Why does preregistration increase the persuasiveness of evidence? A Bayesian

rationalization

Aaron Peikert[1,2,3] & Andreas M. Brandmaier[1, 2, 4]

[1] Max Planck Institute for Human Development

[2] Max Planck UCL Centre for Computational Psychiatry and Ageing Research

[3] Humboldt-Universität zu Berlin

[4] MSB Medical School Berlin

The materials for this article are available on GitHub (Peikert & Brandmaier, 2023a). This

version was created from git commit `52fd282`. The manuscript is available as preprint

(Peikert & Brandmaier, 2023b) and was submitted to Psychological Methods but has not

been peer reviewed.

Author Note

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: 7000

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). Despite the different names, it is fundamentally the same dichotomy that is at stake here. 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

⁶⁵ of preregistration by explicating a more general objective of preregistration that benefits all
⁶⁶ kinds of studies, even those that allow data-dependent decisions.

⁶⁷   One goal of preregistration that has received widespread attention is to clearly
⁶⁸ distinguish confirmatory from exploratory research (Bakker et al., 2020; Mellor & Nosek,
⁶⁹ 2018; Nosek et al., 2018; Simmons et al., 2021; Wagenmakers et al., 2012). In such a
⁷⁰ narrative, preregistration is justified by a confirmatory research agenda. However, two
⁷¹ problems become apparent under closer inspection. First, many researchers do not
⁷² subscribe to a purely confirmatory research agenda. Second, there is no strict mapping of
⁷³ the categories preregistered vs. non-preregistered onto the categories confirmatory
⁷⁴ vs. exploratory research.

⁷⁵   Obviously, researchers can conduct confirmatory research without preregistration —
⁷⁶ though it might be difficult to convince other researchers of the confirmatory nature of
⁷⁷ their research, that is, that they were free of cognitive biases, made no data-dependent
⁷⁸ decisions, and so forth. The opposite, that is, preregistered but not strictly confirmatory
⁷⁹ studies, are also becoming more commonplace (Chan et al., 2004; Dwan et al., 2008; Silagy
⁸⁰ et al., 2002).

⁸¹   This is the result of researchers applying one of two strategies to evade the
⁸² self-imposed restrictions of preregistrations: writing a loose preregistration, to begin with
⁸³ (Stefan & Schönbrodt, 2023) or deviating from the preregistration afterward. Both
⁸⁴ strategies may be used for sensible scientific reasons or with the self-serving intent of
⁸⁵ generating desirable results. Thus, insisting on equating preregistration and confirmation
⁸⁶ has led to the criticism that, all things considered, preregistration is actually harmful and
⁸⁷ neither sufficient nor necessary for doing good science (Pham & Oh, 2021; Szollosi et al.,
⁸⁸ 2020).

⁸⁹   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), which is inversely related to the type-I error rate in null hypothesis testing.

Further, we outline two interpretations 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 is 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 how exploratory.

Epistemic value and the Bayesian rationale

Let us start by defining what we call expected epistemic value. If researchers plan to conduct a study, they usually hope that it will change their assessment of some theory's verisimilitude (Niiniluoto, 1998). In other words, they hope to learn something from conducting the study. The amount of knowledge researchers gain from a particular study concerning the verisimilitude of a specific theory is what we call epistemic value. Researchers cannot know what exactly they will learn from a study before they run it. However, they can develop an expectation that helps them decide about the specifics of a planned study. This expectation is what we term expected epistemic value. To make our three arguments, we must assume three things about 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 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 by investing 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 modeling 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" (Fetzer, 1974; Salmon, 1970). That is, evidence only increases epistemic value for a theory if the evidence is more likely to be observed under the theory than under the alternative.

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 as minimal as possible 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 a confirmatory research agenda to this notion.

167   Let us make some simplifying assumptions and define our notation. To keep the

168   notation simple, we restrict ourselves to evidence of a binary nature (either it was observed

169   or not). We denote the probability of a hypothesis before observing evidence as $P(H)$ and

170   its complement as $P(\neg H) = 1 - P(H)$. The probability of observing evidence under some

171   hypothesis is $P(E|H)$. We can calculate the probability of the hypothesis after observing

172   the evidence with the help of the Bayes formula:

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

173   The posterior probability $P(H|E)$ is of great relevance since it is often used directly

174   or indirectly as a measure of confirmation of a hypothesis. In the tradition of Carnap, in its

175   direct use, it is called confirmation as firmness; in its relation to the a priori probability

176    $P(H)$, it is called increase in firmness (Carnap, 1950, preface to the 1962 edition). As

177   noted before, we concentrate on posterior probability as a measure of epistemic value since

178   no measure shows universally better properties than others. However, it is reasonable that

179   any measure of confirmation increases monotonically with an increase in posterior

180   probability $P(H|E)$, and our argument applies to those measures as well.

181   In short, we want to increase posterior probability $P(H|E)$. Increases in posterior

182   probability $P(H|E)$ are associated with increased epistemic value, of which we want to

183   maximize the expectation. So how can we increase posterior probability? The Bayes

184   formula yields three components that influence confirmation, namely $P(H)$, $P(E|H)$ and

185    $P(E)$. The first option leads us to the unsurprising conclusion that higher a priori

186   probability $P(H)$ leads to higher posterior probability $P(H|E)$. If a hypothesis is more

187   probable to begin with, observing evidence in its favor will result in a hypothesis that is

188   more strongly confirmed, all else being equal. However, the prior probability of a

189   hypothesis is nothing our study design can change. The second option is similarly

190 commonsensical; that is, an increase in $P(E|H)$ leads to a higher posterior probability

191 $P(H|E)$. $P(E|H)$ is the probability of obtaining evidence for a hypothesis when it holds.

192 We call this probability of detecting evidence, given that the hypothesis holds

193 "detectability." Consequently, researchers should ensure that their study design allows them

194 to find evidence for their hypothesis, in case it is true. When applied strictly within the

195 bounds of null hypothesis testing, detectability is equivalent to power (or the complement

196 of type-II error rate). However, while detectability is of great importance for study design,

197 it is not directly relevant to the objective of preregistration. Thus, $P(E)$ remains to be

198 considered. Since $P(E)$ is the denominator, decreasing it can increase the posterior

199 probability. In other words, high risk, high reward.

200   If we equate riskiness with a low probability of obtaining evidence (when the

201 hypothesis is false), the Bayesian rationale perfectly aligns with the observation that risky

202 predictions lead to persuasive evidence. This tension between high risk leading to high gain

203 is central to our consideration of preregistration. A high-risk, high-gain strategy is bound

204 to result in many losses that are eventually absorbed by the high gains. Sustaining many

205 "failed" studies is not exactly aligned with the incentive structure under which many, if not

206 most, researchers operate. Consequently, researchers are incentivized to appear to take

207 more risks than they actually do, which misleads their readers to give their claims more

208 credence than they deserve. It is at this juncture that the practice and mispractice of

209 preregistration comes into play. We argue that the main function of preregistration is to

210 enable proper judgment of the riskiness of a study.

211   To better understand how preregistrations can achieve that, let us take a closer look

212 at the factors contributing to $P(E)$. Using the law of total probability, we can split $P(E)$

213 into two terms:

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

214   We have already noted that there is not much to be done about prior probability

215 $(P(H)$, and hence its counter probability $P(\neg H))$, and that it is common sense to increase

216 detectability $P(E|H)$. The real lever to pull is therefore $P(E|\neg H)$. This probability tells

217 us how likely it is that we find evidence in favor of the theory when in fact, the theory is

218 not true. Its counter probability $P(\neg E|\neg H) = 1 - P(E|\neg H)$ is what we call "theoretical

219 risk", because it is the risk a theory takes on in predicting the occurrence of particular

220 evidence in its favor. We "borrow" the term from Meehl (1978), though he has not

221 assigned it to the probability $P(\neg E|\neg H)$. Kukla (1990) argued that the core arguments in

222 Meehl (1990) can be reconstructed in a purely Bayesian framework. However, while he did

223 not mention $P(\neg E|\neg H)$ he suggested that Meehl (1978) used the term "very strange

224 coincidence" for a small $P(E|\neg H)$ which would imply, that $P(\neg E|\neg H)$ can be related to or

225 even equated to theoretical risk.

226   Let us note some interesting properties of theoretical risk $P(\neg E|\neg H)$. First,

227 increasing theoretical risk leads to higher posterior probability $P(H|E)$, our objective.

228 Second, if the theoretical risk is smaller than detectability $P(E|H)$ it follows that the

229 posterior probability must decrease when observing the evidence. If detectability exceeds

230 theoretical risk, the evidence is less likely under the theory than it is when the theory does

231 not hold. Third, if the theoretical risk equals zero, then posterior probability is at best

232 equal to prior probability but only if detectability is perfect $(P(H|E) = 1)$. In other words,

233 observing a sure fact does not lend credence to a hypothesis.

234   The last statement sounds like a truism but is directly related to Popper's seminal

235 criterion of demarcation. He stated that if it is impossible to prove that a hypothesis is

236 false $(P(\neg E|\neg H) = 0$, theoretical risk is zero), it cannot be considered a scientific

237 hypothesis (Popper, 2002, p. 18). We note these relations to underline that the Bayesian

238 rationale we apply here is able to reconstruct many commonly held views on riskiness and

239 epistemic value.

240   Both theoretical risk $P(\neg E|\neg H)$ and detectability $P(E|H)$ aggregate countless

241 influences; otherwise, they could not model the process of evidential support for theories.

242 To illustrate the concepts we have introduced here, consider the following example of a

243 single theory and three experiments that may test it. The experiments were created to

244 illustrate how they may differ in their theoretical risk and detectability. Suppose the

245 primary theory is about the cognitive phenomenon of "insight." For the purpose of

246 illustration, we define it, with quite some hand-waving, as a cognitive abstraction that

247 allows agents to consistently solve a well-defined class of problems. We present the

248 hypothesis that the following problem belongs to such a class of insight problems:

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

250   We propose three experiments that differ in theoretical risk and detectability. All

251 experiments take a sample of ten psychology students. We present the students with the

252 problem for a brief span of time. After that, the three experiments differ as follows:

253 1. The experimenter gives a hint that the problem is easy to solve when using Roman

254   numerals; if all students come up with the solution, she records it as evidence for the

255   hypothesis.

256 2. The experimenter shows the solution "VIII" and explains it; if all students come up

257   with the solution, she records it as evidence for the hypothesis.

258 3. The experimenter does nothing; if all students come up with the solution, she records

259   it as evidence for the hypothesis.

260   We argue that experiment 1 has high theoretical risk $P(\neg E_1|\neg H)$ and high

261 detectability $P(E_1|H)$. If "insight" has nothing to do with solving the problem $(\neg H)$, then

262 presenting the insight that Roman numerals can be used should not lead to all students

263 solving the problem $(\neg E_1)$; the experiment, therefore, has high theoretical risk

264 $P(\neg E_1|\neg H)$. 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 epistemic

value, it is intuitive to task preregistration with maximizing theoretical risk, i.e., a

confirmatory research agenda. Indeed, limiting the type-I error rate is commonly stated as

the central goal of preregistration (Nosek et al., 2018; Oberauer, 2019; Rubin, 2020). We

argue that while such a conclusion is plausible, we must first consider at least two

constraints that place an upper bound on the theoretical risk.

First, the theory itself limits theoretical risk: Some theories simply do not make

risky predictions, and preregistration will not change that. Consider the case of a

²⁹¹ researcher contemplating the relation between two sets of variables. Suppose each set is

²⁹² separately well studied, and strong theories tell the researcher how the variables within the

²⁹³ set relate. However, our imaginary researcher now considers the relation between these two

²⁹⁴ sets. For lack of a better theory, they assume that some relation between any variables of

²⁹⁵ the two sets exists. This is not a risky prediction to make in psychology (Orben & Lakens,

²⁹⁶ 2020). However, we would consider it a success if the researcher would use the evidence

²⁹⁷ from this rather exploratory study to develop a more precise (and therefore risky) theory,

²⁹⁸ e.g., by using the results to specify which variables from one set relate to which variables

²⁹⁹ from the other set, to what extent, in which direction, with which functional shape, etc., to

³⁰⁰ be able to make riskier predictions in the future. We will later show that preregistration

³⁰¹ increases the degree of belief in the further specified theory, though it remains low till

³⁰² being substantiated by testing the theory again. This is because preregistration increases

³⁰³ the expected epistemic value regardless of the theory being tested, as we will show.

³⁰⁴ Second, available resources limit theoretical risk. Increasing theoretical risk

³⁰⁵ $P(\neg E|\neg H)$ will usually decrease detectability $P(E|H)$ unless more resources are invested.

³⁰⁶ In other words, one cannot increase power while maintaining the same type-I error rate

³⁰⁷ without increasing the invested resources. Tasking preregistration with an increase in

³⁰⁸ theoretical risk makes it difficult to balance this trade-off. Mindlessly maximizing

³⁰⁹ theoretical risk would either never produce evidence or require huge amounts of resources.

Uncertainty about theoretical risk

³¹¹ We have established that higher theoretical risk leads to more persuasive evidence.

³¹² In other words, we have reconstructed the interpretation that preregistrations supposedly

³¹³ work by restricting the researchers, which in turn increases the theoretical risk (or

³¹⁴ equivalently limits the type-I error rate) and thereby creates more compelling evidence.

³¹⁵ Nevertheless, there are trade-offs for increasing theoretical risk. Employing a mathematical

³¹⁶ framework allows us to navigate the trade-offs more effectively and move towards a second,

317 more favorable interpretation. To that end, we incorporate uncertainty about theoretical

318 risk into our framework.

Statistical methods

320   One widely known factor is the contribution of statistical methods to theoretical

321 risk. Theoretical risk $P(\neg E|\neg H)$ is deeply connected with statistical methods, because it is

322 related to the type-I error rate in statistical hypothesis testing $P(E|\neg H)$ by

323 $P(\neg E|\neg H) = 1 - P(E|\neg H)$, if you consider the overly simplistic case where the research

324 hypothesis is equal to the statistical alternative-hypothesis because then the nill-hypothesis

325 is $\neg H$. Because many researchers are familiar with the type-I error rate, it can be helpful

326 to remember this connection to theoretical risk. Researchers who choose a smaller type-I

327 error rate can be more sure of their results, if significant, because the theoretical risk is

328 higher. However, this connection should not be overinterpreted for two reasons. First,

329 according to most interpretations of null hypothesis testing, the absence of a significant

330 result should not generally be interpreted as evidence against the hypothesis (Mayo, 2018,

331 p. 5.3). Second, the research hypothesis seldomly equals the statistical

332 alternative-hypothesis. We argue that theoretical risk (and hence its complement,

333 $P(E|\neg H)$) also encompasses factors outside the statistical realm, most notably the study

334 design and broader analytical strategies.

335   Statistical methods stand out among these factors because we have a large and

336 well-understood toolbox for assessing and controlling their contribution to theoretical risk.

337 Examples of our ability to exert this control are the choice of type-I error rate, adjustments

338 for multiple testing, the use of corrected fit measures (i.e., adjusted R^2), information

339 criteria, or cross-validation in machine learning. These tools help us account for biases in

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

341   The point is that the contribution of statistical methods to theoretical risk can be

342 formally assessed. For many statistical models it can be analytically computed under some

343 assumptions. For those models or assumptions where this is impossible, one can employ

344 Monte Carlo simulation to estimate the contribution to theoretical risk. The precision with

345 which statisticians can discuss contributions to theoretical risk has lured the community

346 concerned with research methods into ignoring other factors that are much more uncertain.

347 We cannot hope to resolve this uncertainty; but we have to be aware of its implications.

348 These are presented in the following.

349 **Sources of Uncertainty**

350 As we have noted, it is possible to quantify how statistical models affect the

351 theoretical risk based on mathematical considerations and simulation. However, other

352 factors in the broader context of a study are much harder to quantify. If one chooses to

353 focus only on the contribution of statistical methods to theoretical risk, one is bound to

354 overestimate it. Take, for example, a t-test of mean differences in two samples. Under ideal

355 circumstances (assumption of independence, normality of residuals, equal variance), it

356 stays true to its type-I error rate. However, researchers may do many very reasonable

357 things in the broader context of the study that affect theoretical risk: They might exclude

358 outliers, choose to drop an item before computing a sum score, broaden their definition of

359 the population to be sampled, translate their questionnaires into a different language,

360 impute missing values, switch between different estimators of the pooled variance, or any

361 number of other things. All of these decisions carry a small risk that they will increase the

362 likelihood of obtaining evidence despite the underlying research hypothesis being false.

363 Even if the t-test itself perfectly maintains its type I error rate, these factors influence

364 $P(E|\neg H)$. While, in theory, these factors may leave $P(E|\neg H)$ unaffected or even decrease

365 it, we argue that this is not the case in practice. Whether researchers want to or not, they

366 continuously process information about how the study is going, except under strict

367 blinding. While one can hope that processing this information does not affect their

368 decision-making either way, this cannot be ascertained. Therefore, we conclude that

369 statistical properties only guarantee a lower bound for theoretical risk. The only thing we

370 can conclude with some certainty is that theoretical risk is not higher than what the

371 statistical model guarantees without knowledge about the other factors at play.

372 **The effects of uncertainty**

373   Before we ask how preregistration influences this uncertainty, we must consider the

374 implications of being uncertain about the theoretical risk. Within the Bayesian framework,

375 this is both straightforward and insightful. Let us assume a researcher is reading a study

376 from another lab and tries to decide whether and how much the presented results confirm

377 the hypothesis. As the researcher did not conduct the study (and the study is not

378 preregistered), they can not be certain about the various factors influencing theoretical risk

379 (researcher degrees of freedom). We therefore express this uncertainty about the theoretical

380 risk as a probability distribution Q of $P(E|\neg H)$ (remember that $P(E|\neg H)$ is related to

381 theoretical risk by $P(E|\neg H) = 1 - P(\neg E|\neg H)$, so it does not matter whether we consider

382 the distribution of theoretical risk or $P(E|\neg H)$). To get the expected value of $P(H|E)$

383 that follows from the researchers' uncertainty about the theoretical risk, we can compute

384 the expectation using Bayes theorem:

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

385   Of course, the assigned probabilities and the distribution Q vary from study to

386 study and researcher to researcher, but we can illustrate the effect of uncertainty with an

387 example. Assuming $P(E|H) = 0.8$ (relective of the typically strived for power of 80%). Let

388 us further assume that the tested hypothesis is considered unlikely to be true by the

389 research community before the study is conducted ($P(H) = 0.1$) and assign a uniform

390 distribution for $P(E|\neg H) \sim U([1-\tau, 1])$ where τ is set to $1 - \alpha$, reflecting our assumption

391 that this term gives an upper bound for theoretical risk $P(\neg E|\neg H)$. We chose this uniform

392 distribution as it is the maximum entropy distribution with support $[1-\tau, 1]$ and hence

393 conforms to our Bayesian framework (Giffin & Caticha, 2007).

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

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

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

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

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 gained epistemic value 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 epistemic value.

The argument for a confirmatory research agenda is that by increasing theoretical risk we increase expected epistemic value, i.e., moving to the right on the x-axis in Figure 1 increases posterior probability (on the y-axis). However, if a hypothesis in a certain study has low theoretical risk, there is not much researchers can do about it. However, studies do not only differ by how high the theoretical risk is but also 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 epistemic value) 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 positive rate of 0.20) moving from uncertain to certain increases the posterior from 0.15 to

0.31.

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 epistemic value for other researchers.

The point we make with these assumptions is that researchers aim to persuade other researchers, for example, the readers of their articles. Not only the original authors are concerned with the process of weighing evidence for or against a theory but really 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.

439 Communicating clearly how authors of a scientific report collected their data and

440 consequently analyzed it to arrive at the evidence they present is crucial for judging the

441 theoretical risk they took. Preregistrations are ideal for communicating just that because

442 any description after the fact is prone to be incomplete. For instance, the authors could

443 have opted for selective reporting, that is, they decided to exclude a number of analytic

444 strategies they tried out. That is not to say that every study that was not-preregistered

445 was subjected to practices of questionable research practices. The point is that we cannot

446 exclude it with certainty. This uncertainty is drastically reduced if the researchers have

447 described what they intended to do beforehand and then report that they did exactly that.

448 In that case, readers can be certain they received a complete account of the situation.

449 They still might be uncertain about the actual theoretical risk the authors took, but to a

450 much smaller extent than if the study would not have been preregistered. The remaining

451 sources of uncertainty might be unfamiliarity with statistical methods or experimental

452 paradigms used, the probability of an implementation error in the statistical analyses, a

453 bug in the software used for analyses, etc. In any case, a well-written preregistration

454 should aim to reduce the uncertainty about the theoretical risk and hence increase the

455 persuasiveness of evidence. Therefore, a study that perfectly adhered to its preregistration

456 will resemble the solid line in Figure 1/2. Crucially, perfect means here that the theoretical

457 risk can be judged with low uncertainty, not that the theoretical risk is necessarily high.

Discussion

459 To summarize, we showed that both higher theoretical risk and lower uncertainty

460 about theoretical risk lead to higher expected epistemic value across a variety of measures.

461 The former result that increasing theoretical risk leads to higher expected epistemic value

462 reconstructs the appeal and central goal of preregistration of confirmatory research

463 agendas. However, theoretical risk is something researchers have only limited control over.

464 For example, theories are often vague and ill-defined, resources are limited, and increasing

465 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).

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

Focusing on uncertainty reduction also explains two common practices of preregistration that do not align with a confirmatory research agenda. First, researchers seldomly predict precise numerical outcomes, instead they use preregistrations to describe the process that generates the results. Precise predictions would have very high theoretical risk (they are likely incorrect if the theory is wrong). A statistical procedure may have high or low theoretical risk depending on the specifics of the model used. Specifying the process, therefore, is in line with the 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 epistemic value is decreased compared to a world where they anticipated the need to deal with missing data. However, it is still justified to deviate. After all they have not anticipated the case and are left with a detectablilty of zero. Any decision will increase detectability to a non-zero value offsetting the increase in uncertainty. The researchers also

⁵¹⁹ may do their best to argue that the deviation was not motivated by increasing theoretical

⁵²⁰ risk, thereby, decreasing the uncertainty. Ideally, there is a default decision that fits well

⁵²¹ with the theory or with the study design. Or, if there is no obvious candidate, the

⁵²² researchers could conduct a multiverse analysis of the available options to deal with

⁵²³ missings to show the influence of the decision (Steegen et al., 2016).

⁵²⁴   As explained above, reduction in uncertainty as the objective for preregistration

⁵²⁵ does not only explain some existing practice, that does not align with confirmation as a

⁵²⁶ goal, it also allows to form recommendations to improve the practice of preregistration.

⁵²⁷ Importantly, we now have a theoretical measure to gauge the functionality of

⁵²⁸ preregistrations, which can only help increase its utility. In particular, a preregistration

⁵²⁹ should be specific about the procedure that is intended to generate evidence for a theory.

⁵³⁰ Such a procedure may accommodate a wide range of possible data, i.e., it may be

⁵³¹ exploratory. The theoretical risk, however low, must be communicated clearly. Parts of the

⁵³² process left unspecified imply uncertainty, which preregistration should reduce. However,

⁵³³ specifying procedures that can be expected to fail will lead to deviation and, subsequently,

⁵³⁴ to larger uncertainty.

⁵³⁵   We have proposed a workflow for preregistration called preregistration as code

⁵³⁶ (PAC) elsewhere (Peikert et al., 2021). In a PAC, researchers use computer code for the

⁵³⁷ planned analysis as well as a verbal description of theory and methods for the

⁵³⁸ preregistration. This combination is facilitated by dynamic document generation, where

⁵³⁹ the results of the code, such as numbers, figures, and tables, are inserted automatically into

⁵⁴⁰ the document. The idea is that the preregistration already contains "mock results" based

⁵⁴¹ on simulated or pilot data, which are replaced after the actual study data becomes

⁵⁴² available. Such an approach dissolves the distinction between the preregistration document

⁵⁴³ and the final scientific report. Instead of separate documents, preregistration, and final

⁵⁴⁴ 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 epistemic value.

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

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.

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

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

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

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

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

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

Kukla, A. (1990). Clinical Versus Statistical Theory Appraisal. *Psychological Inquiry*, *1*(2),

160–161. https://doi.org/10.1207/s15327965pli0102_9

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

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.

618   https://doi.org/10.1207/s15327965pli0102_1

619 Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the

620   slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4),

621   806–834. https://doi.org/10.1037/0022-006X.46.4.806

622 Mellor, D. T., & Nosek, B. A. (2018). Easy preregistration will benefit any research.

623   *Nature Human Behaviour*, *2*(2), 98–98. https://doi.org/10.1038/s41562-018-0294-7

624 Niiniluoto, I. (1998). Verisimilitude: The Third Period. *The British Journal for the*

625   *Philosophy of Science*, *49*(1), 1–29. https://doi.org/10.1093/bjps/49.1.1

626 Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration

627   revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606.

628   https://doi.org/10.1073/pnas.1708274114

629 Oberauer, K. (2019). Preregistration of a forking path – What does it add to the garden of

630   evidence? In *Psychonomic Society Featured Content*.

631 Open Science Collaboration. (2015). Estimating the reproducibility of psychological

632   science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

633 Orben, A., & Lakens, D. (2020). Crud (Re)Defined. *Advances in Methods and Practices in*

634   *Psychological Science*, *3*(2), 238–247. https://doi.org/10.1177/2515245920917961

635 Peikert, A., & Brandmaier, A. M. (2023a). *Supplemental materials for preprint: Why does*

636   *preregistration increase the persuasiveness of evidence? A Bayesian rationalization.*

637   Zenodo. https://doi.org/10.5281/zenodo.7648471

638 Peikert, A., & Brandmaier, A. M. (2023b). *Why does preregistration increase the*

639   *persuasiveness of evidence? A Bayesian rationalization.* PsyArXiv; PsyArXiv.

640   https://doi.org/10.31234/osf.io/cs8wb

641 Peikert, A., van Lissa, C. J., & Brandmaier, A. M. (2021). Reproducible Research in R: A

642   Tutorial on How to Do the Same Thing More Than Once. *Psych*, *3*(4), 836–867.

643   https://doi.org/10.3390/psych3040053

644 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

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

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

Posterior Probability
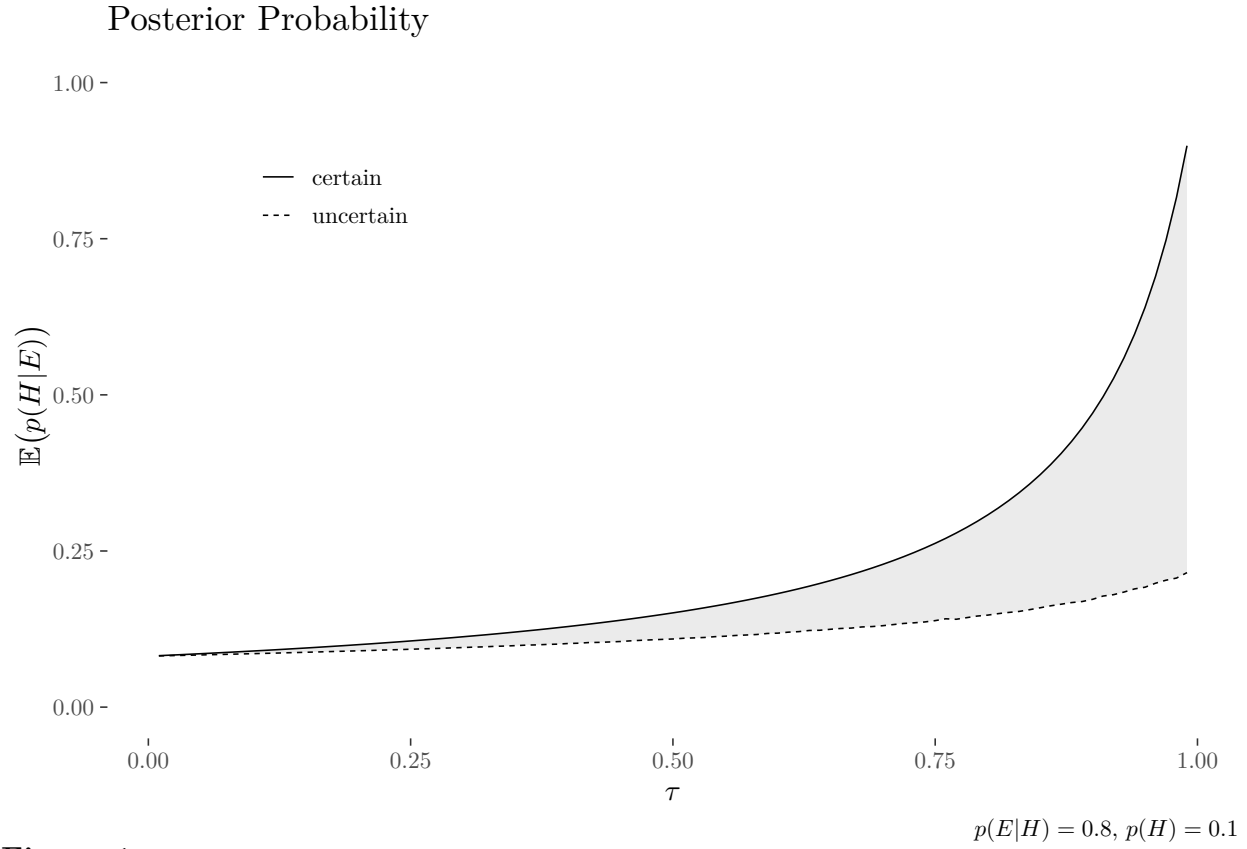
Figure 1

Posterior probability (confirmation as firmness) as a function of theoretical risk τ, where τ is either certain (solid line) or maximally uncertain (dotted line).
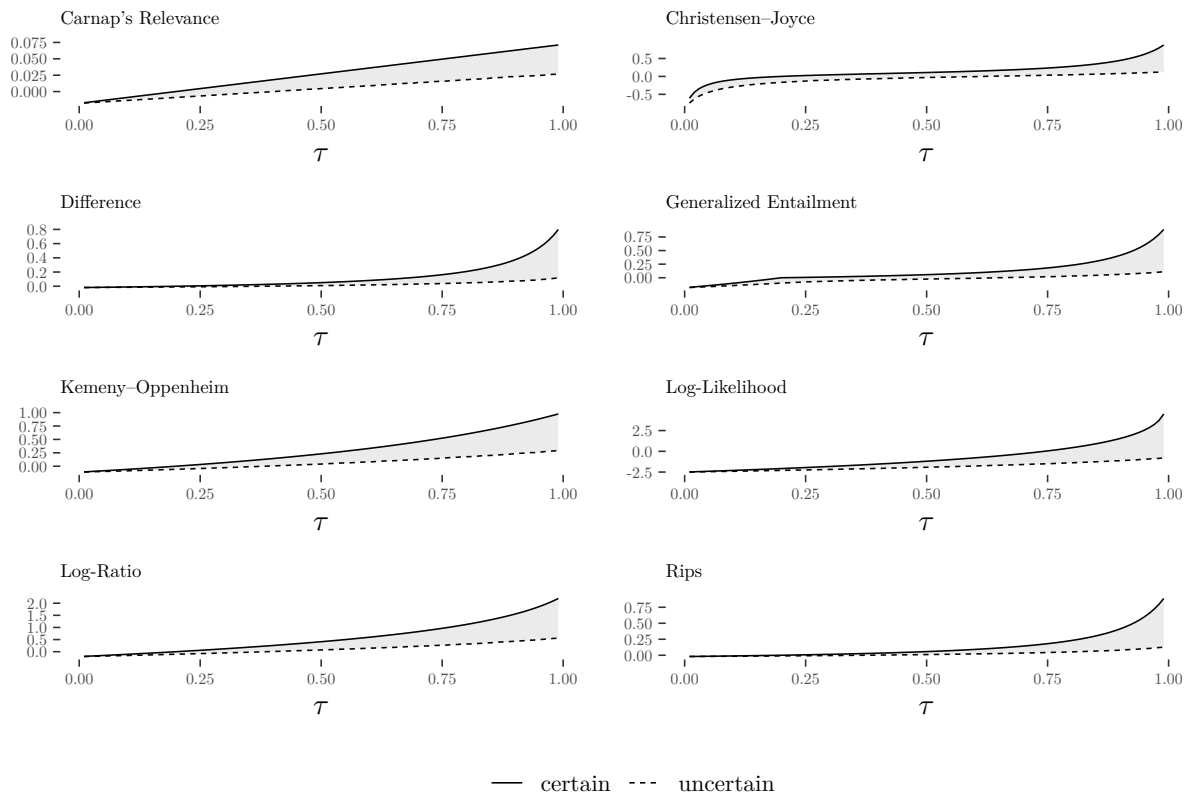
Figure 2

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).