." *Review of General Psychology* 19 (1): 52–68.

Meehl, Paul E. 1978. "Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology." *Journal of Consulting and Clinical Psychology* 46 (4): 806–834.

Meehl, Paul E. 1990. "Appraising and Amending Theories: The Strategy of Lakatosian Defense and Two Principles That Warrant It." *Psychological Inquiry* 1 (2): 108–141.

Mellor, David T., and Brian A. Nosek. 2018. "Easy Preregistration Will Benefit Any Research." *Nature Human Behaviour* 2 (2): 98–98.

Niiniluoto, Ilkka. 1998. "Verisimilitude: The Third Period." *The British Journal for the Philosophy of Science* 49 (1): 1–29.

Nosek, Brian A., Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. 2018. "The Preregistration Revolution." *Proceedings of the National Academy of Sciences* 115 (11): 2600–2606.

Oberauer, Klaus. 2019. "Preregistration of a Forking Path – What Does It Add to the Garden of Evidence?" Jan.

Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349 (6251): aac4716.

Orben, Amy, and Daniël Lakens. 2020. "Crud (Re)Defined." *Advances in Methods and Practices in Psychological Science* 3 (2): 238–247.

Peikert, Aaron, Caspar J. van Lissa, and Andreas M. Brandmaier. 2021. "Reproducible Research in R: A Tutorial on How to Do the Same Thing More Than Once." *Psych* 3 (4): 836–867.

Pham, Michel Tuan, and Travis Tae Oh. 2021. "Preregistration Is Neither Sufficient nor Necessary for Good Science." *Journal of Consumer Psychology* 31 (1): 163–176.

Popper, Karl R. 2002. *The Logic of Scientific Discovery*. London; New York: Routledge.

Rubin, Mark. 2020. "Does Preregistration Improve the Credibility of Research Findings?" *The Quantitative Methods for Psychology* 16 (4): 376–390.

Salmon, Wesley C. 1970. "Statistical Explanation." In *The Nature & Function of Scientific Theories: Essays in Contemporary Science and Philosophy*, University of Pittsburgh Series in the Philosophy of Scienc 4, 173–232. Pittsburgh: University of Pittsburgh Press.

Schönbrodt, Felix, Anne Gärtner, Maximilian Frank, Mario Gollwitzer, Malika Ihle, Dorothee Mischkowski, Le Vy Phan, et al. 2022. "Responsible Research Assessment I: Implementing DORA for Hiring and Promotion in Psychology." Nov.

Shmueli, Galit. 2010. "To Explain or to Predict?" *Statistical Science* 25 (3): 289–310.

Silagy, Chris A., Philippa Middleton, and Sally Hopewell. 2002. "Publishing Protocols of Systematic ReviewsComparing What Was Done to What Was Planned." *JAMA* 287 (21): 2831–2834.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2021. "Pre-Registration: Why and How." *Journal of Consumer Psychology* 31 (1): 151–162.

Steegen, Sara, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. 2016. "Increasing Transparency Through a Multiverse Analysis." *Perspectives on Psychological Science* 11

(5): 702–712.

Stefan, Angelika, and Felix Schönbrodt. 2022. "Big Little Lies: A Compendium and Simulation of p-Hacking Strategies." Mar.

Szollosi, Aba, David Kellen, Danielle J. Navarro, Richard Shiffrin, Iris van Rooij, Trisha Van Zandt, and Chris Donkin. 2020. "Is Preregistration Worthwhile?" *Trends in Cognitive Sciences* 24 (2): 94–95.

van Rooij, Iris, and Giosuè Baggio. 2020. "Theory Development Requires an Epistemological Sea Change." *Psychological Inquiry* 31 (4): 321–325.

van Rooij, Iris, and Giosuè Baggio. 2021. "Theory Before the Test: How to Build High-Verisimilitude Explanatory Theories in Psychological Science." *Perspectives on Psychological Science* 16 (4): 682–697.

Wagenmakers, Eric-Jan, Ruud Wetzels, Denny Borsboom, Han L. J. van der Maas, and Rogier A. Kievit. 2012. "An Agenda for Purely Confirmatory Research." *Perspectives on Psychological Science* 7 (6): 632–638.