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

Aaron Peikert[1,2,3], Maximilian S. Ernst[1, 4], and & Andreas M. Brandmaier[1, 3, 5]

[1] Center for Lifespan Psychology
Max Planck Institute for Human Development
Berlin
Germany

[2] Department of Imaging Neuroscience
University College London
London
UK

[3] Max Planck UCL Centre for Computational Psychiatry and Ageing Research
Berlin
Germany

[4] Max Planck School of Cognition
Leipzig
Germany

[5] Department of Psychology
MSB Medical School Berlin
Berlin

Germany

The materials for this article are available on GitHub (Peikert & Brandmaier, 2023a). This version was created from git commit `eace061`. The manuscript is available as preprint (Peikert & Brandmaier, 2023b).

Submitted to *Meta-Psychology*. Participate in open peer review by sending an email to open.peer.reviewer@gmail.com. The full editorial process of all articles under review at Meta-Psychology can be found following this link: https://tinyurl.com/mp-submissions
You will find this preprint by searching for the first author's name.

Author Note

The authors made the following contributions. Aaron Peikert: Conceptualization, Writing—Original Draft Preparation, Writing—Review & Editing, Methodology, Formal analysis, Software, Visualization, Project administration; Maximilian S. Ernst: Writing—Review & Editing, Formal analysis, Validation; Andreas M. Brandmaier: Writing—Review & Editing, Supervisions.

Correspondence concerning this article should be addressed to Aaron Peikert, Center for Lifespan Psychology, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany. E-mail: peikert@mpib-berlin.mpg.de

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. Finally, we discuss the implications of these insights for the practice of preregistration.

Keywords: preregistration; confirmation; exploration; hypothesis testing; Bayesian; Open Science

Word count: 8390

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

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; Nosek et al., 2018; Shmueli, 2010; Tukey, 1980). 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, to make new discoveries and confirmation, to expose these discoveries to potential falsification, and assess empirical support for the theory. However, mistaking exploratory findings for empirically confirmed results is dangerous. It inflates the likelihood of believing that there is evidence supporting a given hypothesis, 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, evermore researchers preregister their hypotheses and their data collection and analysis plans in advance of their studies (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 have 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 and the corresponding analysis 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 (Bakker et al., 2020; Mellor & Nosek, 2018; Nosek et al., 2018; Simmons et al., 2021; Wagenmakers et al., 2012). In such a narrative, preregistration is justified by a confirmatory research agenda. However, two problems become apparent under closer inspection. First, many researchers do not subscribe to a purely confirmatory research agenda (Baumeister, 2016; Brandmaier et al., 2013; Finkel et al., 2017; Tukey, 1972). Second, there is no strict mapping of 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, and so forth. The opposite, that is, preregistered but not strictly confirmatory studies, are also becoming more commonplace (Chan et al., 2004; Dwan et al., 2008; Silagy et al., 2002).

This is the result of researchers applying one of two strategies to evade the self-imposed restrictions of preregistrations: writing a loose preregistration to begin with (Stefan & Schönbrodt, 2023) or deviating from the preregistration afterward (**lakensWhenHowDeviate2024?**). The latter is a frequent occurrence and, perhaps more worryingly, often remains undisclosed (Akker et al., 2023; Claesen et al., 2021). 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 (Pham & Oh, 2021; Szollosi et al., 2020).

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 pose. However reasonable such a move is, changing the core beliefs 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 might well be the case that psychology needs more confirmatory studies to progress as a science. However, independently of such a goal, 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).

We are interested in why preregistrations should change researchers' evaluation of evidence. Applying a Bayesian framework allows us to investigate our research question most straightforwardly because it directly deals with what we ought to believe, given the evidence presented. Specifically, it allows us to model changes in subjective degrees of belief due to preregistration or, more simply, "persuasion". Please note that our decision to adopt a Bayesian philosophy of science does not make assumptions about the statistical methods researchers use. In fact, this conceptualization is intentionally as minimal as possible to be compatible with a wide range of philosophies of science and statistical methods researchers might subscribe to. One feature of the Bayesian framework, is the strong emphasis on subjective yet rational judgement. Therefore, we assume that researchers will differ significantly in how they value evidence but that by making assumptions about the general process, we can make general statements that apply to all these subjective evaluations. However, we should note that Popperians would be appalled that we are content with positive inductive inferences (but we regard "failing to disprove" as too limited), and Neopopperians would flinch that we assign probabilities to beliefs (we are fond of calculating things). While the latter move is not strictly necessary it allows us to connect the more abstract considerations more closely with what researchers believe.

Now, we outline two possible perspectives on the utility of preregistration. The first one 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). We argue that this view on the utility of preregistration can be interpreted as maximizing theoretical risk, which otherwise may be reduced by researchers' degrees of freedom, p-hacking, and suchlike. The second interpretation is our main contribution: We argue that contrary to the classic view, the objective of preregistration is *not* the maximization of theoretical risk but rather the minimization of uncertainty about the theoretical risk. This interpretation leads to a broad applicability of preregistration to both exploratory and confirmatory studies.

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 the amount of knowledge we gain from a particular study, preregistration is potentially useful for any kind of study, no matter where it falls on the exploratory-confirmatory continuum.

Persuation and the Bayesian rationale

If researchers plan to conduct a study, they usually hope that it will change their assessment of some theory's verisimilitude (Niiniluoto, 1998). Moreover, they hope to convince other researchers can be persuaded to change their believe in a theory as well. Beforehand, researchers cannot know what evidence a study will provide but still must form an expectation in order to decide about the specifics of a planned study, including if they should preregister it. If they can expect that preregistration helps them to persuade other researchers to change their believe, it is only rational to employ preregistration. To make our three arguments, we must assume three things about what an ideal estimation process entails 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 persuasiveness 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 by investing resources and stand to gain evidence for or against a theory 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 modeling what liberties researchers have in what studies to conduct. Without deliberating these concepts further, we find it reasonable 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 persuasiveness in terms of posterior probability; that remains a leap of faith. In fact, persuasiveness depends on how other researchers weigh evidence which differs between individuals.

However, the argument applies to any reward function that satisfies the "statistical relevancy condition" (Fetzer, 1974; Salmon, 1970), that is, evidence only increases believe for a theory if the evidence is more likely to be observed under the theory than under the alternative. In particular, "diagnosticity" (Fiedler, 2017; Oberauer & Lewandowsky, 2019), a concept highlighted in recent psychological literature, seems to adhere to the statistical relevancy condition.

Theoretical risk

Our first argument is that theoretical risk is crucial for judging the persuasiveness of evidence. 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 our notation. To keep the notation simple, we restrict ourselves to evidence of a binary nature (either it was observed or not). 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 confirmation of a hypothesis. In the tradition of Carnap, in its direct use, it is called *confirmation 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). We concentrate on the posterior probability because of its simplicity but take it only as one example of a possible measure. In reality, researchers surely differ in what function they apply to judge evidence and it is often most fruitful to compare more than two competing hypotheses. The goal is therefore to reason about the space of possible measures researchers might apply. However, since any measure fulfilling the statistical relevancy condition increases monotonically with an increase in posterior probability $P(H|E)$, we might well take it to illustrate our reasoning.

In short, we want to increase posterior probability $P(H|E)$. Increases in posterior probability $P(H|E)$ are associated with increases in persuasiveness, of which we want to maximize the expectation. So how can we increase posterior probability? The Bayes formula yields three components that influence confirmation, 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 its favor will result in a hypothesis that is more strongly confirmed, all else being equal. However, the prior probability of a hypothesis is nothing our study design can change. The second option is equally reasonable; that is, an increase in $P(E|H)$ leads to a higher posterior probability $P(H|E)$. $P(E|H)$ is the probability of obtaining evidence for a hypothesis when it holds. We call this probability of detecting evidence, given that the hypothesis 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 complement of type-II error rate). However, while detectability is of great importance for study design, it is not directly relevant to what a preregistration is communicating to other researchers. We later discuss how issues of detectability must be considered in a preregistration. Thus, $P(E)$ remains to be considered. Since $P(E)$ is the denominator, decreasing it can increase the posterior probability. In other words, high risk, high reward.

If we equate riskiness with a low probability of obtaining evidence (when the hypothesis is false), the Bayesian rationale perfectly aligns with the observation that risky predictions lead to persuasive evidence. This tension between high risk leading to high gain is central to our consideration of preregistration. A high-risk, high-gain 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 are incentivized 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 the factors contributing 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 have already noted that there is not much to be done 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 favor. 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 a purely Bayesian framework. However, while he did not mention $P(\neg E|\neg H)$ he suggested 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 (the inverse of statistical relevancy). 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 rationale we apply here is able to reconstruct many commonly held views on how "risky" predictions are valued (but we of course differ from Popper on the central role of induction in science).

Both theoretical risk $P(\neg E|\neg H)$ and detectability $P(E|H)$ aggregate countless influences; otherwise, they could not model the process of evidential support for theories. To illustrate the concepts we have introduced 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 "VIII" 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 can 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 will 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 mentioned earlier, the we restrict ourselves to binary 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.

Preregistration as a means to increase theoretical risk?

Having discussed that increasing the theoretical risk will increase the persuasiveness, 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 (Orben & Lakens, 2020). However, we would consider it a success if the researcher would use the evidence from this rather exploratory study to develop a more precise (and therefore risky) theory, e.g., by using the results to specify which variables from one set relate to which variables from the other set, to what extent, in which direction, with which functional shape, etc., to be able to make riskier predictions in the future. 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 the theory again. This is because preregistration increases the expected persuasiveness regardless of the theory being tested, as we will show.

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. This is similar to the well known tradeoff between type-I error rate and statistical power. 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. As noted before, we strive for high detectability and high theoretical risk in planning, conducting, and analyzing studies. Maximizing one at the expense of the other is not necessarily beneficial for increasing persuasiveness but depends on the specific function they apply to judge evidence and their specific location on the curve. One advantage of our framework is that researchers can employ it to balance the trade-off more effectively assuming they are willing to make some simplifiying assumptions.

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 about theoretical risk into our framework.

Statistical methods

One widely known factor is the contribution of statistical methods to theoretical risk. Theoretical risk $P(\neg E|\neg H)$ is deeply connected with statistical methods, because it is related to the type-I error rate in statistical hypothesis testing $P(E|\neg H)$ by $P(\neg E|\neg H) = 1 - P(E|\neg H)$, 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, this connection should not be overinterpreted for two reasons. First, according to most interpretations of null hypothesis testing, the absence of a significant result should not generally be interpreted as evidence against the hypothesis (Mayo, 2018, p. 5.3). Second, the research hypothesis rarely equals the statistical alternative hypothesis (most research hypothesis are more specific than “any value except zero”). In fact, it is entirely possible to assume the null hypothesis as a research hypothesis, as is commonly done in e.g., structural equation modelling, where the roles of detectability, theoretical risk and type-I/II error rate switch. We argue that theoretical risk (and hence its complement, $P(E|\neg H)$) also encompasses factors outside the statistical realm, most notably the study design and broader analytical strategies. Type-I error rate is the property of a statistical test under some assumptions, whereas theoretical risk is a researchers’ belief. One may take such theoretical properties as a first starting point to form a substantive belief but surely researchers ought to take other factors into consideration. For example, if a researcher believes that there might be confounding variables at play for the relation between two variables, this should decrease theoretical risk; after all they might find an association purely on account of the confounders (Fiedler, 2017).

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 influence theoretical risk (and hence, $P(E|\neg H)$.

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.

Sources of uncertainty

As we have 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 a 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 may 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 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 will 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 ascertained. Therefore, we 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.

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. Let us assume a researcher is reading a study from another lab and tries to decide whether and how much the presented results confirm the hypothesis. As the researcher did not conduct the study (and the study is not preregistered), they can not be certain about the various factors influencing theoretical risk (researcher degrees of freedom). We therefore express this uncertainty about the theoretical risk as a probability distribution Q of $P(E|\neg H)$ (remember that $P(E|\neg H)$ is related to theoretical risk by $P(E|\neg H) = 1 - P(\neg E|\neg H)$, so it does not matter whether we consider the distribution of theoretical risk or $P(E|\neg H)$). To get the expected value of $P(H|E)$ that follows from the researchers' uncertainty about the theoretical risk, we can compute the expectation using Bayes theorem:

$$\mathbb{E}_Q[P(H|E)] = \mathbb{E}_Q\left[\frac{P(H)P(E|H)}{P(H)P(E|H) + P(\neg H)P(E|\neg H)}\right] \tag{3}$$

Of course, the assigned probabilities and the distribution Q vary from study to study and researcher to researcher (and even the measure of confirmation), but we can illustrate the effect of uncertainty with an example. Assuming $P(E|H) = 0.8$ (relective of the typically strived for power of 80%). Let us further assume that the tested hypothesis is considered unlikely to be true by the research community before the study is conducted ($P(H) = 0.1$) and assign a uniform distribution for $P(E|\neg H) \sim U([1-\tau, 1])$ where τ is set to $1-\alpha$, reflecting our assumption that this term gives an upper bound for theoretical risk $P(\neg E|\neg H)$. We chose this uniform distribution as it is the maximum entropy distribution with support $[1-\tau, 1]$ and hence conforms to our Bayesian framework (Giffin & Caticha, 2007).

With this, we derive the expected value of $P(H|E)$ as

$$\mathbb{E}_Q[P(H|E)] = \mathbb{E}_Q\left[\frac{P(H)P(E|H)}{P(H)P(E|H) + P(\neg H)P(E|\neg H)}\right] \tag{4}$$

$$= \int_{[1-\tau,1]} \tau^{-1} \frac{P(H)P(E|H)}{P(H)P(E|H) + P(\neg H)P(E|\neg H)} \, \mathrm{d}P(E|\neg H) \tag{5}$$

$$= \frac{P(H)P(E|H)}{P(\neg H)\tau} \ln\left(\frac{P(H)P(E|H) + P(\neg H)}{P(H)P(E|H) + P(\neg H)(1-\tau)}\right) \tag{6}$$

Figure 1 shows exemplary the effect of theoretical risk (x-axis) on the posterior probability (y-axis) being certain (solid line) or uncertain (dashed line) about the theoretical risk of a study. Our expectation of the persuasiveness varies considerably depending on how uncertain we are about the theoretical risk a study took on. Mathematically, uncertainty about theoretical risk is expressed through the variance (or rather entropy) of the distribution. The increase in uncertainty (expressed as more entropic distributions) leads to a decreased expected persuasiveness.

The argument for a confirmatory research agenda is that by increasing theoretical risk we increase expected persuasiveness, i.e., moving to the right on the x-axis in Figure 1

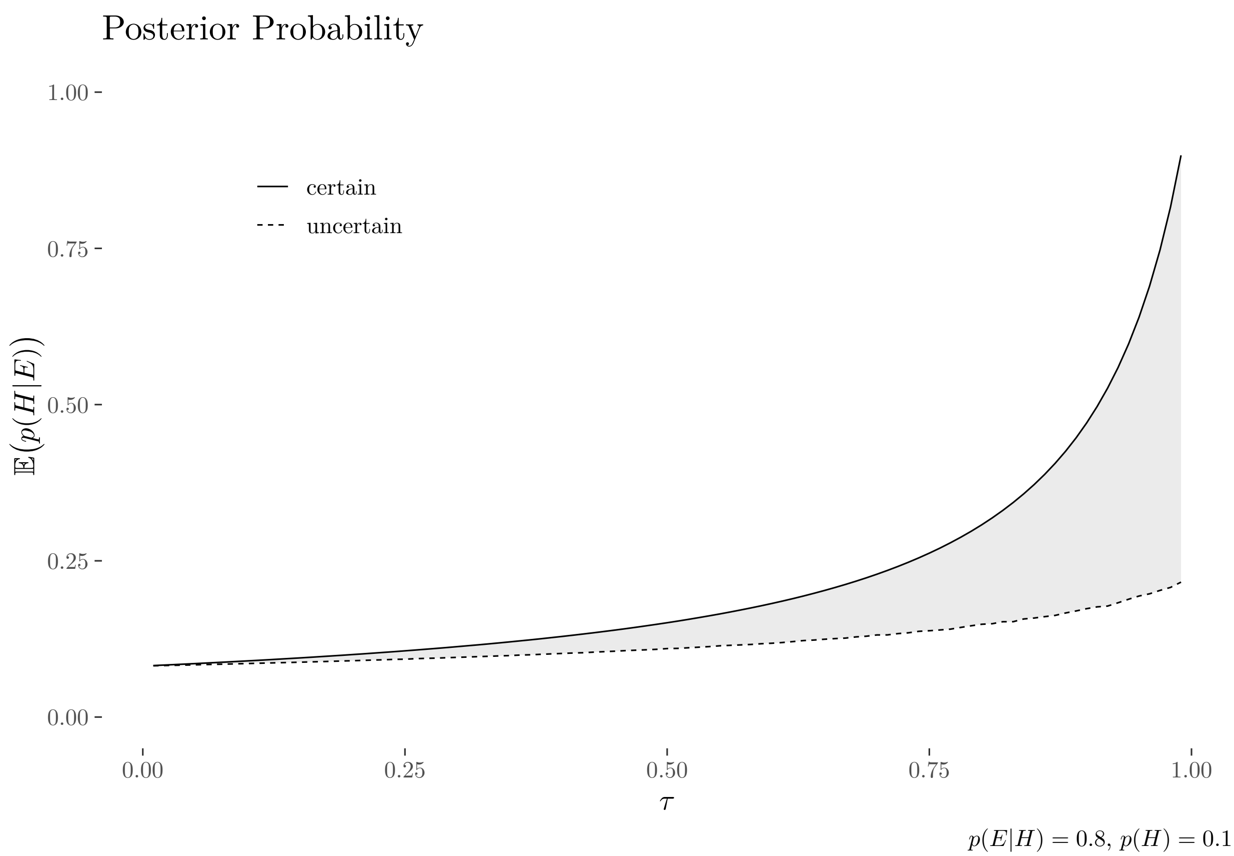

Figure 1

Posterior probability (confirmation as firmness) as a function of theoretical risk τ, where τ is either certain (solid line) or maximally uncertain (dotted line).

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 by how certain the recipient is about the theoretical risk. A study that has a very high theoretical risk (e.g., 1.00% chance that if the hypothesis is wrong, evidence in its favor will be observed,) 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 confirmation) and, therefore, the extent of the benefit varies, 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 $\tau = 0.80$ (false

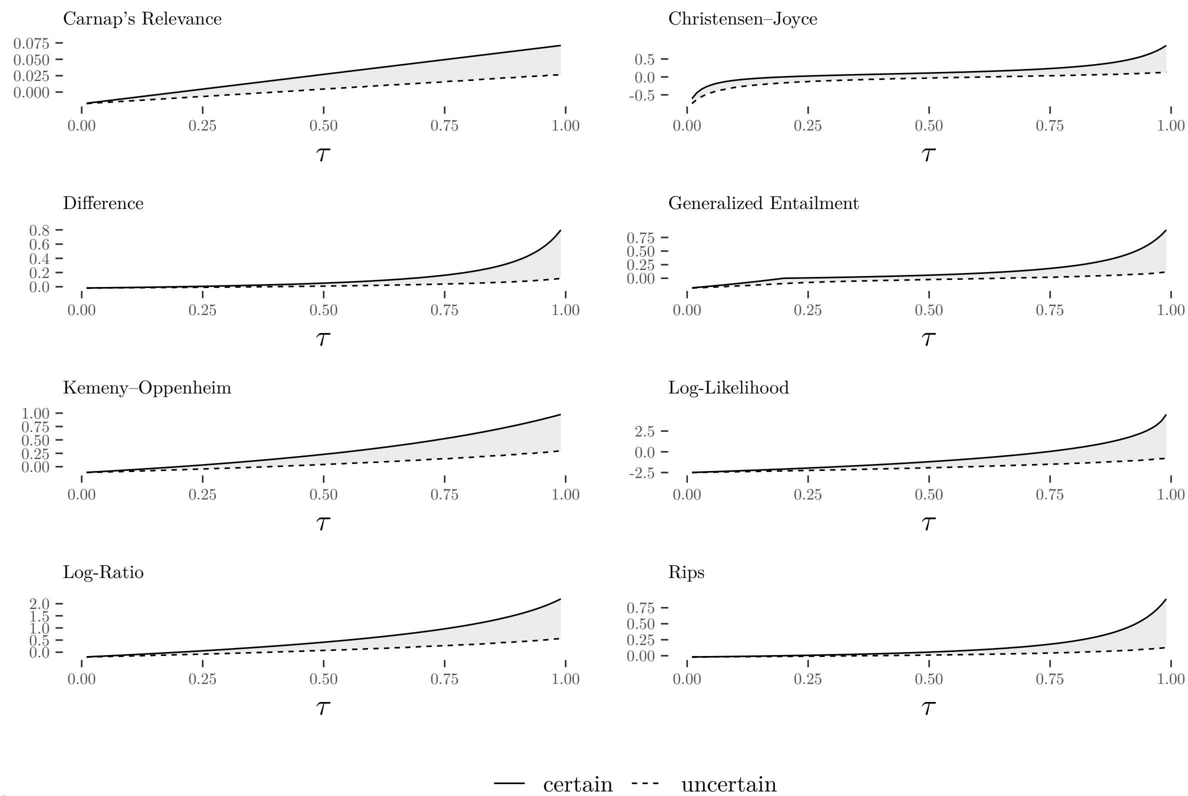

Figure 2

Several measures for confirmation as an increase in firmness as a function of τ, where τ is either certain (solid line) or maximally uncertain (dotted line). Measures taken from Sprenger and Hartmann (2019), Table 1.3, p. 51.

positive rate of 0.20) moving from uncertain to certain increases the posterior from 0.15 to 0.31. We find it helpful to calculate an example because of the nonlinear nature of the evidence functions.

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 a similar rational process.
3. Researchers try to maximize the expected persuasiveness 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 the whole scientific community the study authors hope to persuade. Unfortunately, readers of a scientific article (or, more generally, any consumer of a research product) will likely 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 skeptic, they may claim maximum uncertainty.

Communicating clearly how authors of a scientific report collected 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 for communicating just that because any description after the fact is prone 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, readers can be certain they 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 extent than if the study would not have been preregistered.

The 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. To further reduce the uncertainty about theoretical risk, researchers must therefore publish code and ideally data. After all, computational reproducibility is only possible if the data analytic procedure was communicated clearly enough to allow others to retrace the computational steps (Peikert & Brandmaier, 2021).

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.

Hacking, harking, and other harms

The importance of distinguishing between low and highly uncertain theoretical risk becomes perhaps clearer if we consider a few hypothetical cases for illustration.

1. We know with absolute certainty that researchers will revert to p-hacking to create evidence that is favorable for the theory.
2. A hypothesis was picked to explain reported results after the fact (HARKing, Kerr, 1998).
3. We cannot exclude the possibility of p-hacking having led to the reported results.
4. Reported results were obtained by planned exploration.

5. Reported results were obtained by unplanned exploration.

In case 1, there is no theoretical risk ($P(\neg E|\neg H) = 0$). If we know that the results will be engineered to support the hypothesis no matter what, there is no reason to collect data. A prime example of this case is the $p_{\text{pointless}}$ metric (Hussey, 2021). Case 2 has a similar problem. After all, the hypothesis that it had to happen the way it did happen is irrefutable. In fact, both cases should be problematic to anyone who subscribes to the statistical relevancy condition because if we choose the hypothesis in accordance with the data or vice versa, without restrictions, they are not related anymore (i.e., observing the data does not tell us anything about the hypothesis and the other way around). Case 3 is different since here the theoretical risk is not necessarily low but simply uncertain (and perhaps best represented by the dotted line in Figure 1/2). In case 4, the theoretical risk is neither zero (unless the researcher plans to do run variations of analyses until a favourable outcome is obtained, then we have a particular instance case of 1) nor high (as this is the nature of exploratory approaches). However, we can take advantage of computational reproducibility, use statistical properties, simulation or resampling methods, together with scientific reasoning, to get a reasonably certain evaluation of the theoretical risk. Low uncertainty about high theoretical risk is a somewhat favourable position (i.e., close to the solid line in Figure 1/2). This favorable position leads us to recommend preregistration of exploratory studies. Case 5 shares the neither zero nor high theoretical risk of case 4 but has additional uncertainty about how much exploration was going on (how hard exactly did the researchers try to come up with favourable results). Its low *and uncertain* theoretical risk make it difficult to produce compelling evidence.

Discussion

To summarize, we showed that both higher theoretical risk and lower uncertainty about theoretical risk lead to higher persuasiveness across a variety of measures. The former result that increasing theoretical risk leads to higher expected persuasiveness reconstructs the appeal and central goal of preregistration of confirmatory research agendas. However, theoretical risk is something researchers 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, 2020b; van Rooij & Baggio, 2021, 2020). 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 & Lakens, 2012).

Theoretical risk may conceptually be related to the framework of "severity" (Mayo, 2018; Mayo & Spanos, 2011). Severity, is a Neopopperian view which asserts that there is evidence for a hypothesis just to the extent that it survives stringent scrutiny. However, there are crucial differences between the two. First, our perspective on theoretical risk is not primarily concerned with avoiding inductive reasoning but with subjective changes of belief. This is important because, while severity is calculable, it remains unclear how severity should be valued, e.g. if an increase in severity from .80 to .81 should be as impressive as from .99 to .999. Second, severity considerations are mainly after the fact. Severity, a measure with which we can rule out alternative explanations, can only be calculated after evidence was observed. This makes it difficult to guide a priori decisions in planning a study, after all severity disregards power, if we observe evidence, and disregards Type I error rate when we do not. This implies that for a priori balancing Type I and Type II error rate, a researcher must assign a priori probabilities to, for example, the size of an effect. Since such a move not in line with frequentist rationale there is no guidelines available on how to do this. Third, we would argue that severity considerations assume full information about how the evidence came about and hence imply axiomatically the need for perfect preregistration. This comes down to frequentist understanding of probability as the outcome of a well defined random experiments. When judging a particular study, a frequentist, and hence a severe tester, may not assign probability to the event that the researchers did, for example, p-hack. The lack of knowledge on the readers side does not turn the p hacking into a random event of which we can calculate the long run frequency aka frequentist probability. A severe test, hence, must assume that they know the Type I and Type II error rate precisely. Full transparency, is hence assumed, and we can not imagine many ways except preregistration that get close to this ideal. This assumptions also makes it difficult to deal with less than perfect preregistrations and post-hoc changes without appealing to principles outside the core philosophy of severity.

However, there also are communalities, like the strong emphasis on counterfactual consideration (imagining the hypothesis was false), and there are even proposals to reconcile Bayesian and severity considerations (van Dongen et al., 2023).

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 persuasiveness by virtue of reducing uncertainty about theoretical risk. Nevertheless, exploratory studies will have a lower expected persuasiveness 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 rationale we propose here, but is less reasonable when the goal of preregistration is supposed to be a strictly confirmatory research agenda.

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 theoretical 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 possible under our proposed 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 persuasiveness 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 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). In any case, deviations must be transparently reported and we applaud recent developments to standardize and normalize this process (Willroth & Atherton, 2023).

As explained above, reduction in uncertainty as the objective for preregistration does not only explain some existing practice, 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 a 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 preregistration should reduce. However, specifying procedures that can be expected to fail will lead to deviation and, subsequently, to larger uncertainty.

Our emphasis on transparency aligns with other justifications of preregistration, especially those put forth by Lakens (2019)'s, although based on quite different philosophical foundations. Our goal is to contribute a rationale that more comprehensively captures the spectrum of exploration and confirmation in relation to preregistrations, post-hoc changes of preregistrations, and subjective evaluations of evidence. We find it difficult to content ourselves with vague terms like "control" or "transparency" if they ultimately remain unconnected to how much researchers believe in a theory. Within our framework, researchers have the ability to input their assumptions regarding the perspectives of other researchers and calculate the potential impact of their actions on their readership, whether these actions relate to study design, to the preregistration itself, or subsequent deviations from it. We put subjective evaluations at the center of our considerations; we deal explicitly with researchers who are proponents of some theory (they have higher priors for the theory being true), researchers who suspect confounding variables (they assume lower theoretical risk), or those who remain doubtful if everything relevant was reported (they have higher uncertainty about theoretical risk) or even those who place greater value on incongruent evidence than others (they differ in their confirmation function). We, therefore, hope to not only provide a rationale for preregistration for those who subscribe to a Bayesian philosophy of science but also a framework to navigate the complicated questions that arise in the practice of preregistration.

At the same time, approaching the evaluation of evidence using a Bayesian formalism is far from novel Fiedler (2017). To our knowledge, it was not yet applied to the problem of preregistration. However, Oberauer and Lewandowsky (2019) made use of the formalism to model the relation between theory, hypothesis, and evidence. In the context of this conceptualization, they discussed the usefulness of preregistration, though without applying the formalism there. Most importantly, they are rather critical of the idea that preregistration has tangible benefits. Instead, they prefer multiverse analyses but contend that those could be preregistered if one fancies it. Their reasoning is based on two intuitions about what should *not* influence the evaluation of evidence: temporal order and the mental state of the originator. In our opinion, they disregard the temporal order a bit too hastily, as it is a long-standing issue in Bayesian philosophy of science known as the "problem of old evidence" (Chihara, 1987). However, we agree that not the temporal order is decisive but if the researchers incorporated the information into the hypothesis the evidence is supposed to confirm. For the other, we argue that the mental state of the originator does matter. Suppose there are $k = 1, 2, \dots, K$ ways to analyze data, where each k has a $P(E_k|\neg H) > 0$. If they intend to try each way after another but happen to be "lucky" on the first try and stop, should we then apply $P(E|\neg H) = P(E_1|\neg H)$ or $P(E|\neg H) = P(E_1 \vee \dots \vee E_k|\neg H)$? We think the latter. However, this "Defeatist" intuition is not universally warranted and depends on what we take H to mean specifically (Kotzen, 2013). Addressing, this problem might benefit from combining Oberauer and Lewandowsky (2019)'s idea of updating on two nested levels (theory-hypothesis layered on top of hypothesis-evidence) with our approach to modelling uncertainty.

Whatever the difference in evaluating preregistration as a tool, maybe conceptually more profound is that Oberauer and Lewandowsky (2019) conceptualizes "discovery-oriented research" differently than we do "exploratory". They assume the same theoretical risk ($P(\neg E|\neg H) = .05$) and detectability ($P(E|H)) = .8$) in their calculation example as we do but assign different prior probabilities, namely .06 for discovery versus .6 for theory testing. Then, they conclude that discovery-oriented researcher requires a much lower type-I error rate to control false positive in light of the low prior probability. This runs counter to our definition of exploratory research having low theoretical risk. Of course, we agree that low priors require more persuasive evidence; our disagreement, therefore, lies mainly in terminology. They imagine discovery-oriented researchers to conduct experiments where they have low expectations that they obtain positive evidence ($.06 \cdot .8 + .94 \cdot .05 = 0.095$), but if they do, it raises the posterior significantly (from .06 to .51) In our view, researchers who set out to explore a data set often find "something" (due to low $P(\neg E|\neg H)$); therefore, it should only slightly raise your posterior if they do. On a substantive matter, we believe both kinds of research are common in psychology. It is, therefore, mostly a disagreement on terminology. This disagreement only highlights why using a mathematical framework to investigate such things is so useful and ultimately indispensable because we can clearly see where and how we differ in our reasoning.

We believe that our reasoning is quite similar to Höfler et al. (2022), who call for transparent exploration using preregistration. We could be more sure of our agreement, if they had formulated their arguments within a mathematical framework, which would also have helped to dissolve an apparent conflict in their definitions of confirmation, exploration, and transparency. On the one hand, they define "The principle difference between confirmation and exploration is that confirmation adheres to an evidential norm for the test of a hypothesis to pass.", but then suggest that transparent exploration can be conducted using inferences tests as a filtering mechanism. Their distinction between confirmation, intransparent and transparent exploration are otherwise just as well placed along the dimensions, theoretical risk and uncertainty about theoretical risk.

With the goal to facilitate rigorous exploration, we have proposed a workflow for preregistration called *preregistration as code* (PAC) elsewhere (Peikert et al., 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 an approach dissolves the distinction between the preregistration document and the final scientific report. Instead of 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 does not limit the extent of exploration, it is very specific about the probability 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 persuasiveness.

Acknowledgement

We thank Leo Richter, 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. We are grateful to Julia Delius for her helpful assistance in language and style editing.

Declarations

All code and materials required to reproduce this article are available under https://github.com/aaronpeikert/bayes-prereg (Peikert & Brandmaier, 2023a). The authors have no competing interests to declare that are relevant to the content of this article.

References

Akker, O. van den, Bakker, M., Assen, M. A. L. M. van, Pennington, C. R., Verweij, L., Elsherif, M., Claesen, A., Gaillard, S. D. M., Yeung, S. K., Frankenberger, J.-L., Krautter, K., Cockcroft, J. P., Kreuer, K. S., Evans, T. R., Heppel, F., Schoch, S. F., Korbmacher, M., Yamada, Y., Albayrak-Aydemir, N., … Wicherts, J. (2023, May 10). *The effectiveness of preregistration in psychology: Assessing preregistration strictness and preregistration-study consistency.* https://doi.org/10.31222/osf.io/h8xjw

Bakker, M., Veldkamp, C. L. S., Assen, M. A. L. M. van, Crompvoets, E. A. V., Ong, H. H., Nosek, B. A., Soderberg, C. K., Mellor, D., & Wicherts, J. M. (2020). Ensuring the quality and specificity of preregistrations. *PLOS Biology*, *18*(12), e3000937. https://doi.org/10.1371/journal.pbio.3000937

Baumeister, R. F. (2016). Charting the future of social psychology on stormy seas: Winners, losers, and recommendations. *Journal of Experimental Social Psychology*, *66*, 153–158. https://doi.org/10.1016/j.jesp.2016.02.003

Brandmaier, A. M., Oertzen, T. von, Ghisletta, P., Hertzog, C., & Lindenberger, U. (2015). LIFESPAN: A tool for the computer-aided design of longitudinal studies. *Frontiers in Psychology*, *6*, 272.

Brandmaier, A. M., von Oertzen, T., McArdle, J. J., & Lindenberger, U. (2013). Structural equation model trees. *Psychological Methods*, *18*(1), 71–86. https://doi.org/10.1037/a0030001

Cagan, R. (2013). San Francisco Declaration on Research Assessment. *Disease Models & Mechanisms*, dmm.012955. https://doi.org/10.1242/dmm.012955

Carnap, R. (1950). *Logical Foundations of Probability.* Chicago, IL, USA: Chicago University of Chicago Press.

Chan, A.-W., Hróbjartsson, A., Haahr, M. T., Gøtzsche, P. C., & Altman, D. G. (2004). Empirical Evidence for Selective Reporting of Outcomes in Randomized TrialsComparison of Protocols to Published Articles. *JAMA*, *291*(20), 2457–2465.

https://doi.org/10.1001/jama.291.20.2457

Chihara, C. S. (1987). Some Problems for Bayesian Confirmation Theory. *The British Journal for the Philosophy of Science*, *38*(4), 551–560. https://doi.org/10.1093/bjps/38.4.551

Christensen, D. (1991). Clever Bookies and Coherent Beliefs. *The Philosophical Review*, *100*(2), 229–247. https://doi.org/10.2307/2185301

Claesen, A., Gomes, S., Tuerlinckx, F., & Vanpaemel, W. (2021). Comparing dream to reality: An assessment of adherence of the first generation of preregistered studies. *Royal Society Open Science*, *8*(10), 211037. https://doi.org/10.1098/rsos.211037

Dwan, K., Altman, D. G., Arnaiz, J. A., Bloom, J., Chan, A.-W., Cronin, E., Decullier, E., Easterbrook, P. J., Elm, E. V., Gamble, C., Ghersi, D., Ioannidis, J. P. A., Simes, J., & Williamson, P. R. (2008). Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias. *PLOS ONE*, *3*(8), e3081. https://doi.org/10.1371/journal.pone.0003081

Fetzer, J. H. (1974). Statistical Explanations. In K. F. Schaffner & R. S. Cohen (Eds.), *PSA 1972: Proceedings of the 1972 Biennial Meeting of the Philosophy of Science Association* (pp. 337–347). Springer Netherlands. https://doi.org/10.1007/978-94-010-2140-1_23

Fiedler, K. (2017). What Constitutes Strong Psychological Science? The (Neglected) Role of Diagnosticity and A Priori Theorizing. *Perspectives on Psychological Science*, *12*(1), 46–61. https://doi.org/10.1177/1745691616654458

Finkel, E. J., Eastwick, P. W., & Reis, H. T. (2017). Replicability and other features of a high-quality science: Toward a balanced and empirical approach. *Journal of Personality and Social Psychology*, *113*(2), 244–253. https://doi.org/10.1037/pspi0000075

Fried, E. I. (2020a). Lack of Theory Building and Testing Impedes Progress in The Factor and Network Literature. *Psychological Inquiry*, *31*(4), 271–288. https://doi.org/10.1080/1047840X.2020.1853461

Fried, E. I. (2020b). Theories and Models: What They Are, What They Are for, and What They Are About. *Psychological Inquiry*, *31*(4), 336–344. https://doi.org/10.1080/1047840X.2020.1854011

Giffin, A., & Caticha, A. (2007). Updating Probabilities with Data and Moments. *AIP Conference Proceedings*, *954*, 74–84. https://doi.org/10.1063/1.2821302

Höfler, M., Scherbaum, S., Kanske, P., McDonald, B., & Miller, R. (2022). Means to valuable exploration: I. The blending of confirmation and exploration and how to resolve it. *Meta-Psychology*, *6*. https://doi.org/10.15626/MP.2021.2837

Hoyningen-Huene, P. (2006). Context of Discovery Versus Context of Justification and Thomas Kuhn. In J. Schickore & F. Steinle (Eds.), *Revisiting Discovery and Justification: Historical and philosophical perspectives on the context distinction* (pp. 119–131). Springer Netherlands. https://doi.org/10.1007/1-4020-4251-5_8

Hussey, I. (2021). A method to streamline p-hacking. *Meta-Psychology*, *5*. https://doi.org/10.15626/MP.2020.2529

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLOS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Koole, S. L., & Lakens, D. (2012). Rewarding Replications: A Sure and Simple Way to Improve Psychological Science. *Perspectives on Psychological Science*, *7*(6), 608–614. https://doi.org/10.1177/1745691612462586

Kotzen, M. (2013). Multiple Studies and Evidential Defeat. *Noûs*, *47*(1), 154–180. http://www.jstor.org/stable/43828821

Kukla, A. (1990). Clinical Versus Statistical Theory Appraisal. *Psychological Inquiry*, *1*(2), 160–161. https://doi.org/10.1207/s15327965pli0102_9

Lakens, D. (2019). The value of preregistration for psychological science: A conceptual analysis. , *62*(3), 221–230. https://doi.org/10.24602/sjpr.62.3_221

Lishner, D. A. (2015). A Concise Set of Core Recommendations to Improve the Dependability of Psychological Research. *Review of General Psychology*, *19*(1), 52–68. https://doi.org/10.1037/gpr0000028

Mayo, D. G. (2018). *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars* (First). Cambridge University Press. https://doi.org/10.1017/9781107286184

Mayo, D. G., & Spanos, A. (2011). Error Statistics. In *Philosophy of Statistics* (pp. 153–198). Elsevier. https://doi.org/10.1016/B978-0-444-51862-0.50005-8

Meehl, P. E. (1990). Appraising and Amending Theories: The Strategy of Lakatosian Defense and Two Principles that Warrant It. *Psychological Inquiry*, *1*(2), 108–141. https://doi.org/10.1207/s15327965pli0102_1

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4), 806–834. https://doi.org/10.1037/0022-006X.46.4.806

Mellor, D. T., & Nosek, B. A. (2018). Easy preregistration will benefit any research. *Nature Human Behaviour*, *2*(2), 98–98. https://doi.org/10.1038/s41562-018-0294-7

Niiniluoto, I. (1998). Verisimilitude: The Third Period. *The British Journal for the Philosophy of Science*, *49*(1), 1–29. https://doi.org/10.1093/bjps/49.1.1

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Oberauer, K. (2019). Preregistration of a forking path – What does it add to the garden of evidence? In *Psychonomic Society Featured Content*.

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, *26*(5), 1596–1618. https://doi.org/10.3758/s13423-019-01645-2

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Orben, A., & Lakens, D. (2020). Crud (Re)Defined. *Advances in Methods and Practices in Psychological Science*, *3*(2), 238–247. https://doi.org/10.1177/2515245920917961

Peikert, A., & Brandmaier, A. M. (2023a). *Supplemental materials for preprint: Why does preregistration increase the persuasiveness of evidence? A Bayesian rationalization*. Zenodo. https://doi.org/10.5281/zenodo.7648471

Peikert, A., & Brandmaier, A. M. (2023b). *Why does preregistration increase the persuasiveness of evidence? A Bayesian rationalization*. PsyArXiv; PsyArXiv. https://doi.org/10.31234/osf.io/cs8wb

Peikert, A., & Brandmaier, A. M. (2021). A Reproducible Data Analysis Workflow With R Markdown, Git, Make, and Docker. *Quantitative and Computational Methods in Behavioral Sciences*, 1–27. https://doi.org/10.5964/qcmb.3763

Peikert, A., van Lissa, C. J., & Brandmaier, A. M. (2021). Reproducible Research in R: A Tutorial on How to Do the Same Thing More Than Once. *Psych*, *3*(4), 836–867. https://doi.org/10.3390/psych3040053

Pham, M. T., & Oh, T. T. (2021). Preregistration Is Neither Sufficient nor Necessary for Good Science. *Journal of Consumer Psychology*, *31*(1), 163–176. https://doi.org/10.1002/jcpy.1209

Popper, K. R. (2002). *The logic of scientific discovery*. Routledge.

Rubin, M. (2020). Does preregistration improve the credibility of research findings? *The Quantitative Methods for Psychology*, *16*(4), 376–390. https://doi.org/10.20982/tqmp.16.4.p376

Salmon, W. C. (1970). Statistical Explanation. In *The Nature & function of scientific theories: Essays in contemporary science and philosophy* (pp. 173–232). University of Pittsburgh Press.

Schönbrodt, F., Gärtner, A., Frank, M., Gollwitzer, M., Ihle, M., Mischkowski, D., Phan, L.

V., Schmitt, M., Scheel, A. M., Schubert, A.-L., Steinberg, U., & Leising, D. (2022). *Responsible Research Assessment I: Implementing DORA for hiring and promotion in psychology*. PsyArXiv. https://doi.org/10.31234/osf.io/rgh5b

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, *25*(3), 289–310. https://doi.org/10.1214/10-STS330

Silagy, C. A., Middleton, P., & Hopewell, S. (2002). Publishing Protocols of Systematic Reviews Comparing What Was Done to What Was Planned. *JAMA*, *287*(21), 2831–2834. https://doi.org/10.1001/jama.287.21.2831

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2021). Pre-registration: Why and How. *Journal of Consumer Psychology*, *31*(1), 151–162. https://doi.org/10.1002/jcpy.1208

Sprenger, J., & Hartmann, S. (2019). *Bayesian Philosophy of Science*. Oxford University Press. https://doi.org/10.1093/oso/9780199672110.001.0001

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Stefan, A. M., & Schönbrodt, F. D. (2023). Big little lies: A compendium and simulation of p-hacking strategies. *Royal Society Open Science*, *10*(2). https://doi.org/10.1098/rsos.220346

Szollosi, A., Kellen, D., Navarro, D. J., Shiffrin, R., Rooij, I. van, Zandt, T. V., & Donkin, C. (2020). Is Preregistration Worthwhile? *Trends in Cognitive Sciences*, *24*(2), 94–95. https://doi.org/10.1016/j.tics.2019.11.009

Tukey, J. W. (1972). Exploratory data analysis: As part of a larger whole. *Proceedings of the 18th Conference on Design of Experiments in Army Research and Development and Training*, 1–18. https://apps.dtic.mil/sti/tr/pdf/AD0776910.pdf

Tukey, J. W. (1980). We Need Both Exploratory and Confirmatory. *The American Statistician*, *34*(1), 23–25. https://doi.org/10.2307/2682991

van Dongen, N., Sprenger, J., & Wagenmakers, E.-J. (2023). A Bayesian perspective on severity: Risky predictions and specific hypotheses. *Psychonomic Bulletin & Review*, *30*(2), 516–533. https://doi.org/10.3758/s13423-022-02069-1

van Rooij, I., & Baggio, G. (2021). Theory Before the Test: How to Build High-Verisimilitude Explanatory Theories in Psychological Science. *Perspectives on Psychological Science*, *16*(4), 682–697. https://doi.org/10.1177/1745691620970604

van Rooij, I., & Baggio, G. (2020). Theory Development Requires an Epistemological Sea Change. *Psychological Inquiry*, *31*(4), 321–325. https://doi.org/10.1080/1047840X.2020.1853477

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*(6), 632–638. https://doi.org/10.1177/1745691612463078

Willroth, E. C., & Atherton, O. E. (2023). *Best Laid Plans: A Guide to Reporting Preregistration Deviations* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/dwx69