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

Abstract

The replication crisis spurned many researchers to preregister their analyses before acquiring data. However, there is no agreement on what preregistration should accomplish and why it is uniquely suited to these goals. A widespread view is that preregistration should limit how much the data may influence the hypotheses tested on the same data (i.e., restrict researchers' degrees of freedom, alpha error, and theoretical risk). If no such influence occurs, an analysis is generally considered confirmatory. Consequently, many researchers believe that only confirmatory studies benefit from preregistration. Hence, they struggle to preregister their research, as many study designs and theories require adapting to the data. We show that limiting preregistration to confirmatory research is unnecessarily restrictive. To that end, we formalize the objective of preregistration and demonstrate that exploratory studies also benefit from the practice. Drawing on Bayesian philosophy of science, we argue that preregistration should aim to reduce uncertainty about the inferential procedure used to derive the results. Crucially, this objective separates preregistration from the goal of confirmatory research and provides a principled justification in its own right. While the extent to which a study is exploratory is important, certainty about the inferential procedure is a precondition for persuasive evidence. Lastly, we discuss what implications these insights have for the practice of preregistration.

The scientific community has long pondered the vital distinction between exploration and confirmation, discovery and justification, hypothesis-generating and hypothesis-testing, or prediction and postdiction Nosek et al. (2018). There is a broad consensus that both 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 it has been corroborated. Mistaking exploratory findings for empirically confirmed results inflates the the likelihood of believing that there is evidence for a given explanation to be true if it is false. A variety of practices, 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 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 and to produce results that were clearly and transparently to be labeled 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 promised false-positive rates. This ability to provide 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 leave data-dependent decisions unspecified.

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 narrative, preregistration is justified by a confirmatory research agenda. However, two problems emerge 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, etc. The opposite would be preregistered but not strictly confirmatory studies and those are possible as well (Chan et al., 2004; e.g., Dwan et al., 2008; Silagy et al., 2002). Researchers may apply two strategies to evade the self-imposed restrictions of preregistrations: writing a loose preregistration to begin with (Stefan & 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 (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 researcher 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 a confirmatory research agenda does not only imply changing *how* they investigate but also *what* research questions. 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. While preregistration is especially useful for dividing confirmatory and exploratory research, we argue that preregistration can be useful for any kind of study in the continuum between purely confirmatory and fully exploratory.

To that end, 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. Theoretical risk serves as a general measure of juding how much evidence supports a given theory.

We then 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 or against a theory, preregistration is applicable and potentially useful for any kind of study, whether it is one of prediction, postdiction, or a mixture of both.

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 it will change their assessment of some theory's verisimilitude. 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 which study to conduct in what manner. 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 the same rational process.
3. All else being equal, 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" (XXX). That is epistemic value is gained if evidence is observed, given that it is more likely to observe evidence when the theory holds than if the theory does not hold.

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.

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 preregistration to this fact.

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 (XXX), 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. 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 generally expected that any measure of corroboration increases monotonically with an increase in posterior probability ($P(H|E)$). We assume that any measure shows gain if, and only if, evidence is observed assuming $P(E|H) > P(E|\neg H)$. This assumption is sometimes called statistical relevance condition (XXX).

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 $P(\neg E|\neg H)$. In a response to the original paper, Kukla (1990) argued that the arguments layed out in Meehl (1978) 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 (I I I I I) 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 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. 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.

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. Indeed, limiting the type-I error rate is commonly stated as a goal of preregistration (XXX). 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 & 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.

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.

Statistical methods

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, 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 toolbox for assessing and controlling their contribution to theoretical risk. Examples of our ability to exert this control are the setting of type-I error rate, 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.

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, 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 to the quantity we seek to minimize. 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. 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 $p(E|\neg H)$ 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 & Caticha, 2007).

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, 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.

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.

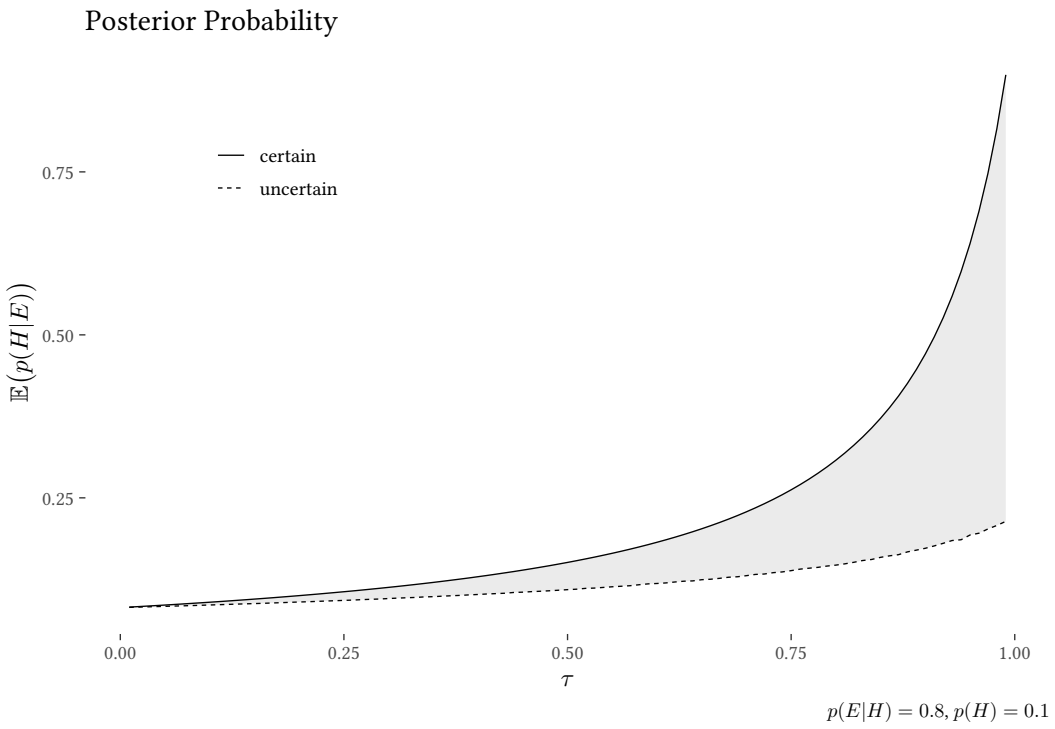

Figure 1: Posterior Probability (corroboration as firmness) as a function of τ, where τ is either certain (solid line) or maximally uncertain (dotted line).

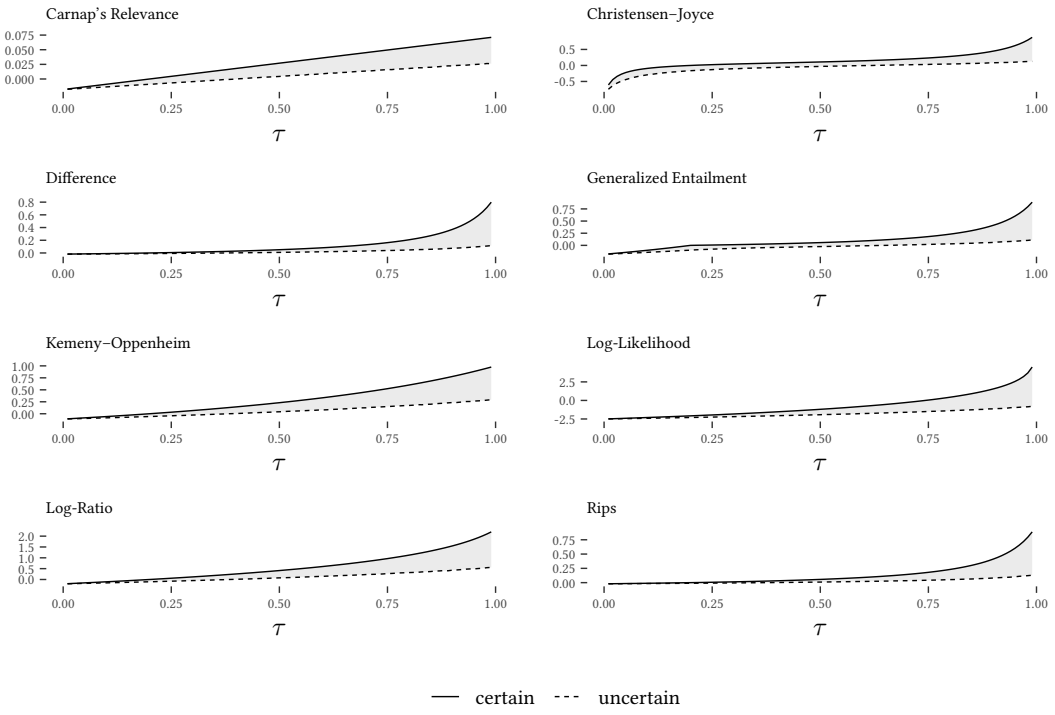

Figure 2: Several measures for corroboration as increase in firmness as a function of τ, where τ is either certain (solid line) or maximally uncertain (dotted line).

Recollect, 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, the readers of their articles. We have to remember that the process of weighing evidence for or against a theory extends beyond the original authors to all the people they hope to persuade. Unfortunately, the case for lack of insight into the myriad of factors that influence theoretical risk is particularly strong for the reader of a resulting article (or, more generally, the consumer of the research product). While the authors may have deep insight into what they did and how it might influence 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, intentionally or not. From the perspective of an ultimate sceptic, they may claim maximum uncertainty.

Communicating clearly how the authors gathered the 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 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 p hacking. The point is, that we cannot exclude this and innumerable possibilities and, hence, are left uncertain. 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.

Discussion

We started out with the observation that preregistrations do not always cleanly divide preregistration from confirmation. <–AB: can you rewrite this and be more clear? Isn't this more about researchers who do not distinguish these concepts? –> Subscribing to this distinction would mean to use preregistrations solely as a means to increase the theoretical risk researchers are willing to accept. Undoubtedly, this results in more trustworthy

evidence and may even increase the effectiveness of scientific undertakings in general. However, as argued earlier, directly maximizing theoretical risk increases the likelihood that a study fails to deliver results favoring the theory (a special instance of this problem is the well-known trade-off of statistical power and type-I error rate). As we further argued, minimizing the uncertainty about theoretical risk (without necessarily increasing the theoretical risk itself) has benefits for the epistemic value, too, but without increasing the likelihood of producing a "failed" study. Note that any single study that does not produce evidence in favor of a research hypothesis is still precious for the scientific process. Under the current incentive structure, however, many researchers have reason to avoid such outcomes and few, if any, have the privilege to operate outside of these incentives.

In our view, there are two advantages to focusing on uncertainty over theoretical risk itself. First, aiming to reduce uncertainty is beneficial across a wide range of potential theoretical risks one encounters as researcher. That is, researchers benefit from preregistering their study, regardless of whether or not the study is tagged "exploratory". If researchers conduct a more exploratory study they can clearly communicate how exploratory they aim to be and their results can be judged accordingly. Second, if researchers realize after the fact that a deviation from their preregistration in certain details is sensible (and sometimes, the only reasonable way forward), they are explicitly allowed to change their strategy. Allowing researchers to deviate from the preregistration is controversial (XXX) because the authors might sift through all the possibilities and then select the most beneficial for their theory. In its full consequence, this argument might result in the same uncertainty that we argued plagues not-preregistered studies. However, if authors of study offer a convincing argument for their deviation, their readers might may judge it to be a reasonable explanation for the deviation. In fact, using the here presented framework, we can distinguish reasonable from unreasonable deviations. If a deviation increases detectability enough to offset the increase in uncertainty in theoretical risk, it is a deviation worth doing.

References

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

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

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

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

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

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

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

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

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

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

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

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.

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 ReviewsComparing 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

Stefan, A., & Schönbrodt, F. (2022). *Big Little Lies: A Compendium and Simulation of p-Hacking Strategies*. PsyArXiv. https://doi.org/10.31234/osf.io/xy2dk

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

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