The Objectivity of Subjective Bayesianism

Jan Sprenger*

July 12, 2017

Abstract

Subjective Bayesianism is a major school of uncertain reasoning and statistical inference. It is often criticized for a lack of objectivity: (i) it opens the door to the influence of values and biases, (ii) evidence judgments can vary substantially between scientists, (iii) it is not suited for informing policy decisions. My paper rebuts these concerns by bridging the debates on scientific objectivity and statistical method. First, I show that the above concerns arise equally for standard frequentist inference. Second, I argue that the involved senses of objectivity are epistemically inert. Third, I show that Subjective Bayesianism promotes other, epistemically relevant senses of scientific objectivity—most notably by increasing the transparency of scientific reasoning.

1 Introduction

How, and in what sense, can statistical inference be objective? In times of questionable research practices, frequent replication failures, and dwindling trust in science, this question is of the utmost importance for philosophy of statistics, and science in general.

The present paper studies the objectivity of Subjective Bayesianism, a major school of uncertain reasoning. It models an agent's rational degrees of belief in a hypothesis as following the laws of probability. The version of Subjective Bayesianism that I defend in this paper is shared by many practitioners (e.g., Goodman, 1999; Howson and

*Contact information: Tilburg Center for Logic, Ethics and Philosophy of Science (TiLPS), Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. Email: j.sprenger@uvt.nl. Webpage: www.laeuferpaar.de

Urbach, 2006; Lee and Wagenmakers, 2013): empirical evidence informs (prior) degrees of belief, but it is usually not sufficient to determine them in a uniquely rational way.

Subjective Bayesians revise their degrees of belief in a proposition (e.g., a scientific hypothesis) by the principle of Bayesian Conditionalization. Assume, for example, that you want to assess a hypothesis H and that $p(H)$ describes your prior degree of belief in H. You observe data D. Then, your posterior degree of belief in H after learning D is equal to the conditional probability of H given D: $p(H|D)$. This posterior probability is given by **Bayes' Theorem**:

$$p(H|D) = \frac{p(H)p(D|H)}{p(D)} \tag{1}$$

where $p(D) = \sum_{H_i \in \mathcal{H}} p(D|H_i)p(H_i)$ is the marginal probability of data D relative to the elements of the hypothesis space $\mathcal{H}$.

After observing D, the posterior distribution $p(\cdot|D)$ serves as a basis for inference and decision-making. For example, if H is the hypothesis that a new medical drug is no better than a placebo, and if H is probable given D, then we will stop developing the drug any further. In addition, the divergence between prior and posterior distribution can be used for quantifying the degree to which D confirms H (Fitelson, 2001; Crupi, 2013).

There is a prima facie tension between Subjective Bayesianism and the pursuit of scientific objectivity: the subjective elements in Bayesian reasoning, above all the choice of a prior distribution, barely match widely endorsed senses of scientific objectivity, such as intersubjectivity, value freedom, and conformity to standardized inference protocols. Since the epistemic authority of science in guiding public policy leans on the objectivity of scientific inference, it is sometimes claimed that "a notion of probability as personalistic degree of belief [...], by its very nature, is not focused on the extraction and presentation of evidence of a public and objective kind" (Cox and Mayo, 2010, 298). This view is echoed in writings of well-known statisticians, methodologists and philosophers of science such as Fisher (1956), Mayo (1996), Popper (2002) and Senn (2011).

These objections are prima facie plausible, but they are rarely buttressed by a conceptual analysis of scientific objectivity. By transferring insights from the objectivity

debate (e.g., Longino, 1990; Megill, 1994; Douglas, 2009) to the context of statistical inference, I try to assess the probative value of the above objections, and the degree to which Subjective Bayesianism can be objective. First, I present the criticisms of Subjective Bayesianism in somewhat greater detail (Section 2) and I show that two tempting responses are insufficient (Section 3). Then I argue that the criticisms apply equally to the main competitor of Subjective Bayesianism—frequentist inference with significance tests (Section 4).[1] What is more, the criticisms are based on outdated (yet popular) and restrictive readings of scientific objectivity (Section 5) with dubious epistemic value. Section 6 explains why Bayesian inference promotes relevant senses of scientific objectivity (e.g., robustness, transparency, facilitating discussion and criticism) that are not captured by traditional accounts. I support my argument with a case study from social psychology. Section 7 wraps up the main insights. Thus, the paper does not only defend the objectivity of Subjective Bayesianism: it also shows how conceptual work on scientific objectivity bears on questions in the methodology of statistical inference.

2 The Objections

Objectivity is a label that can be attached to different aspects of science: to the claims of a theory in relation to the world, to the process of gathering data, to individual reasoning about scientific theories, and to the social dimension of producing scientific knowledge (Longino, 1990; Douglas, 2004, 2009; Reiss and Sprenger, 2014). These carriers correspond to diverse senses of scientific objectivity. I follow Heather Douglas's taxonomy which distinguishes eight senses of objectivity in science, and I adapt them (where appropriate) to the context of statistical inference. Three senses pose obvious challenges for Subjective Bayesianism:

Concordant Objectivity (Intersubjectivity) Different speakers or community members agree on the reality of an observation, an evidence claim or a judgment on a theory. This sense of objectivity is purely factual, not about the way agreement is reached.

[1] A comparison of Subjective and Objective Bayesianism would also be of great interest, but beyond the scope of this already lengthy paper—especially because Objective Bayesianism is no monolithic block, but contains different varieties and approaches (e.g. Jeffreys, 1961; Jaynes, 1968; Williamson, 2010; Bernardo, 2012).

Value-Free Objectivity Values and subjective judgments are banned from the process of scientific reasoning (e.g., in assessing theories on the basis of observed evidence). This sense of scientific objectivity sees values as detrimental to the unbiasedness and impartiality of scientific research.

Procedural Objectivity Experimentation and reasoning processes are standardized according to specific protocols. The point of this sense of objectivity is to eliminate individual idiosyncrasies and to obtain always the same result, regardless of who performs an experiment or data analysis (see also Porter, 1996). It is particularly influential in the life sciences, where strict standards regulate the design, conduct and interpretation of medical trials.

Let us begin with concordant objectivity, or equivalently, intersubjectivity. This notion has a long philosophical tradition, e.g., Quine (1992, 5) stated that "the requirement of intersubjectivity is what makes science objective". Subjective Bayesian inference violates concordant objectivity because different scientists typically use different priors for analyzing one and the same dataset, leading to different conclusions. Since consensus remains elusive, it is open which (and whose) probability assessments should inform judgments on theories, and evidence-based public policy.

The failure of Subjective Bayesianism with respect to value-free objectivity may be even more worrying. In sensitive areas such as climate science and the biomedical sciences, financial and ethical stakes are high, and consequences of wrong decisions are severe. These fields strive for inference methods that are as impartial and evidence-based as possible, and the pronounced role of personal degrees of belief in Subjective Bayesianism seems to jeopardize that aim. In the words of the medical methodologist Lemuel Moyé:

> Without specific safeguards, use of Bayesian procedures will set the stage for the entry of non-fact-based information that, unable to make it through the "evidence-based" front door, will sneak in through the back door of "prior distributions". There, it will wield its influence, perhaps wreaking havoc on the research's interpretation. (Moyé, 2008, 476)

The objection can be rephrased as saying that the choice of a prior, which cannot always be based on hard information, will bias the final result in a particular direction.

It is clear that such a liberal procedure cannot be objective in the sense of being value-free. Similarly, the discretion to choose a prior distribution at will is at odds with the goal of attaining procedural objectivity by means of standardized, uniform statistical analysis procedures.

All these tensions between Subjective Bayesianism and various senses of scientific objectivity support Cox and Mayo's intuition that (subjective) Bayesian methods fail to quantify objective evidence for use in science and public policy. Indeed, one may even conclude that Subjective Bayesians commit a category mistake. Their formalism answers the question of what we may reasonably believe, but it does not quantify the (objective) evidence for a scientific claim (Royall, 1997, 4).

The following sections respond to these worries. Before that, however, I would like to explain why two popular defenses of Subjective Bayesianism fail to counter the objections.

3 Convergence Theorems and Bayes Factors

A standard reply to the above worries contends that concordant objectivity may not hold at the beginning of a research process, but it will be attained in the long run: "If a fairly sharp consensus of views emerges from a rather wide spread of initial opinions, then, and only then, might it be meaningful to refer to 'objectivity'." (Smith, 1986, 10). And indeed, the famous merging-of-opinions or washing-out theorems (Blackwell and Dubins, 1962; Gaifman and Snir, 1982) show that Bayesians eventually reach such a consensus. The theorems study the limiting behavior of two agents' degrees of belief when they are informed by the same body of evidence. In a nutshell, the posterior degrees of belief p^N and q^N will converge when collecting more and more information ($N \to \infty$), as long as their prior degrees of belief p and q assign probability zero to the same propositions.[2] This means that differences in prior probability will eventually wash out.

Unfortunately, this observation fails to alleviate the worries from the previous section. In practice, we deal with finite datasets to which the merging theorems do not apply. They are purely asymptotic claims. Neither do they bound the speed of

[2]This means that $p(X) = 0 \Leftrightarrow q(X) = 0$; a property known as absolute continuity of probability measures. Notably, the convergence is uniform, that it, it holds simultaneously for all elements of the probability space.

convergence, preventing an application to small to medium-sized datasets (see also Earman, 1992, 148–149). The merging-of-opinion theorems do not state sufficient, but only *necessary* conditions for a calculus of degree of belief that pursues the goal of objectivity—just as statistical consistency (=convergence to the true value as sample size increases) is a necessary, but not a sufficient property for the goodness of a statistical estimator. In particular, the merging theorems do not justify the claim that Subjective Bayesianism achieves concordant objectivity.

Another reply proposes to shift the objectivity discourse from posterior distributions to **Bayes factors**. While posterior distributions are an important basis for action and decision-making, they do not make any claims about the weight of a particular body of evidence. But it is those claims that we want to be objective—especially because they allow non-experts to assess the statistical evidence and to make informed policy decisions, without committing themselves to contentious prior assumptions. Bayes factors (Kass and Raftery, 1995; Rouder et al., 2009) address this question by describing how much data D favors hypothesis H_1 over its competitor H_0.

$$BF_{10}(D) \quad := \quad \frac{p(H_0|D)}{p(H_1|D)} \cdot \frac{p(H_1)}{p(H_0)} = \frac{p(D|H_1)}{p(D|H_0)} \tag{2}$$

In other words, the Bayes factor measures the discriminative power of D with respect to H_1 and H_0 by comparing the (average) probability of D under H_1 to the (average) probability of D under H_0. The higher $BF_{10}(D)$ is, the more the data speak for H_1, and vice versa. Table 1 provides a standardized interpretation scheme for Bayes factors.

Notably, the Bayes factor is independent of how strongly one is convinced of H_0 as opposed to H_1 a priori because $p(H_0)$ and $p(H_1)$ cancel out when applying Bayes' Theorem to $p(H_0|D)$ and $p(H_1|D)$. However, there *is* a crucial dependence on prior probabilities. Assume that μ is a real-valued parameter and that we test the point hypothesis $H_0 : \mu = 0$ against the alternative $H_1 : \mu \neq 0$. This is a very common setting in science. In such a case, the Bayes factor will typically depend on how spread out the prior distribution for H_1 is: the extreme values of μ will badly fit the observed data, driving down $p(D|H_1) = \int_{\mu \in \mathbb{R}} p(D|\mu)\, p(\mu)\, d\mu$ and thus, also $BF_{10}(D)$. The phenomenon is reversed for priors that are concentrated around $\mu = 0$. Hence, individual differences in shaping the prior distribution influence—and possibly bias—the final evidential claims. While I agree that Subjective Bayesians should use Bayes

Bayes Factor BF_{10}	Interpretation
$>$100	Extreme evidence for H_1
30–100	Very strong evidence for H_1
10–30	Strong evidence for H_1
3–10	Moderate evidence for H_1
1–3	Anecdotal evidence for H_1
1	No evidence for either hypothesis
1/3–1	Anecdotal evidence for H_0
1/3–1/10	Moderate evidence for H_0
1/10–1/30	Strong evidence for H_0
1/30–1/100	Very strong evidence for H_0
$<$1/100	Extreme evidence for H_0

Table 1: Classification of Bayes Factors according to Lee and Wagenmakers (2013), adjusted from Jeffreys (1961).

factors for quantifying statistical evidence, this move alone is not enough to rebut the objections concerning lack of value freedom, procedural uniformity and intersubjective agreement.

4 Frequentism and Scientific Objectivity

In this section, I argue that the concerns about the objectivity of Subjective Bayesianism apply equally to its main competitor: frequentist inference. In that school of statistics, hypotheses are either true or false, and no objects of subjective uncertainty. Inferences are justified on the basis of their long-run properties, not on the basis of posterior probabilities. Among the many varieties of frequentist inference (Fisher, 1956; Neyman and Pearson, 1967; Mayo, 1996), I focus on the most widespread one: Null Hypothesis Significance Testing (NHST), and the use of p-values for quantifying statistical evidence. (I will comment on extensions of NHST and alternative frequentist approaches at the end of this section.) Notably, none of the arguments below depend on the frequent misuse and misinterpretation of NHST in statistical practice, or on objections to the logic of NHST. I focus on the classical problem of testing a point null hypothesis H_0 against an unspecific alternative H_1.

For the sake of simplicity, suppose that we analyze (approximately) Normally distributed data $D = (X_1, \ldots, X_N)$ with unknown mean μ—the parameter of interest—

and unknown variance σ^2.[3] We calculate the sample mean $\bar{X}$ and the (corrected) standard deviation S by

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i \qquad\qquad S = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

Now, the statistic

$$t = \frac{\overline{X} - \mu_0}{S/\sqrt{N}}$$

measures the divergence of the data from the null hypothesis $H_0 : \mu = \mu_0$. Under H_0, it follows Student's t-distribution with $N - 1$ degrees of freedom. This allows us to calculate the p-value: the probability that, if H_0 were true, t would display an even higher divergence from μ_0 than the actually observed discrepancy.

$$p := p_{H_0}(|t(X)| \geq |t|).$$

In this case, we conduct a two-tailed test and consider divergences in both directions from $\mu = \mu_0$. $p < .05$ is commonly interpreted as significant evidence against the null hypothesis. Even smaller p-values denote strong ($p < .01$) and very strong ($p < .001$) evidence against the null hypothesis. Standardly, researchers "reject the null" and infer the alternative on the basis of such p-values.

On the face of it, p-values are way more objective measures of evidence than Bayes factors: they do not involve subjective judgments or degrees of belief and they can be calculated straightforwardly from the statistical model and the observed data. Each researcher will obtain the same p-value when performing the same test. However, these impressions are superficial and dissolve upon closer inspection.

First, NHST and p-values take an asymmetrical stance on the hypothesis testing problem. There is no systematic guidance how we should interpret a non-significant result ($p \geq .05$). Statistics textbooks (e.g., Chase and Brown, 2000; Wasserman, 2004) restrict themselves to a purely negative interpretation: $p \geq .05$ means failure to demonstrate a statistically significant phenomenon. The founding father of NHST, R.A. Fisher (1935) even stressed that the only purpose of an experiment is to *disprove* the null hypothesis, and that we cannot say whether the results confirm the null hypothesis.

[3] Data are assumed to be independent and identically distributed (i.i.d.).

Obviously, this leads to problems whenever the null hypothesis is of substantial scientific interest, e.g., independence of two variables in a causal model, the safety of a medical drug, or absence of parapsychological forces (Gallistel, 2009; Wetzels et al., 2009; Morey et al., 2014; Sprenger, 2017).

In other words, phrasing a scientific inference problem in terms of NHST introduces a value judgment by ruling out the possibility of evidence in favor of the null hypothesis. Especially when the topic under investigation is delicate and politically contentious (e.g., the null hypothesis claims that a particular factor does not contribute to global warming), the asymmetry of NHST undermines their value-freedom: there is no way how we could ever quantify the evidence for that claim and inform policymakers accordingly. NHST is tied to a purist falsificationist methodology which fails to be adequate when similarly important hypotheses are pitched against each other.[4]

Second, frequentist inferences are justified by the long-run properties of the procedures that generate them. This makes good experimental design vital for the justification of the entire inference. In particular, experiments are considered to be reliable when the type I error—the probability of observing significant evidence against a true null hypothesis—is bounded at a low level (e.g., $\alpha = .05$), and the power of the experiment to appraise a true alternative is reasonably high (e.g., $1 - \beta = .8$). Power is always relative to a particular, representative effect size. It needs to be decided beforehand which effect sizes are plausible alternatives to the null such that one does not end up with an underpowered experiment and an unreliable inference (Cohen, 1988; Ioannidis, 2005). There is no surrogate for sound individual scientific judgment in this task. Subjective plausibility judgments are not only essential to Bayesian inference: they are part and parcel of NHST, and in fact, any method of scientific inference (for a practitioner's perspective, see Gelman and Hennig, 2017).

Third, p-values are computed relative to a statistical test, the direction of which is a matter of subjective judgment. In a recent experiment, the well-known social psychologist Daryl J. Bem (2011) published a study which presented evidence for various precognitive skills ("extrasensory powers", "*psi*"), including retroactive influence of a future event on an individual's behavior. In a Bernoulli (success/failure) experi-

[4]Sometimes, meta-analysis is supposed to fill this gap, e.g., failure to find significant evidence against the null in a series of experiments counts as evidence for the null. But first, this move does not provide a systematic, principled theory of statistical evidence, and second, it fails to answer the important question how data support the null hypothesis in a single experiment.

ment, Bem tested the null hypothesis $H_0 : \mu = .5$ that success and failure are equally likely against the (one-sided) alternative $H_1 : \mu > .5$ that precognition would lead to higher success than failure rates. In most experiments, the null was rejected at the $\alpha = .05$ level ($p < .05$). However, this finding was sensitive to whether the test was conducted as a two-sided test ($H_1 : \mu \neq .5$) or as a one-sided test where only departures in a specific direction are considered ($H_1 : \mu > .5$). The authors who critically discussed Bem's findings (Rouder and Morey, 2011; Wagenmakers et al., 2011a) insisted that a more stringent two-tailed test be performed, where deviations in both directions would count as evidence against the null. Ultimately, the evaluation of the statistical evidence depends on how plausible it is that only positive departures from chance need to be considered as a serious alternative to the null hypothesis. This case shows that p-values are not as intersubjectively agreed, mechanically reproducible and value-free as a look at statistics textbooks may suggest.[5]

Fourth and last, supporting scientific judgments on the basis of p-values is not straightforward. Critics of Bem's experiment insist that given the extraordinary nature of Bem's theoretical claims, p-values against the null of no precognition have to be much more convincing than the conventional $p < .05$ (Wagenmakers et al., 2011a). Indeed, what counts as substantial evidence against the null seems to be highly context-sensitive. While the psychological community usually conforms to the $p < .05$ criterion, standards are much more demanding in discplines such as particle physics where a divergence of five standard deviations ($p \approx \mathcal{O}(10^{-6})$) would be required before considered evidence for a major finding, such as the recent discovery of the Higgs Boson. It may be argued that this is just a contingent sociological problem, but if so, it is highly persistent!

The more general issue hiding here is the well-known problem of inductive risk in the assessment of scientific theories. Since the works of Rudner (1953) and Hempel (1965) it is known that weighing uncertainties in statistical reasoning involves value judgments—be it by the individual scientist or the scientific community in a field (Levi, 1960; Douglas, 2000). Both Bayesians and frequentists face this problem: translating p-values into a scientifically meaningful conclusion requires no less subjective judgment

[5]Of course, the problem is more general: for both Bayesians and frequentists, the choice of a statistical test demands a lot of subjective judgment. Often, these choices are nontrivial even in simple problems, e.g., in deciding whether to analyze a contingency table with Fisher's exact test, Pearson's χ^2-test or yet another method.

and consideration of context than for Bayes factors.

We have named four factors that impair the objectivity of frequentist inference with NHST: (1) the asymmetrical design of NHST and the impossibility to confirm the null hypothesis, (2) the need for plausibility judgments in power analysis and experimental design, (3+4) the contentious calculation and interpretation of p-values, including the general problem of inductive risk. All three senses of objectivity from Section 5—value-free, procedural and concordant objectivity—are affected by these factors.

It could be argued that my arguments make a good case against the classical NHST method, but fail to hold for more sophisticated forms of frequentism. For instance, Cohen's (1988) power-centered perspective constructs frequentist inference as a decision procedure where it is also possible to (pragmatically) accept the null hypothesis when it fits the data better than the alternative hypothesis of a scientifically meaningful effect. But this is essentially a *decision procedure* for statistical inference and gives up on attempts to quantify *evidence* for the null—which is even conceded by proponents of that paradigm (Machery, 2012, 816–818). The same holds true for the estimation-centered paradigm that proposes to replace NHST by estimation with confidence intervals (Cumming, 2014). Since confidence intervals have a valid pre-experimental, but no valid post-experimental interpretation, the estimation paradigm does not answer the important question of how to quantify statistical evidence for and against the tested null hypothesis (see also Gallistel, 2009; Morey et al., 2014)—whatever its other merits may be.

All in all, Subjective Bayesianism and (NHST) frequentism face similar problems when they are evaluated in terms of concordant, value-free and procedural objectivity. Now I will argue that these criteria are epistemically inert and a poor basis for assessing the objectivity of statistical inference procedures in the first place.

5 Beyond Concordant, Value-Free, and Procedural Objectivity

In this section, I question the epistemic value of concordant, value-free, and procedural objectivity, motivating that we have to consider other senses of objectivity, too.

First, the pursuit of procedural objectivity by means of banning subjective judgment and promoting standardized protocols has often contributed to a mindless use of statistical techniques (Cohen, 1994; Gigerenzer, 2004; Ziliak and McCloskey, 2008):

significance levels expressed by p-values replace proper scientific thinking and lead to the suppression of scientifically valuable, but statistically insignificant research (e.g., Rosenthal, 1979; Ioannidis, 2005). Here is a particularly illuminating quote:

> All psychologists know that statistically significant does not mean plain-English significant, but if one reads the literature, one often discovers that a finding reported in the Results sections studded with asterisks becomes in the Discussion section highly significant or very highly significant, important, big! (Cohen, 1994, 1001)

In other words, procedural objectivity may have its merits in highly policitized or bias-prone areas of research, but at the same time, it tends to promote mechanical, mindless and possibly misleading use of statistical inference procedures. It is therefore highly context-dependent whether procedural objectivity is epistemically beneficial.

This diagnosis has implications for concordant objectivity, too. This sense of objectivity as intersubjectivity is purely factual; it does not make claims to intrinsic epistemic value. After all, scientists often differ in their disciplinary training, experience, or methodological approach, and these differences will justifiably lead them to different assumptions (e.g., prior distributions), and different conclusions. Meta-analysis and evidence aggregation is a more promising place for concordant objectivity: under ideal circumstances, a free exchange of information and argument may lead to individually rational belief states and intersubjective agreement at the same time (e.g., Lehrer and Wagner, 1981). If we want concordant objectivity to be epistemically valuable, it must not be seen as a constraint on data analysis, but act at the level of amalgamating research findings.

Finally, there is value-free objectivity. We have already mentioned the problem of inductive risk. By now, it is commonly accepted that complete value freedom cannot be achieved in scientific reasoning, and statistical reasoning in particular (e.g., Rudner, 1953; Douglas, 2000). As a consequence, Douglas (2004, 2009) proposes to replace value-free objectivity by **detached objectivity**: values may have a place in scientific reasoning as long as they do not *replace* the evidence. This proposal implements the idea that objectivity implies impartiality in a more modest way than complete value freedom, and it explains why we prioritize evidence over values in forming a judgment on scientific hypotheses. With respect to detached objectivity, however, the criticism of Subjective Bayesianism loses much of its sting: the shape of the prior distribution

affects the outcomes of a Bayesian analysis, but it is less clear why this is problematic. As long as the prior distribution can be justified by reference to past experience or theoretical considerations, the particular choice of the data analyst does not replace evidence by values and violate the ideal of detached objectivity.

Similarly, Subjective Bayesianism fits the bill with respect to Douglas' **value-neutral objectivity**, which means "taking a position that is balanced or neutral with respect to a spectrum of values" and avoiding positions that "are more extreme than they are supportable" (Douglas, 2004, 460). For sure, prior probabilities can express extreme positions, but criticizing and varying them is a routine part of Bayesian inference (more on this will be said in Section 6). If value-neutral objectivity fails in practice, it is not because Subjective Bayesianism is methodologically flawed, but because its methods are abused—just as NHST and p-values are often abused, too.

Of course, this brief overview cannot replace a thorough discussion of scientific objectivity (e.g., McMullin, 1982; Megill, 1994; Longino, 1990; Lacey, 1999; Douglas, 2009; Reiss and Sprenger, 2014). Here, I only wanted to motivate why value freedom, concordance and procedural standardization may be epistemically inert senses of objectivity: they do not contribute to the reliability and epistemic authority of scientific research, or at least not as much as it may appear at first sight. Other conceptions, such as detached and value-neutral objectivity, appear more reasonable, but then it is less clear why Subjective Bayesianism struggles to meet these ideals.

I will now introduce two other senses of scientific objectivity and show that on these counts, subjective Bayesian inference outperforms the rivalling frequentist framework. A case study—Bem's experiment on extrasensory powers—shall buttress my claims.

6 Interactive and Convergent Objectivity

This section moves our discussion of scientific objectivity from the level of individual reasoning to the social aspects of knowledge production. A prominent sense of objectivity in that domain is Helen Longino's

Interactive Objectivity "A method of inquiry is objective to the degree that it permits transformative criticism" (Longino, 1990, 76). This includes, among others, the existence of (1) avenues for criticism of the obtained results; (2) shared standards

for assessing theories; and (3) equality of intellectual authority among qualified practitioners (see also Harding, 1991; Douglas, 2004).

In the first place, the concept of interactive objectivity aims at the social structures that regulate and facilitate scientific communication. That is, science must be structured in such a way that arguments can flow freely, that constructive criticism is possible, and that discussions are not stifled by the power and authority of a particular subgroup in the scientific community.

Interactive objectivity can be applied at the level of statistical inference, too. Subjective Bayesianism promotes interactive objectivity because it makes crucial assumptions behind a statistical inference *transparent*, such as structural assumptions on the unknown parameter, or expectations on the observed effect size. This move opens exactly those avenues for mutual criticism that are demanded from objective science (cf. Gelman and Hennig, 2017). In frequentist inference, however, such assumptions are often hidden behind the curtain (see Section 4).

The debate about the Bem (2011) study on precognition, mentioned in Section 4, illustrates these abstract considerations. Bem conducted a series of nine similar experiments that tested the null hypothesis of no precognition. In one experiment, participants were asked which of two pictures on a computer screen they liked better. Later, the computer would randomly select one of the pictures as the "target" that would be displayed subliminally during the rest of the experiment. More often than chance would allow for, participants preferred the target which would later be selected by the computer. Also the other experiments tested for the existence of retroactive influence patterns with binary response variables (success/failure).

Bem's study stirred a great deal of controversy. There were doubts about the interpretability of some of Bem's experiments (e.g., Rouder and Morey, 2011), and indications for questionable research practices, such as selective reporting and selling exploratory as confirmatory research (Wagenmakers et al., 2011a; Francis, 2012). Moreover, subsequent experiments failed to replicate Bem's effects (Galak et al., 2012). For the sake of the argument, we will not belabor these points, assume that the experiments have been conducted and reported in an orderly fashion and focus on the statistical analysis of the data. Bem pitched the null hypothesis $H_0 : \mu = .5$ (=no precognition, participants are guessing) against the alternative $H_1 : \mu > .5$ (=participants do systematically better than guessing). Using a one-sided t-test as a large sample ap-

proximation of testing the mean of a Binomial distribution, Bem observed statistically significant evidence against the null hypothesis ($p < .05$) in eight out of nine experiments. The mean effect size over all experiments was $\delta = 0.22$, a small to moderate value. Overall, the results were supposed to indicate strong evidence against the null hypothesis of no precognition.

Wagenmakers et al. (2011a,b) conducted a Bayesian re-analysis of Bem's original data. In their critique of Bem's original data analysis, Wagenmakers et al. used a hierarchical Bayesian model where uncertainty about the true effect size δ is described by a Normal distribution $N(\mu, \sigma^2)$, centered around zero ($\mu = 0$) and with unknown variance $\sigma_\delta^2 \sim 1/\chi^2(1)$. When integrating out the influence of the variance σ^2, the prior for the effect size δ follows a Cauchy distribution with probability density f_r given below. The slope of the distribution is described by a scale parameter r (Rouder et al., 2009).

$$f_r(\delta) = \frac{1}{\pi r} \cdot \frac{r^2}{r^2 + \delta^2}$$

Using the default choice $r = 1$, whose theoretical motivation goes back to Harold Jeffreys (1961), Wagenmakers and colleagues obtained a specific prior distribution for δ: $f_1(\delta) = 1/(\pi \cdot (1 + \delta^2))$.[6] With that distribution, they performed a two-tailed Bayesian t-test and concluded that a majority of experiments report evidence in favor of the null hypothesis. Only one experiment provides moderate evidence for the alternative hypothesis H_1 ($10 > BF_{10} > 3$—see Table 1 and 2), four further experiments report anecdotal evidence in favor of H_0 ($3 > BF_{10} > 1$) or no evidence at all ($BF_{10} \approx 1$). The other experiments support H_0 to various degrees, leading to the overall conclusion that the evidence did not favor H_1.

In their response to the paper by Wagenmakers and colleagues, Bem et al. (2011) argued that this Bayesian analysis was based on prior distributions which are not suitable for analyzing data in parapsychological research. They noted that effect sizes in psychology are typically small to moderate ($\delta \approx 0.25$, Bornstein, 1989; Richard et al., 2003), and even smaller for parapsychological effects. This does not square well with the assumptions of Wagenmakers et al. that under H_1, we believe with probability .43 that the absolute value of the effect will be greater than .8. If we could rationally

[6] The use of default priors in Bayesian inference raises a number of interesting philosophical questions (e.g., Sprenger, 2012) which go beyond the scope of this paper. That said, for the given (Binomial) dataset, the chosen approach looks adequate.

Experiment No.	1	2	3	4	5	6(a+b)	7	8	9
BF_{10}, Wagenmakers et al.	1.64	1.05	1.82	0.58	0.88	1.10	0.13	0.47	5.88
BF_{10}, Bem, Utts and Johnson	4.94	3.45	5.35	1.76	2.74	3.78	0.50	1.62	1012

Table 2: The Bayes factor BF_{10} for the alternative hypothesis $H_1 : \mu \neq .5$ vs. the null hypothesis $H_0 : \mu = .5$ in a two-tailed test, according to the Cauchy prior of Wagenmakers et al. (2011a) and the knowledge-based prior of Bem et al. (2011).

expect to observe such large effect sizes, there would be no debate about the reality of parapsychological phenomena.

The same argument is applied to argue against the use of Cauchy priors with their relatively thick, diffuse tails (see Figure 1). Placing any substantial probability on really large effects seems to be plainly inconsistent with the inconclusive and disputed evidence from parapsychological experiments (Hyman and Honorton, 1986; Utts, 1991; Storm et al., 2010). Yet, the prior distribution which is chosen by Wagenmakers and colleagues place a 6% chance on effects greater than 10. Also this is clearly an unrealistic assumption. Choosing wide priors tends to favor the null hypothesis since the alternative hypothesis contains lots of extreme hypotheses which have, under spread-out priors, a substantial weight in the calculation of the Bayes factor.

Instead, Bem et al. (2011, 717–718) advocate a "knowledge-based prior" on δ which differs in two crucial respects: (1) it is based on a Normal instead of a Cauchy distribution, leading to flat tails and highly improbable observations of large effects; (2) the variance σ^2 is chosen such that one's degree of belief in $\delta \in [-0.5, 0.5]$ is equal to 90%, provided there is an effect at all. Figure 1 plots this prior distribution against the prior used by Wagenmakers and colleagues. Using their "knowledge-based prior", Bem, Utts and Johnson obtain moderate evidence in favor of the alternative hypothesis in most experiments. Multiplying the Bayes factors from the individual experiments, they argue that the total picture provides overwhelming evidence against the null hypothesis.

Unsurprisingly, Wagenmakers and colleagues fail to be convinced. They object to the multiplication of Bayes factors across experiments. Moreover, they dispute the assumption that effect sizes should be that small, citing a survey of published articles in *Psychonomic Bulletin & Review* and *Journal of Experimental Psychology: Learning, Memory and Cognition* (Wetzels and Wagenmakers, 2012). These data suggest that substantial effects are as likely as small effects.

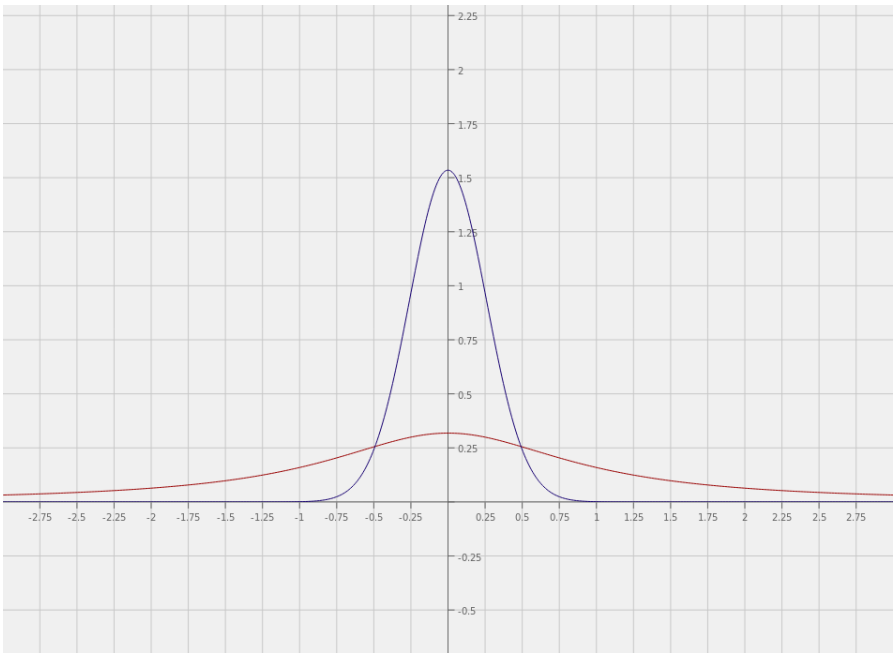

Figure 1: The default Cauchy prior ($r = 1$) advocated by Rouder et al. (2009) and used by Wagenmakers et al. (2011a) versus the "knowledge-based prior" based on the Normal distribution by Bem et al. (2011).

The point of reporting this debate is not to argue that either party is right or wrong. Rather, I would like to illuminate how the subjective Bayesian framework enhances transparency in statistical reasoning, and how it facilitates reasoning and argumentation about the assumptions involved. Phrasing their assumptions in terms of a prior distribution enables two parties to trace the roots of their disagreement in a principled way, to identify the scientific propositions on which they disagree, and to correct potential errors. In the Bem case, the disagreement is primarily fed by three factors: (1) the technical question whether Bayes factors can be multiplied across experiments; (2) the mixed evidence about the distribution of effect sizes in psychology; (3) the methodological question whether the specific nature of parapsychological experiments should be taken into account when shaping the alternative hypothesis.

It is not easy to imagine a similarly productive discourse in frequentist statistics, where a mechanistic interpretation of p-values prevails, and no (probabilistic) judgments about the plausibility of specific effect sizes are allowed. Subjective Bayesian inference does a much better job at enabling readers, stakeholders policy makers to form their own opinion about the implications of an experiment: by making prior distributions transparent, they can decide to what extent they agree with the conclusions. Both sides in the debate agree, incidentally, that this transparency is a great advantage of Bayesian statistics, both epistemically and socially, and they also confirm that subjective assumptions are inevitable in scientific modeling (Bem et al. (2011, 718–719) and Wagenmakers et al. (2011b, 11–12)).

I would also like to point out a particular technique used by Wagenmakers and colleagues: *robustness analysis*, that is, assessing the sensitivity of the conclusions with respect to the prior assumptions. Since any Bayesian inference relies on a prior probability distribution, subjective Bayesians usually try to secure their evidential claims by showing that it is invariant under a variety of prior distributions. This technique is supposed to dispel worries that the final result is the product of idiosyncratic or extreme assumptions. Regulatory bodies even regard robustness analysis as an essential part of Bayesian reasoning:

> We recommend you be prepared to clinically and statistically justify your choices of prior information. In addition, we recommend that you perform sensitivity analysis to check the robustness of your models to different choices of prior distributions. (US Food and Drug Administration, 2010)

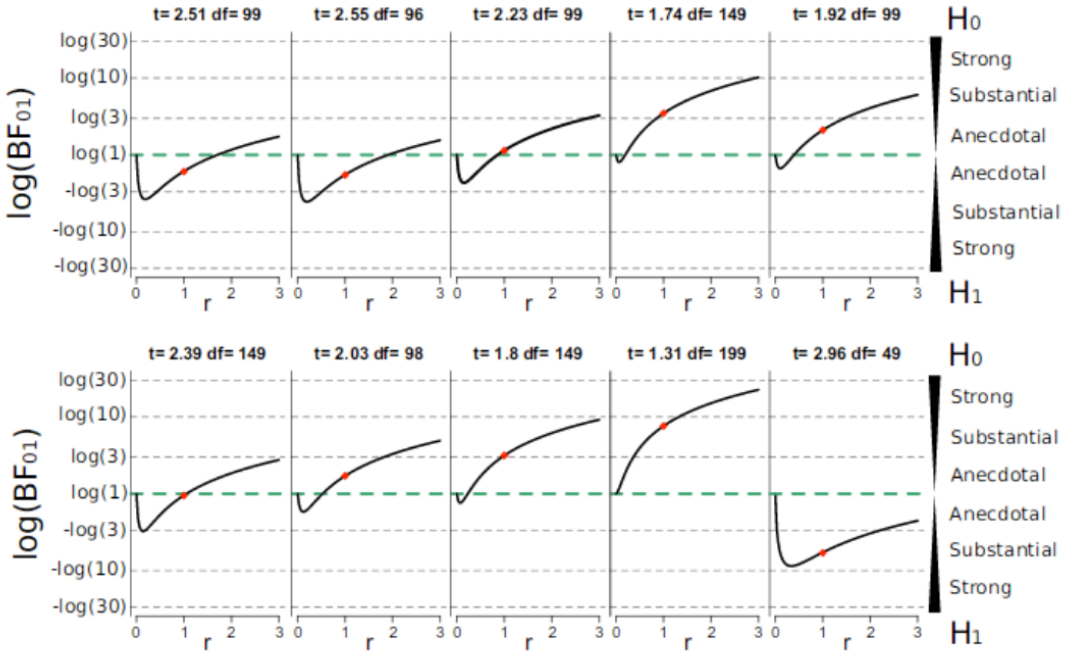

Figure 2: The robustness analysis of Wagenmakers et al. (2011a) regarding the Bayes factors for the null hypothesis in Bem's (2011) experiments. Experiment 6 was split into two datasets.

Checking the robustness of a Bayesian inference with respect to the priors is built into the epistemic framework of Subjective Bayesianism, up to the point that it is a standard option in Bayesian statistics packages (e.g., JASP). Given the contentious nature of expectations about effect sizes in parapsychology, Wagenmakers and colleagues conducted a robustness analysis that varied the value of the scale parameter r. The results are taken from the online appendix of Wagenmakers et al. (2011a) and reproduced in Figure 2.

From these figures it becomes clear that only for a single experiment—Experiment #9, reproduced in the bottom right corner—, there is stable evidence in favor of H_1. By contrast, most experiments show that there is stable evidence in favor of H_0, independent of the choice of r. Only for very specific choices of r, the overall picture favors H_1. Also, they point out that the choice of the scale parameter in Bem et al.'s

knowledge-based prior clearly favors the alternative: almost all other values of that parameter would, within the Gaussian model chosen by Bem et al. (2011), lead to a more favorable assessment of the null hypothesis. These analyses suggest that the evidence claim in favor of H_0 is more robust, and therefore more stable, than Bem's claim in favor of H_1.

Robustness analysis contributes to the assessment of the resilience of an evidential claim, and it is a good check against violations of value-neutral objectivity (Section 5). Evidence claims that require very specific, and possibly extreme prior assumptions, such as in the case of Bem (2011) and Bem et al. (2011), hardly qualify for the "balanced judgment" that distinguishes value-free objectivity. That said, robustness analysis also matches another important sense of scientific objectivity (Douglas, 2004, 2009):

Convergent Objectivity A scientific result is objective to the extent that it is validated from independent assumptions and perspectives. Stability of a result under those variations increases confidence in its reliability.

Originally, this definition was meant to apply in a wide sense, e.g., to invariance of a phenomenon or inference under different experimental designs and theoretical models. For example, Scheele, Priestley and Lavoisier independently conducted groundbreaking work that led to the recognition of oxygen as a chemical element, and the rejection of the phlogiston theory. But there is no reason why this concept should not encompass sensitivity analysis with respect to Bayesian priors, too, and we have seen how robustness considerations in Subjective Bayesianism, and their entrenchment in regulatory constraints, foster the pursuit of convergent objectivity.

Of course, robustness analysis in statistics is more general than this particular application (e.g., Huber, 2009; Staley, 2012), but frequentist theory mostly deals with *distributional robustness*: deviations from the assumed sampling distribution, such as violations of Normality or homoscedasticity. It does not involve robustness with respect to expectations on the size of the observed effect (though see Mayo and Spanos, 2006).

All in all, Subjective Bayesianism promotes objectivity in various ways, especially in the social dimension of scientific inquiry. By adopting the ideal of robustness to variations in the priors, subjective Bayesian inference secures the stability of evidential claims and contributes to convergent as well as value-neutral objectivity. By means of transparent prior distributions, it opens up avenues for criticism by scientific peers and

contributes to interactive objectivity. Arguments for or against specific priors have to be justified by theoretical considerations or past data, not by authority or fiat. Choices are only as good and reasonable as the arguments that support them. These aspects of Subjective Bayesianism facilitate the "transformative criticism" (Longino, 1990, 76) that distinguishes interactive objectivity in the production of scientific knowledge.

7 Conclusions

This paper has argued that charging subjective Bayesianism with lack of objectivity is based on a misunderstanding. The counterargument was based on a threefold strategy. First, I have shown that frequentist (NHST) inference suffers from the very same problems as Subjective Bayesianism, albeit in a more hidden way (Section 4). Second, the senses of objectivity that suggest a critical judgment on Subjective Bayesianism (concordance, value freedom, procedural uniformity) have dubious epistemic value. Realizing this point makes us less confident that subjective Bayesian inference undermines the epistemic authority of science (Section 5). Finally, Subjective Bayesianism promotes two epistemically and socially relevant senses of scientific objectivity, namely interactive and convergent objectivity (Section 6). Subjective Bayesianism makes robustness analysis an integral part of assessing evidential claims, it increases the transparency of statistical reasoning, and it facilitates a critical discussion of crucial modeling assumptions, such as the shape of the prior distribution.

Obviously, the investigations in this paper should be extended to other schools of statistical inference, such as Objective Bayesianism, and other varieties of frequentist inference. For the time being, the present research supports the conclusion that Subjective Bayesianism is surprisingly objective (in the senses specified above), and that it compares favorably to standard frequentist inference.

References

Bem, D. J. (2011). Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, 100(3):407–425.

Bem, D. J., Utts, J., and Johnson, W. O. (2011). Must psychologists change the way they analyze their data? *Journal of Personality and Social Psychology*, 101(4):716–719.

Bernardo, J. M. (2012). Integrated objective Bayesian estimation and hypothesis testing. In *Bayesian Statistics 9: Proceedings of the Ninth Valencia Meeting*, pages 1–68 (with discussion). Oxford University Press, Oxford.

Blackwell, D. and Dubins, L. (1962). Merging of Opinions with Increasing Information. *The Annals of Mathematical Statistics*, 33(3):882–886.

Bornstein, R. (1989). Exposure and affect: Overview and meta-analysis of research, 1968–1987. *Psychological Bulletin*, 106:265–289.

Chase, W. and Brown, F. (2000). *General Statistics*. Wiley, New York.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence & Erlbaum, Newark/NJ.

Cohen, J. (1994). The Earth is Round (p<.05). *Psychological Review*, 49:997–1001.

Cox, D. and Mayo, D. G. (2010). Objectivity and Conditionality in Frequentist Inference. In Mayo, D. G. and Spanos, A., editors, *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability and the Objectivity and Rationality of Science*, chapter 2, pages 276–304. Cambridge University Press, Cambridge.

Crupi, V. (2013). Confirmation. In *The Stanford Encyclopedia of Philosophy*.

Cumming, G. (2014). The New Statistics: Why and How. *Psychological Science*, 25:7–29.

Douglas, H. (2000). Inductive Risk and Values in Science. *Philosophy of Science*, 67:559–579.

Douglas, H. (2004). The irreducible complexity of objectivity. *Synthese*, 138(3):453–473.

Douglas, H. (2009). *Science, Policy, and the Value-Free Ideal*. Pittsburgh University Press, Pittsburgh.

Earman, J. (1992). *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. MIT Press, Cambridge, Mass.

Fisher, R. A. (1935). *The design of experiments.* Oliver & Boyd, Edinburgh.

Fisher, R. A. (1956). *Statistical methods and scientific inference.* Hafner, New York.

Fitelson, B. (2001). *Studies in Bayesian Confirmation Theory*. PhD thesis, University of Wisconsin–Madison.

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic B*, 19:975–991.

Gaifman, H. and Snir, M. (1982). Probabilities Over Rich Languages, Testing and Randomness. *The Journal of Symbolic Logic*, 47(3):495–548.

Galak, J., LeBoeuf, R. A., Nelson, L. D., and Simmons, J. P. (2012). Correcting the Past: Failures to Replicate Ψ. *Journal of Personality and Social Psychology*, 103:933–948.

Gallistel, C. R. (2009). The importance of proving the null. *Psychological review*, 116(2):439–453.

Gelman, A. and Hennig, C. (2017). Beyond objective and subjective in statistics. *Journal of the Royal Statistical Society, Series A*.

Gigerenzer, G. (2004). Mindless Statistics. *Journal of Socio-Economics*, 33:587–606.

Goodman, S. (1999). Toward Evidence-Based Medical Statistics. 2: The Bayes factor. *Annals of Internal Medicine*, 130:1005–1013.

Harding, S. (1991). *Whose Science? Whose Knowledge? Thinking from Women's Lives*. Cornell University Press, Ithaca.

Hempel, C. G. (1965). Science and human values. In *Aspects of Scientific Explanation*. The Free Press, New York.

Howson, C. and Urbach, P. (2006). *Scientific Reasoning: The Bayesian Approach*. Open Court, La Salle, IL, 3rd edition.

Huber, P. J. (2009). *Robust Statistics*. Wiley, New York, 2nd edition.

Hyman, R. and Honorton, C. (1986). A joint communiqué: The psi ganzfeld controversy. *Journal of Parapsychology*, 50:351–364.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2:e124.

Jaynes, E. T. (1968). Prior Probabilities. In *IEEE Transactions on Systems Science and Cybernetics (SSC-4)*, pages 227–241.

Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press, Oxford, 3rd editio edition.

Kass, R. E. and Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90:773–795.

Lacey, H. (1999). *Is Science Value Free? Values and Scientific Understanding*. Routledge, London.

Lee, M. D. and Wagenmakers, E.-J. (2013). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press, Cambridge.

Lehrer, K. and Wagner, C. (1981). *Rational Consensus in Science and Society*. Reidel, Dordrecht.

Levi, I. (1960). Must the Scientist Make Value Judgments? *Journal of Philosophy*, 11:345–357.

Longino, H. (1990). *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton University Press, Princeton, NJ.

Machery, E. (2012). Power and Negative Results. *Philosophy of Science*, 79:808–820.

Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. University of Chicago Press, Chicago.

Mayo, D. G. and Spanos, A. (2006). Severe Testing as a Basic Concept in a Neyman-Pearson Philosophy of Induction. *British Journal for the Philosophy of Science*, 57:323–357.

McMullin, E. (1982). Values in Science. In *Proceedings of the Biennal Meeting of the PSA*, pages 3–28.

Megill, A., editor (1994). *Rethinking Objectivity*. Duke University Press, Durham & London.

Morey, R. D., Rouder, J. N., Verhagen, J., and Wagenmakers, E.-J. (2014). Why hypothesis tests are essential for psychological science: a comment on Cumming (2014). *Psychological science*, 25(6):1289–1290.

Moyé, L. A. (2008). Bayesians in clinical trials: Asleep at the switch. *Statistics in Medicine*, 27:469–482.

Neyman, J. and Pearson, E. S. (1967). *Joint Statistical Papers*. University of California Press, Berkeley/CA.

Popper, K. R. (2002). *The Logic of Scientific Discovery*. Routledge, London. Reprint of the revised English 1959 edition. Originally published in German in 1934 as "Logik der Forschung".

Porter, T. (1996). *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton University Press, Princeton.

Quine, W. V. O. (1992). *Pursuit of Truth*. Harvard University Press, Cambridge MA.

Reiss, J. and Sprenger, J. (2014). Scientific Objectivity. In *The Stanford Encyclopedia of Philosophy*.

Richard, F. D., Bond, C. F. J., and Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, 7:331–363.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3):638–641.

Rouder, J. N. and Morey (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review*, 18:682–689.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., and Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16(2):225–237.

Royall, R. (1997). *Statistical Evidence: A Likelihood Paradigm*. Chapman & Hall, London.

Rudner, R. (1953). The scientist qua scientist makes value judgments. *Philosophy of Science*, 30:1–6.

Senn, S. (2011). You may believe you are a Bayesian but you are probably wrong. *Rationality, Markets and Morals*, 2:48–66.

Smith, A. (1986). Why isn't everyone a Bayesian? Comment. *American Statistician*, 40:10.

Sprenger, J. (2012). The Renegade Subjectivist : José Bernardo's Reference Bayesianism. *Rationality, Markets and Morals*, 3:1–13.

Sprenger, J. (2017). Two impossibility results for Popperian corroboration. *British Journal for the Philosophy of Science*.

Staley, K. (2012). Strategies for securing evidence through model criticism. *European Journal for Philosophy of Science*, 2:21–43.

Storm, L., Tressoldi, P., and Di Risio, L. (2010). Meta-analysis of free response studies 1992–2008: Assessing the noise reduction model in parapsychology. *Psychological Bulletin*, 136:471–485.

US Food and Drug Administration (2010). Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials.

Utts, J. (1991). Replication and Meta-Analysis in Parapsychology. *Statistical Science*, 6:363–403 (with discussion).

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., and van der Maas, H. L. J. (2011a). Why Psychologists Must Change the Way They Analyze Their Data: The Case of Psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, 100(3):426–432.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., and van der Maas, H. L. J. (2011b). Yes, Psychologists Must Change the Way They Analyze Their Data: Clarifications for Bem, Utts and Johnson (2011).

Wasserman, L. (2004). *All of Statistics*. Springer, New York.

Wetzels, R., Raaijmakers, J. G. W., Jakab, E., and Wagenmakers, E.-J. (2009). How to quantify support for and against the null hypothesis: a flexible WinBUGS implementation of a default Bayesian t test. *Psychonomic Bulletin & Review*, 16:752–760.

Wetzels, R. and Wagenmakers, E.-J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review*, pages 1057–1064.

Williamson, J. (2010). *In Defence of Objective Bayesianism*. Oxford University Press, Oxford.

Ziliak, S. and McCloskey, D. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. University of Michigan Press, Ann Arbor.