A Skeptical Eye on Psi

Chapter · January 2016

CITATIONS
3

READS
2,920

5 authors, including:

Denny Borsboom
University of Amsterdam
171 PUBLICATIONS **4,629** CITATIONS
SEE PROFILE

Rogier A. Kievit
MRC Cognition and Brain Sciences Unit
55 PUBLICATIONS **724** CITATIONS
SEE PROFILE

Han L J van der Maas
University of Amsterdam
152 PUBLICATIONS **3,370** CITATIONS
SEE PROFILE

Eric-Jan Wagenmakers
University of Amsterdam
193 PUBLICATIONS **7,319** CITATIONS
SEE PROFILE

A Skeptical Eye on Psi

Eric–Jan Wagenmakers[1], Ruud Wetzels[1], Denny Borsboom[1], Rogier Kievit[2], & Han L. J. van der Maas[1]
1 University of Amsterdam
2 Cambridge University

> "I assume that the reader is familiar with the idea of extra-sensory perception, and the meaning of the four items of it, viz. telepathy, clairvoyance, precognition and psycho-kinesis. These disturbing phenomena seem to deny all our usual scientific ideas. How we should like to discredit them! Unfortunately the statistical evidence, at least for telepathy, is overwhelming." (Turing, 1950, p. 453).

Research on extra-sensory perception or *psi* is contentious and highly polarized. On the one hand, its proponents believe that evidence for psi is overwhelming, and they support their case with a seemingly impressive series of experiments and meta-analyses. On the other hand, psi skeptics believe that the phenomenon does not exist, and the claimed statistical support is entirely spurious. We are firmly in the camp of the skeptics. However, the main goal of this chapter is not to single out and critique individual experiments on psi. Instead, we wish to highlight the many positive consequences that psi research has had on more traditional empirical research, an influence that we hope and expect will continue in the future.

In the first section below, we assume that psi does not exist. Under this assumption, the entire field is spurious, and the literature on psi is a perfect reflection of what happens when a large group of researchers fool themselves into believing that an imaginary phenomenon is real. Thus, each and every effect is a fluke; all meta-analyses reflect bias, deception, or even outright fraud; none of the experiments are replicable; no practical application is ever successful. Such a state of affairs resembles the fictitious situation on planet F345 in the Andromeda galaxy in the year 3045268, where an intelligent humanoid species has developed an unfortunate preoccupation with *null fields*, research fields that have no basis in reality; hence "(...) whatever claims for discovery are made are mostly just the result of random error, bias, or both. The produced discoveries are just estimating the net bias operating in each of these null fields. Examples of such null fields are nutribogus

---

A small portion of the content of this chapter is taken from an online, unpublished rejoinder to Bem, Utts, and Johnson (2011). This work was supported by an ERC grant from the European Research Council. Correspondence concerning this article may be addressed to Eric-Jan Wagenmakers, University of Amsterdam, Department of Psychology, Weesperplein 4, 1018 XA Amsterdam, the Netherlands. Email address: EJ.Wagenmakers@gmail.com.

epidemiology, pompompomics, social psychojunkology (...) Unfortunately, F345 scientists do not know that these are null fields and don't even suspect that they are wasting their effort and their lives in these scientific bubbles." (Ioannidis, 2012, p. 647).

Research on psi saves us a trip to Ioannidis' planet F345, as the effects of researcher bias can be gauged simply by studying the earthly literature on psi. In a sense, research on psi can be considered the drosophila for questionable research practices and the deleterious effects of hindsight bias and confirmation bias. In other words, if a particular research methodology (e.g., meta-analysis on a series of studies mostly conducted by proponents) reliably shows the existence of psi, this means that the same methodology cannot be safely used in other fields to demonstrate the existence or impact of a particular phenomenon. Such methodology can be said not to be *psi-resistent*. In addition, we can propose and develop methods that are able to expose psi findings as false, and apply these psi-resistent methods in traditional research fields as well. In short, psi research is an excellent *control condition for science.*[1]

In the second section below, we assume that psi may exist, although the odds in favor of its existence are slim. We outline the principles of Bayesian belief updating, and show its consequences for research on psi and other implausible phenomena. In the third section below, we present a case study and present a critical analysis of the recent work by Daryl Bem, who arguably tried to bamboozle readers and reviewers by presenting nine precognition experiments with over 1,000 participants. We conclude by outlining a five-step program that proponents of psi —or any other implausible empirical phenomenon— can follow in order to begin to convince skeptics of their position.

Part I. What Can We Learn From Research on Psi, Assuming Psi Does Not Exist?

In 1650, Oliver Cromwell tried to convince the General Assembly of the Church of Scotland that their support for Charles II was misguided: "I beseech you, in the bowels of Christ, think it possible you may be mistaken."[2] Cromwell's plea formed the basis of what Bayesian statistician Dennis Lindley called *Cromwell's Rule*, the principle that no proposition, however implausible, should be assigned zero probability: "In other words, if a decision-maker thinks something cannot be true and interprets this to mean it has zero probability, he will never be influenced by any data, which is surely absurd. So leave a little probability for the moon being made of green cheese; it can be as small as 1 in a million, but have it there since otherwise an army of astronauts returning with samples of the said cheese will leave you unmoved." (Lindley, 1991, p. 104)

When it comes to research on psi, Cromwell's Rule is a two-sided sword: psi skeptics should reserve a little probability that the phenomenon is true, but psi proponents should reserve a little probability that the phenomenon is false. Here we swing the sword in the direction of the proponents, and we beseech them to think it possible that they are mistaken, and psi does not exist. Under this assumption, psi is a control condition for science, an

[1] This idea has been proposed earlier, for instance by Diaconis (1991), by Eliezer Yudkowsky at `http://lesswrong.com/lw/1gc/frequentist_statistics_are_frequently_subjective/`, by Allan Crossman at `http://lesswrong.com/lw/1ib/parapsychology_the_control_group_for_science/`, and by Scott Alexander at `http://slatestarcodex.com/2014/04/28/the-control-group-is-out-of-control/`.

[2] The Scots were not convinced and answered: "would you have us to be sceptics in our religion?"

unwitting jester in the court of academia. A jester, however, does serve several important functions: he is supposed to critique his masters and their guests, and he is often the first to deliver unwelcome news.[3]

Fully consistent with its role as court jester, research on psi has contributed greatly to the recent surge of much-needed doubt and self-reflection that has rippled through the empirical sciences. This "crisis of confidence" originated from a combination of events, but one the most important was the 2011 publication by Daryl Bem, who managed to have a flagship academic journal –the *Journal of Personality and Social Psychology*– accept an article claiming that people can look into the future (Bem, 2011; Miller, 2011). Another important contributing event was the fraud case of Diederik Stapel, a Dutch social psychologist who, for many years until the denouement in 2011, was able to fabricate data on a massive scale (e.g., Stroebe, Postmes, & Spears, 2012). These two major events were shortly followed by others; for instance, Simmons, Nelson, and Simonsohn (2011) reminded researchers how significant results can be obtained by "p-hacking": a combination of questionable research practices (QRPs) that can be executed with the sole goal to pass the standard .05 threshold of statistical significance (see also De Groot, 1956/2014; John, Loewenstein, & Prelec, 2012; Kerr, 1998; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011).

More recently, it has become clear that many empirical results cannot be replicated when QRPs have been eliminated by preregistering the data analysis plan in advance of data collection. Prominent failures to replicate include preclinical findings in cancer biology (e.g., Begley & Ellis, 2012), findings on social priming (e.g., Donnellan, Lucas, & Cesario, in press; Doyen, Klein, Pichon, & Cleeremans, 2012; Harris, Coburn, Rohrer, & Pashler, 2013; LeBel & Wilbur, in press; LeBel & Campbell, in press; Pashler, Rohrer, & Harris, 2013; Shanks et al., 2013) and in social psychology in general (Nosek & Lakens, 2014). The extent and seriousness of these failures to replicate is still unclear and await a more thorough investigation (e.g., Open Science Collaboration, 2012), but the intermediate results have already caused widespread concern.

Thus, substantial credit for the current "crisis of confidence" goes to psi researchers. It is their work that has helped convince other researchers that the academic system is broken, for if our standard scientific methods allow one to prove the impossible, than these methods are surely up for revision. Partly because of research on psi, there is now a deeper appreciation for the fact that science is not necessarily self-correcting; that researchers can insert subtle biases in data analysis, invalidating even the most sophisticated statistical procedures; and that the published literature provides only a contorted reflection of the true state of affairs. Some of these insights are not new, but their widespread appreciation is certainly due in part to psi researchers.

The next section provides a general overview of the different threats to the validity of empirical findings. These threats affect current research practices throughout the empirical sciences (see also Goldacre, 2008). Countering these threats is essential in any academic endeavor, but even more so for research on hypotheses that are highly implausible *a priori*.

[3] For example, in 1340 the English navy destroyed the French fleet at the Battle of Sluys. The French king's jester told him the English sailors "don't even have the guts to jump into the water like our brave French" (Otto, 2001, p. 113), as cited at `http://en.wikipedia.org/wiki/Jester`.

Ten Threats to The Validity of Research Findings

Research findings can be compromised by a variety of mistakes and biases. Below we provide a selective and systematic overview of ten common threats.

1. **Poor design**. When a study is poorly designed, its conclusions are jeopardized. Sometimes poor design completely invalidates the results (e.g., when a critical control condition is omitted, or when the stimulus material is inadequate), and sometimes it only makes the results less compelling than they would otherwise be. A general example of the latter is the failure to counterbalance items or to the failure to consider a within-subject design instead of a between-subject design. For psi, a good design also features a follow-up test using the most successful participants from the first test. If there are inter-individual differences in the sensitivity to psi, participants who score above chance in the first test should also score above chance, on average, in the follow-up test (Lee & Wagenmakers, 2013).

2. **Mistakes in data collection**. These include mistakes in the software programs responsible for computerized testing and experimenter mistakes in assigning participants to conditions. Perhaps the most common mistake in data collection is to allow the presence of demand effects, where the experimenter is aware of the research hypothesis and interacts with the participants in a between-participant design.

3. **Wonky analyses, that is, using the data twice**. In the words of Ben Goldacre: "You cannot find your starting hypothesis in your final results. It makes the stats go all wonky." (Goldacre, 2008, p. 221). This is by far the most dangerous threat to the validity of research findings, and the next section discusses it in some detail.

4. **Publication bias**. Publication bias is related to cherry-picking through multiple testing. The cause is a pervasive "prejudice against accepting the null hypothesis" (Greenwald, 1975). Much has been written about this issue, but the core problem is highlighted by the Randall Munroe cartoon, "significance" or "do jelly beans cause acne?" (`http://xkcd.com/882/`). In other words, the researcher's intention is one of "seek and you will find": either across experiments, leading to publication bias, or within an experiment, leading to multiple testing.

5. **The infamous p-value**.[4] The p-value is the probability of encountering a test statistic as least as extreme as the one that was observed, given that the null hypothesis is true and the sampling is correctly specified. Misinterpretations of the p-value are legion. One of its main structural limitations is that it does not consider how extreme the observed data are under the alternative hypothesis; hence, it is possible to reject the null hypothesis for data that are much more likely under the null hypothesis than under the alternative hypothesis. In an early critique of p-values, (Berkson, 1938, p. 531) stated: "My view is that *there is never any valid reason for rejection of the null hypothesis except on the willingness to embrace an alternative one.* No matter how rare an experience is under a null hypothesis, this does not warrant logically, and in practice we do not allow it, to reject the null hypothesis if, for any reasons, no alternative hypothesis is credible." This comment seems particularly apt with respect to research on psi.[5] For this and other reasons, p-values do not quantify evidence for or against a null hypothesis. Relevant references include Berger

[4] Unfortunately, some authors of this chapter are stubborn frequentists who wish to distance themselves from the sweeping negative statements about p-values below.

[5] The extent to which one deems an alternative hypothesis "credible" is subjective. The later section on Bayesian inference discusses the prior plausibility of competing hypotheses in more detail.

and Wolpert (1988), Johnson (2013), Nickerson (2000), Nuzzo (2014), Royall (1997), and Wagenmakers (2007).

6. **BEM: adjusting the narrative to fit the results**. "There are two possible articles you can write: (1) the article you planned to write when you designed your study or (2) the article that makes the most sense now that you have seen the results. They are rarely the same, and the correct answer is (2)." (Bem, 2003). This is what we have termed the Bem Exploration Method (BEM; Wagenmakers et al., 2011), and it should be clear that its use constitutes an important threat to the validity of one's research findings (see also Kerr, 1998).

7. **Meta-analyses on biased studies**. The literature on psi features several meta-analyses (e.g., Bem, Tressoldi, Rabeyron, & Duggan, 2014; Mossbridge et al., 2014; Utts, 1991; Storm, Tressoldi, & Di Risio, 2010, but see Diaconis, 1991; Hyman, 2010; Rouder, Morey, & Province, 2013; Schwarzkopf, 2014). However, meta-analyses conducted on a set of biased studies can only affirm the existence of that bias. "As Henri Poincaré put it: 'The physicist is persuaded that one good measurement is worth many bad ones'." (Jaynes, 2003, p. 258). In the context of psi, Price (1955, p. 367) concluded "(...) the only answer that will impress me is an adequate experiment. Not 1000 experiments with 10 million trials and by 100 separate investigators giving total odds against chance of 10^{1000} to 1—but just one good experiment."[6]

8. **Conceptual replications instead of direct replications**. To protect and solidify the status quo, inoculating it against empirical disconfirmation, there is nothing quite as effective as discouraging direct replications and promoting conceptual replications instead. As argued by Pashler and Harris (2012, p. 533): "The inevitable conclusion, it seems to us, is that a scientific culture and an incentive scheme that promotes and rewards conceptual rather than direct replications amplifies the publication bias problem in a rather insidious fashion. Such a scheme currently exists in every area of research of which we are aware." As outlined by Pashler and Harris, the problem with conceptual replications is that they, in practice, will never be taken as evidence *against* a substantive hypothesis: If the conceptual replication fails, the divergence from the original experiment is considered too extreme. However, if the conceptual replication succeeds, it is taken as strong evidence in favor of the original hypothesis. In the immortal words of Yazz then, for exotic findings, "the only way is up".

9. **Skeptical attitude towards skeptics**. Once a group of researchers has repeatedly published on a particular phenomenon, they tend to view outside criticism with suspicion and sometimes even hostility. Skeptics may be accused of being incompetent, of being unaware of the crucial but hidden "tricks of the trade", of bullying the proponents, and of being overly negative. By downplaying the ability of the skeptics to collect and analyze data competently, and by questioning their motives, the proponents can continue to oppress doubt and act as if the empirical phenomenon of interest are reliable, robust, and reproducible.

[6]The Price article lead to a lively discussion (Meehl & Scriven, 1956; Price, 1956; Rhine, 1956; Soal, 1956). Many years later, Price (1972) issued a brief apology for having suggested that Rhine and Soal engaged in research fraud. Ironically, the current consensus is that Soal did in fact commit research fraud (see `http://en.wikipedia.org/wiki/Samuel_Soal`). For a skeptical review of Rhine's work see `http://en.wikipedia.org/wiki/Joseph_Banks_Rhine`.

10. **Favoritism**. Researcher Y submits an article on phenomenon X for publication in an academic journal. This article is likely to be handled be an action-editor who is familiar with X. This action-editor will in turn select expert reviewers, most of whom also work on X and are likely to know Y personally. The reviewers may like the conclusion of the article because it agrees with their own work. In addition, they realize that another prestigious publication on X will have a positive impact on their own work and career. Consequently, these reviewers are less critical than if they had been confronted with an article questioning the existence or importance of phenomenon X. The same considerations that hold for the review process also hold for the process of grant evaluation.

In recent years, the existence of the above threats has become increasingly clear, thanks in no small part to the persistent efforts of researchers attempting to demonstrate the presence of psi. In the next section we discuss the threat that has received the most attention: the double use of data.

The Danger of Using the Data Twice

It often happens that researchers conduct an experiment without knowing in advance how they will analyze the resulting data *exactly*. For instance, a psi researcher may wish to study whether participants have precognition by making them guess repeatedly where on the computer screen a picture will appear next. The researcher may test male and female participants, and may include neutral pictures, erotic pictures, romantic pictures, and negative pictures as stimulus materials. In addition, the researcher may administer a test that yields an extraversion score for each participant. The general hypothesis is that people have precognition. A more specific hypothesis, perhaps, is that the effect is larger for extravert people. The researcher then analyzes the data and finds that the effect of precognition is statistically significant ($p = .032$), but only for extravert women confronted with erotic pictures, and only after log-transforming the dependent variable and removing participants whose guesses were slower than 3 sec.

In this example, the p-value of .032 is deeply misleading, greatly overstating the case against the null hypothesis. This happens because the p-value hypothesis test is only valid for analyses that were pre-planned and not motivated by the data itself. That is, the p-value of .032 is only statistically valid if the researcher had decided, in advance of data collection, to test precognition for extravert women confronted with erotic pictures, after log-transforming the dependent variable and removing participants whose guesses were slower than 3 sec. But if this was the hypothesis of interest, why bother to test male participants, or participants who score low on extraversion? And why include the neutral pictures, the romantic pictures, and the negative pictures? In this case, the study design reveals the truth: the researcher did not have strong prior expectations at all. Instead, the researcher analyzed the data, found that the effect of precognition was highest for extravert women confronted with erotic pictures (after log-transforming the dependent variable and removing participants whose guesses were slower than 3 sec), and then went on to test this specific hypothesis on the very same data that inspired that hypothesis in the first place. This is statistical heresy.

At the heart of this heresy lies a blurred distinction between exploratory or *hypothesis-generating* research versus confirmatory or *hypothesis-testing* research. This distinction has been made by various authors (including ourselves, Wagenmakers et al., 2011; Wagenmak-

ers, Wetzels, Borsboom, van der Maas, & Kievit, 2012), but here we wish to highlight the early contribution by Adriaan de Groot (e.g., De Groot, 1956/2014, 1969). In his monograph "Methodology", De Groot pointed out that "Exploratory investigations differ from hypothesis testing in that the canon of the inductive method of testing is not observed, at least not in its rigid form. The researcher does take as his starting-point certain expectations, a more or less vague theoretical framework; he is indeed out to find certain kinds of relationships in his data, but these have not been antecedently formulated in the form of precisely stated (testable) hypotheses. Accordingly they cannot, in the strict sense, be put to the test." (De Groot, 1969, p. 306). Consequently, "One 'is allowed' to apply statistical tests in exploratory research, just as long as one realizes that they do not have evidential impact. They have also lost their forceful nature, regardless of their outcomes." (De Groot, 1956/2014).

The core problem, as De Groot explained in an example on psychokinesis, is that exploratory research implies a multiple comparison procedure with an unknown number of comparisons, as it cannot be discerned from the final result how many data patterns the researcher considered before testing the one that ends up reported (De Groot, 1956/2014).[7] Thus, exploratory research allows the researcher many degrees of freedom, but the price that has to be paid for this freedom is that statistical testing has become all but impossible. It should be emphasized that exploratory research is a noble and worthwhile endeavor, and we have nothing against it; the concern is not with the nature of the research –confirmatory or exploratory–, but rather with the lack of a proper distinction between the two. It is dishonest and misleading to conduct exploratory research but carry out a hypothesis test and pretend the result has the same evidential impact as it does for confirmatory work.

Figure 1 provides an illustration of the exploration-confirmation continuum (Wagenmakers et al., 2012). On the far right, research is purely transparent and confirmatory, and the statistical results have their intended meaning. On the far left, the Texas sharpshooter has first fired shots randomly at a fence, and then walked up to draw the targets around his shots. This is an extreme example of exploration and double use of the data, and for such cases the statistical test is misleading and meaningless (or "wonky", in Ben Goldacre's terms). In between the extremes lies a continuum of exploration. Here the data are "massaged" or "tortured" to an extent that is often difficult to discern from the published article. The statistical results will therefore be wonky to an unknown degree.

Despite a researcher's good intentions, it is almost impossible to prevent exploration in data analysis. Who would not be tempted to conduct a hypothesis test on a phenomenon that appears promising? Without conducting the test, how can researchers tell whether or not they are looking at noise? Many researchers may not even realize the distinction between exploratory and confirmatory research, and simply analyze the data with the goal to find interesting patterns, and confirm the reality of these patterns by a statistical test. That such a procedure is misleading is probably not always evident to the researcher. In addition, hindsight is 20/20, and after the data have been looked at it is difficult to resist the myriad of human biases that pull the researcher away from skepticism and towards accepting the hypothesis that was expected, in some form or other, and so clearly exert its effect.

The uncomfortable truth is that, in order to safeguard the purely confirmatory na-

[7] From a Bayesian perspective, the multiple comparisons implicit in exploratory research should be reflected in the prior plausibility of the hypotheses under consideration.

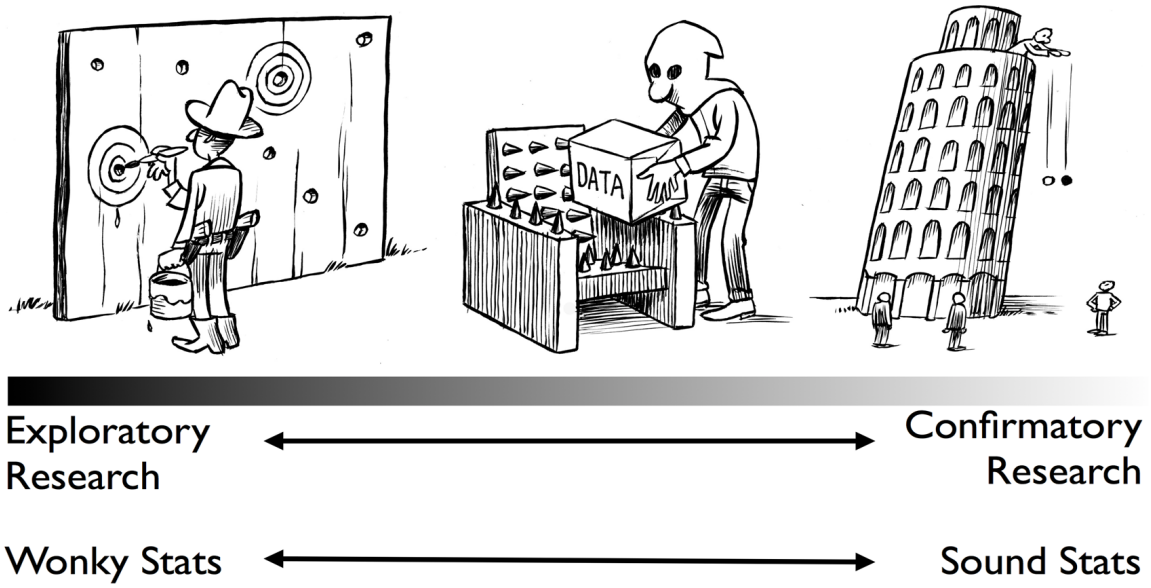

Figure 1. A continuum of experimental exploration and the corresponding continuum of statistical wonkiness (Wagenmakers et al., 2012). "On the far left of the continuum, researchers find their hypothesis in the data by post hoc theorizing, and the corresponding statistics are 'wonky', dramatically overestimating the evidence for the hypothesis. On the far right of the continuum, researchers preregister their studies such that data collection and data analyses leave no room whatsoever for exploration; the corresponding statistics are 'sound' in the sense that they are used for their intended purpose. Much empirical research operates somewhere in between these two extremes, although for any specific study the exact location may be impossible to determine. In the grey area of exploration, data are tortured to some extent, and the corresponding statistics are somewhat wonky." Figure downloaded from Flickr, courtesy of Dirk-Jan Hoek.

ture of the statistical inference, it is crucial to preregister the analysis plan in all of its excruciating details. As mentioned by De Groot (1969, p. 69): "If an investigation into certain consequences of a theory or hypothesis is to be designed as a genuine testing procedure (and not for exploration), a precise *antecedent formulation* must be available, which permits testable consequences to be deduced."

In fact, De Groot believed that an antecedent formulation or preregistration document should adhere to a comprehensive set of guidelines. Specifically, De Groot (1969, p. 136) stated:

> "Foremost (...) is the recommendation to *work out* in advance the investigative procedure (or experimental design) *on paper to the fullest possible extent.* This 'blueprint' should comprise: a brief exposition of the theory; a formulation of the hypothesis to be tested; a precise statement of the deductions leading up to the predictions to be verified; a description of the instruments – in the broadest sense – to be used, complete with instructions for their manipulation; detailed operational definitions of the variables to be used; a statement about the measurement scales (nominal, ordinal, interval, ratio) in which the respec-

> tive variables are to be read (...); a clearly defined statement of the respective universes to which the hypothesis and the concrete prediction(s) apply; an exact description of the manner in which the samples are to be drawn or composed; a statement of the confirmation criteria, including formulation of null hypotheses, if any, choice of statistical test(s), significance level and resulting confirmation intervals (...); for each of the details mentioned, a brief note on their rationale, i.e., a justification of the investigator's particular choices."

In sum, confirmation bias and hindsight bias make it difficult for researchers to respect a clear distinction between exploratory and confirmatory work. If statistical inference is desired, then the associated research must be confirmatory, that is, of the hypothesis-testing type. To guarantee that this is the case, researchers can preregister their analysis plan, for instance on the Open Science Framework (`https://osf.io/`). Due in large part to the efforts of Chris Chambers[8], many academic journals have now adopted a new two-step review format that includes preregistration as a central component (e.g., Chambers, 2013; Wolfe, 2013); in the first step, the researcher submits a preregistration document that contains a motivation for the study and a detailed plan of analysis; in the second step, after the data are collected, the researcher submits the article reporting the results. The central idea of this format is to eliminate all QRPs and researcher degrees of freedom, and obtain an honest impression of the phenomenon of interest. In the context of this chapter, it is relevant to mention that a similar preregistration format was used in the 1980s by the *European Journal of Parapsychology*.[9]

Part II. Does Psi Exist?

Implausible does not mean impossible. The strange, counter-intuitive behavior of the quantum world is but one example of this general rule. Also, even if many researchers believe in the absence/presence of a particular phenomenon or the validity of a particular procedure, this does not necessarily mean that they are right. After all, science is not a democracy.

Nevertheless, even the most vociferous proponent would agree that the hypothesis of psi is highly implausible *a priori*. For example: no psi organ or brain area has yet been discovered; if psi existed, our world would look very different than it does today; no plausible mechanism has been proposed to generate psi; if psi existed, casinos and betting office should not make as much money as they do; if psi existed, why are there no practical applications that make use of this extraordinary human facility? The latter argument is made convincingly by another Randall Munroe cartoon, "the economic argument", shown here as Figure 2.

To set the stage for the discussion below, we need to formalize the way in which data revise the belief of a rational agent. This can be done via Bayesian inference, as follows. Before any empirical data are in, the relatively plausibility of two competing hypotheses is quantified by the *prior odds*, $p(\mathcal{H}_0)/p(\mathcal{H}_1)$. Here, $\mathcal{H}_0$ represents the hypothesis "psi does not exist" and $\mathcal{H}_1$ represent the hypothesis "psi exists". It is evident that the prior odds are

[8] See for instance `http://neurochambers.blogspot.nl/`, `http://www.theguardian.com/science/head-quarters/2014/may/20/psychology-registration-revolution`.

[9] We thank Dick Bierman for pointing this out.

Figure 2. "The economic argument". Reprinted with permission from XKCD (`http://xkcd.com/808/`) under a Creative Commons Attribution-NonCommercial 2.5 License.

highly subjective, and the prior odds for a psi proponent will differ greatly from the prior odds of a psi skeptic. In other words, the prior odds quantify one's initial enthusiasm or skepticism with respect to the phenomenon under consideration. When empirical data D come in, the prior odds are updated to *posterior odds*, $p(\mathcal{H}_0 \mid D)/p(\mathcal{H}_1 \mid D)$, which quantify the relatively plausibility of the two competing hypotheses after taking into account the observed data D. The posterior odds inherit some of the subjectivity of the prior odds, and consequently proponents and skeptics may disagree about the posterior odds almost as much as they disagree about the prior odds.

What skeptics and proponents can agree upon, at least in principle, is the extent to which the observed data change the prior odds to the posterior odds. This change in odds is called the *Bayes factor* (Jeffreys, 1961; Kass & Raftery, 1995):

$$\mathrm{BF}_{01} = \frac{p\left(D \mid \mathcal{H}_0\right)}{p\left(D \mid \mathcal{H}_1\right)}. \tag{1}$$

Thus, if $\mathrm{BF}_{01} = 9$, the observed data are nine times more likely under the null hypothesis $\mathcal{H}_0$ than under the alternative hypothesis $\mathcal{H}_1$; if $\mathrm{BF}_{01} = .2$, the observed data are five times more likely under the alternative hypothesis $\mathcal{H}_1$ than under the null hypothesis $\mathcal{H}_0$. Hence the function of the Bayes factor is to "grades the decisiveness of the evidence" (Jeffreys,

1961) that the data provide for the two competing hypothesis. An in-depth discussion of Bayes factors is beyond the scope of this chapter, but relevant details can be found in Rouder, Speckman, Sun, Morey, and Iverson (2009), Rouder, Morey, Speckman, and Province (2012), Rouder and Morey (2012), Wetzels and Wagenmakers (2012), Wetzels et al. (2011).

The only remaining complication is to specify exactly what we mean by $\mathcal{H}_1$: "psi is present". We could pick a single effect size, say $d = .25$, but this ignores our uncertainty – only in exceptional cases do we have the opportunity and knowledge to specify the alternative hypothesis exactly, as a single point. In realistic applications, our uncertainty about the effect size expected under $\mathcal{H}_1$ is quantified by a *prior distribution*. One popular default prior distributions is similar to a standard Gaussian, centered on zero but with fat tails (i.e., the Cauchy distribution). Another default choice for effect size is the standard normal. In another valuable contribution from psi researchers, Bem et al. (2011) proposed a *psi prior*, a prior on effect size that was deemed suitable for very small effects. This prior is a normal distribution with mean 0 and standard deviation 0.3. The value of the psi prior transcends the field of psi, as it provides a useful lower bound on effect sizes that may be expected under any $\mathcal{H}_1$. It is of course still possible for researchers to argue for effect sizes smaller than psi, but they would have some explaining to do.

In sum, the Bayesian perspective on belief revision can account for the fact that skeptic and proponent have a subjectively different assessment of prior plausibility for psi. The key element, however, is the Bayes factor: the extent to which the observed data change the prior odds to the posterior odds. In other words, skeptic and proponent may not agree on the our evidential starting position, but they can at least agree on the evidential flow, the direction and extent to which the data change one's initial opinion (Jefferys, 1990). Of course, this strategy only applies under the assumption that the data were collected in a confirmatory fashion, and that no deception was present or mistakes were made. Also, the Bayesian perspective is not the only one that allows multiple hypotheses to be pitted against each other, or to confer evidence on the null hypothesis, and various alternative inference schemes are available to establish the objective of weighing the evidence in a sensible fashion.

Extraordinary Claims...

Many philosophers and researchers have considered the case of psi or extra-sensory perception (ESP) as an example of a highly implausible hypothesis. The Scottish philosopher David Hume (1711–1776) stated: "no testimony is sufficient to establish a miracle, unless the testimony be of such a kind, that its falsehood would be more miraculous, than the fact, which it endeavors to establish; and even in that case there is a mutual destruction of arguments, and the superior only gives us an assurance suitable to that degree of force, which remains, after deducting the inferior." The first real Bayesian statistician, Pierre-Simon Laplace (1749–1827), formulated the same sentiment: "The weight of evidence for an extraordinary claim must be proportioned to its strangeness." American astronomer Carl Sagan (1934–1996) coined the famous phrase "extraordinary claims require extraordinary evidence." In terms of Bayesian belief updating, this means that implausible hypotheses have the prior odds stacked against them, and in order to overcome this initial disadvantage they require relatively compelling evidence from the observed data (i.e., convincing

Bayes factors).

Now suppose that a psi enthusiast conducts an experiment and reports a Bayes factor of one billion in favor of $\mathcal{H}_1$: "psi exists" over $\mathcal{H}_0$: "psi does not exist". Should psi skeptics now adjust their prior belief by a factor of one billion? This key question was considered in some detail by Jaynes (2003, pp. 123-124):

> "Probability theory gives us the results of consistent plausible reasoning from the information *that was actually used* in our calculation. It can lead us wildly astray (...) if we fail to use all the information that our common sense tells us is relevant to the question we are asking. When we are dealing with some extremely implausible hypothesis, recognition of a seemingly trivial alternative possibility can make many orders of magnitude difference in the conclusions (...) innocent possibilities, such as unintentional error in the record keeping (...) to deliberate falsification of the whole experiment for wholly reprehensible motives. Let us call them all, simply, 'deception' (...)
>
> Indeed, the very evidence which the ESP'ers throw at us to convince us, has the opposite effect on our state of belief; issuing reports of sensational data defeats its own purpose. For if the prior probability for deception is greater than that of ESP, then the more improbable the alleged data are on the null hypothesis of no deception and no ESP, the more strongly we are led to believe, not in ESP, but in deception. For this reason, the advocates of ESP (or any other marvel) will never succeed in persuading scientists that their phenomenon is real, until they learn how to eliminate the possibility of deception in the mind of the reader (...)[10]

At this point the psi proponent might feel caught between a rock and a hard place: if the evidence from the data is not compelling, then it is dominated by prior skepticism; if the evidence from the data is compelling, then the findings are deemed be the result of deception. Is there no way to win? We believe that, just as with any other highly implausible hypothesis, there are ways for proponents to convince the skeptics. We discuss some of these tactics in a separate section below. First, however, we discuss a famous case study on psi that brings together many of the considerations outlined earlier.

Part III. Bamboozled?

In an article for *Journal of Personality and Social Psychology*, Bem (2011) presented nine experiments that test for the presence of *precognition*, the phenomenon that people's current thinking and behavior is affected by events that occur later the future. Bem argued that in eight out of the nine experiments, the data supported the presence of precognition with one-sided p values smaller than .05.

[10] Jaynes goes on to note that this observation was already made by Laplace in his *Essai Philosophique sur les Probabilités*, when, after calling attention to "the immense weight of testimonies necessary to admit a suspension of natural laws", Laplace observed that those who claim to have witnessed a miracle, "decrease rather than augment the belief which they wish to inspire; for then those recitals render very probable the error or the falsehood of their authors. But that which diminishes the belief of educated men often increases that of the uneducated, always avid for the marvellous.".

Bem's findings —and, perhaps more importantly, the fact that they were published in a major journal— created a storm of media attention. In the *New York Times*, several researchers voiced strong opinions: Dr. Ray Hyman, a long-time critic of ESP research, questioned the quality of the refereeing process as he believed that the publication of Dr. Bem's article was "(...) pure craziness (...) an embarrassment for the entire field"[11], and Dr. Douglas Hofstadter argued for "(...) a cutoff for craziness, and when that threshold is exceeded, then the criteria for publication should get far, far more stringent." Bem's article was also discussed in *Science* (Miller, 2011) and many other media throughout the world. A Google search on "Bem" and "feeling the future" generates over 400,000 hits.[12] Bem himself appeared on the popular US television show *The Colbert Report*, where the host described Bem's work as "extrasensory pornception" referring to the fact that Experiment 1 in Bem (2011) found that precognition was present only for erotic pictures. In the *New York Times*, Bem was quoted as saying "What I showed was that unselected subjects could sense the erotic photos, but my guess is that if you use more talented people, who are better at this, they could find any of the photos."

In our reply to Bem (Wagenmakers et al., 2011), we implicitly suggested that the reviewers and editor had been bamboozled by the sheer number of experiments and participants. A careful evaluation of the Bem article showed that the evidence is much less conclusive than it may appear at first. Specifically, we first noted that the analysis of the experiments had been partly exploratory, whereas the statistical analysis assumed a fully confirmatory approach. That is, as discussed above in more detail, we argued that the data had been used twice: once to draw attention to an interesting result, and then to test it. In support of our claim, we pointed to several instances where it was clear that the analysis had been exploratory.

Next we used the principles of Bayesian belief revision to argue that the bar for publishing should be set higher for claims that are outlandish or improbable (see above: extraordinary claims require extraordinary evidence). Third, we used a default Bayesian t test (Rouder et al., 2009) to highlight that Bem's one-sided p values overestimate the evidence against the null; in fact, our default test indicated little evidence in favor of precognition—only one of Bem's nine experiments yielded data substantially more likely under $\mathcal{H}_1$ (i.e., the hypothesis of precognition) than under $\mathcal{H}_0$.

Despite the advantages of specifying a default (objective) test, we also realized that $\mathcal{H}_1$ can be specified in a different, more subjective manner. To examine the robustness of our conclusions we systematically varied our specification of $\mathcal{H}_1$. The results confirmed that for a wide range of different, non-default prior distributions on effect size the evidence for precognition is either non-existent or negligible.[13]

In their rejoinder, Bem and two reputable statisticians took issue with several of our claims (Bem et al., 2011). The rejoinder sidestepped the issue of exploration[14] and focused

[11] Dr. Hyman did not question the publication of a parapsychological article as such. Instead, Dr. Hyman was puzzled that JPSP had accepted an article with so many departures from accepted methodological practice (Dr. Hyman, personal communication).

[12] Query issued on July 1st, 2014.

[13] The online appendix is available on the first author's website or at `http://www.ruudwetzels.com/articles/Wagenmakersetal_robust.pdf`.

[14] A painfully detailed analysis of the Bem experiments by James Alcock is available at `http://www.csicop.org/specialarticles/show/back_from_the_future`. Bem's response and Alcock's reply can also

mostly on the manner in which the Bayes factors had been calculated. In the first place, Bem et al. (2011) proposed a psi prior that assigns a lot of mass to relatively small effect sizes. In fact, the Bem psi prior happens to resemble the prior that maximizes the evidence against the null hypothesis in the Bem experiments. Nevertheless, even this unrealistic maximum level of evidence was still unimpressive on an experiment-by-experiment basis, confirming the results from the initial robustness analysis (Wagenmakers et al., 2011).

Thus, even when we ignore the issue of exploration, and even when we assume a psi prior that almost maximizes the evidence against the null hypothesis, the results for the individual experiments are still not compelling. At this point, Bem et al. (2011) argued that, instead of assessing the evidential impact of each experiment in isolation, one should combine the evidence across experiments by multiplying the separate Bayes factors. This is a conceptual concession —in the original article, the evidential impact was evaluated for each experiment in isolation— but it is also a serious statistical error.[15] The experiments are only conditionally independent, meaning that knowledge about effect size should be updated across the nine experiments. When one wishes to combine evidence across experiments, it can be appropriate, for instance, to construct a hierarchical model and treat effect size in the different experiments as a random effect. However, given the doubtful origin of the data and the experimental design we suggest that additional analysis efforts are probably misplaced.[16]

In order to resolve such conflicting opinions over the evidential impact of a particular set of studies, the royal road is to conduct replication studies in a purely confirmatory design. Ever since the publication of the Bem study, several high-powered experiments have failed to replicate the effect (Galak, LeBoeuf, Nelson, & Simmons, 2012; Ritchie, Wiseman, & French, 2012; Wagenmakers et al., 2012); nevertheless, Bem and colleagues have recently argued that a meta-analysis supports the presence of the effect (i.e., Bem et al., 2014; for a skeptical review see `http://osc.centerforopenscience.org/2014/06/25/a-skeptics-review/`).

In sum, our discussion with Bem was informative but it was not as productive as we would have hoped. We do believe that a productive exchange between skeptics and proponents is possible in principle, and the next section outlines our ideas of how this can be accomplished.

Part IV. How to Convince a Skeptic in Five Easy Steps

In order to convince the skeptics, proponents of psi —or any other implausible hypothesis— face an uphill battle. If the psi truly exists, then this battle can be won, and it can be won easily. Below is a five-step proposal program that can be implemented without much effort.

1. **Publicly preregister the experiments, and report the results regardless**

be found online.

[15] In their own response to Bem (2011), Rouder and Morey (2011, p. 685) state "Meta analysis seems like it should be a strong point of the Bayes factor. If one has several replicate experiments, it seems reasonable that the posterior odds from the first can serve as the prior for the second, and so on. Under this framework, the combined evidence across all the replicate experiments is simply the product of the Bayes factors. This intuition that the meta-analytic Bayes factor is the product of individual Bayes factors is not correct."

[16] For a more detailed response to Bem's rejoinder see `https://dl.dropboxusercontent.com/u/1018886/ClarificationsForBemUttsJohnson.pdf`.

of outcome. In the first stage of the research program the proponents are free to run exploratory pilot studies at will, cherry-picking phenomena and fine-tuning the paradigm until it reliably produces the effect. But once the research transitions from hypothesis-generating to hypothesis-testing, a preregistration document is required to proceed. Moreover, any preregistration event needs to be publicly announced in order to ensure that the results will be reported even if they do not show the result desired by the proponent.

2. **Involve qualified skeptics**. As pointed out by Diaconis (1991, p. 386): "Since the field has so far failed to produce a replicable phenomena, it seems to me that any trial that asks us to take its findings seriously should include full participation by qualified skeptics. Without a magician and/or knowledgeable psychologist skilled at running experiments with human subjects, I don't think a serious effort is being made."

3. **No p-values**. As hinted at above, p-values overestimate the evidence against the null. In addition, there is an argument that they cannot be used to test hypotheses that are highly implausible (Pandolfi & Carreras, in press). Instead, the evidence for the presence or absence of psi can be quantified by the Bayes factor. Attention needs to be spent on the specification of the prior distribution for effect size. One could either choose the Bem psi prior, or opt for a default prior coupled with a robustness check. These priors can be defined to be one-sided. Importantly, the choice of prior should be outlined in the preregistration document.

4. **Single high-quality experiments**. The focus of the work should be on single, high-quality, high-power experiments; meta-analyses of biased, low-quality studies are not reliable and will not convince a skeptical audience.

5. **Make money**. Munroe's "economic argument" (see Figure 2) plays on a serious implication of real science. Contrary to popular belief, successful applications of scientific theories are not mere add-ons to science; instead, such applications carry a significant evidentiary load. Newton's mechanics has gained a lot of credibility from its engineering applications, and the atomic bomb has conferred enormous evidence upon Einstein's theories as the ultimate application his $e = mc^2$. It is evident that even the smallest psi effects are worth millions if applied in games of chance. An online betting facility that requires the gambler to predict, say, whether a picture will appear at the left or on right of the screen should go bankrupt as soon as extraverted females are allowed to predict the location of erotic pictures (we are happily prepared to invest in such a casino). If psi is demonstrably effective in generating cash, it would quickly be accepted in the pantheon of scientifically credible phenomena.

Concluding Remarks

Many insights have been gained from research on psi. For instance, research on psi has helped make it clear that any statistical method, no matter how sophisticated, can be brought to heel by a combination of selective reporting and motivated analysis procedures. In order to prevent hypothesis-generating research from masquerading as hypothesis-testing research, strict control over the data analysis process is necessary. Such control can only be accomplished by study preregistration, a methodology that has undergone a recent surge of interest across the empirical sciences.

It has become increasingly obvious that all of the concerns that skeptics have raised for research on psi are also relevant for more pedestrian forms of research. Consider, for

instance, the phenomenon of social priming, where a subtle cognitive or emotional manipulation influences overt behavior. The prototypical example is the elderly walking study from Bargh, Chen, and Burrows (1996); in the priming phase of this study, students were either confronted with neutral words or with words that are related to the concept of the elderly (e.g., "Florida", "bingo"). The results showed that the students' walking speed was slower after having been primed with the elderly-related words. In his recent book on decision making, Kahneman (2011, pp. 56-57) writes: "When I describe priming studies to audiences, the reaction is often disbelief (...) The idea you should focus on, however, is that disbelief is not an option. The results are not made up, nor are they statistical flukes. You have no choice but to accept that the major conclusions of these studies are true." At the 2014 APS annual meeting in San Francisco, however, Hal Pashler presented a long series of failed replications of social priming studies, conducted together with Christine Harris, the upshot of which was that disbelief does in fact remain an option.[17]

Psi research, as a phenomenon, thus has substantial implications for the way we organise our scientific discipline. For if psi does not exist, as we have assumed here, this shows that the current methodological framework of scientific psychology is unable to stand its ground against standard research practices involving data selection, hypothesising after the data come in, and similar QRPs. The antidote against these QRPs is well known and can be implemented simply by insisting on a separation of explanatory and confirmatory work, and public preregistration of experiments designed to serve a hypothesis-testing function. As will be evident, we are unconvinced by the need for the drastic revision of belief that psi enthusiasts advocate, but we do observe Cromwell's rule, and have clearly described the manner in which our opinions may be swayed.

The consequences of the new wave of direct replications and renewed methodological rigor are as yet unknown, but one thing is certain: they can only improve the quality and dependability of our science. And for this alone we should be grateful that psi research exist. As academic jesters, psi researchers could not have done their job any better.

[17] For concrete examples see `http://laplab.ucsd.edu/publications`.

References

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype-activation on action. *Journal of Personality and Social Psychology, 71*, 230–244.

Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature, 483*, 531–533.

Bem, D., Tressoldi, P. E., Rabeyron, T., & Duggan, M. (2014). Feeling the future: A meta-analysis of 90 experiments on the anomalous anticipation of random future events. *Manuscript submitted for publication.*

Bem, D. J. (2003). Writing the empirical journal article. In J. M. Darley, M. P. Zanna, & H. L. Roediger III (Eds.), *The compleat academic: A career guide* (pp. 171–201). Washington, DC: American Psychological Association.

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

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

Berger, J. O., & Wolpert, R. L. (1988). *The likelihood principle (2nd ed.)*. Hayward (CA): Institute of Mathematical Statistics.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association, 33*, 526–536.

Chambers, C. D. (2013). Registered Reports: A new publishing initiative at Cortex. *Cortex, 49*, 609–610.

De Groot, A. D. (1956/2014). The meaning of "significance" for different types of research. Translated and annotated by Eric-Jan Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier Kievit, Marjan Bakker, Angelique Cramer, Dora Matzke, Don Mellenbergh, and Han L. J. van der Maas. *Acta Psychologica, 148*, 188–194.

De Groot, A. D. (1969). *Methodology: Foundations of inference and research in the behavioral sciences.* The Hague: Mouton.

Diaconis, P. (1991). Comment. *Statistical Science, 6*, 386.

Donnellan, M. B., Lucas, R. E., & Cesario, J. (in press). On the association between loneliness and bathing habits: Nine replications of Bargh and Shalev (2012) Study 1. *Emotion.*

Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE, 7*, e29081.

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

Goldacre, B. (2008). *Bad science.* London: Fourth Estate.

Greenwald, A. G. (1975). Significance, nonsignificance, and interpretation of an ESP experiment. *Journal Of Experimental Social Psychology, 11*, 180–191.

Harris, C. R., Coburn, N., Rohrer, D., & Pashler, H. (2013). Two failures to replicate high-performance-goal priming effects. *PLoS ONE, 8*, e72467.

Hyman, R. (2010). Meta–analysis that conceals more than it reveals: Comment on Storm et al. (2010). *Psychological Bulletin, 136*, 486–490.

Ioannidis, J. P. A. (2012). Why science is not necessarily self–correcting. *Perspectives on Psychological Science*, *7*, 645–654.

Jaynes, E. T. (2003). *Probability theory: The logic of science.* Cambridge: Cambridge University Press.

Jefferys, W. H. (1990). Bayesian analysis of random event generator data. *Journal of Scientific Exploration*, *4*, 153–169.

Jeffreys, H. (1961). *Theory of probability* (3 ed.). Oxford, UK: Oxford University Press.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth–telling. *Psychological Science*, *23*, 524–532.

Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, 19313–19317.

Kahneman, D. (2011). *Thinking, fast and slow.* London: Allen Lane.

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

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*, 196–217.

LeBel, E. P., & Campbell, L. (in press). Heightened sensitivity to temperature cues in highly anxiously attached individuals: Real or elusive phenomenon? *Psychological Science.*

LeBel, E. P., & Wilbur, C. J. (in press). Big secrets do not necessarily cause hills to appear steeper. *Psychonomic Bulletin & Review.*

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian modeling for cognitive science: A practical course.* Cambridge University Press.

Lindley, D. V. (1991). *Making decisions* (2 ed.). New York: Wiley.

Meehl, P. E., & Scriven, M. (1956). Compatibility of science and ESP. *Science*, *123*, 14–15.

Miller, G. (2011). News of the week: ESP paper rekindles discussion about statistics. *Science*, *331*, 272–273.

Mossbridge, J. A., Tressoldi, P., Utts, J., Ives, J. A., Radin, D., & Jonas, W. B. (2014). Predicting the unpredictable: Critical analysis and practical implications of predictive anticipatory activity. *Frontiers in Human Neuroscience*, *8:146.*

Nickerson, R. S. (2000). Null hypothesis statistical testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241–301.

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, *45*, 137–141.

Nuzzo, R. (2014). Statistical errors. *Nature*, *506*, 150–152.

Open Science Collaboration, T. (2012). An open, large–scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, *7*, 657–660.

Otto, B. K. (2001). *Fools are everywhere: The court jester around the world.* Chicago: University of Chicago Press.

Pandolfi, M., & Carreras, G. (in press). The faulty statistics of complementary alternative medicine (CAM). *European Journal of Internal Medicine.*

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, *7*, 531–536.

Pashler, H., Rohrer, D., & Harris, C. R. (2013). Can the goal of honesty be primed? *Journal of Experimental Social Psychology*, *49*, 959–964.

Price, G. R. (1955). Science and the supernatural. *Science*, *122*, 359–367.

Price, G. R. (1956). Where is the definitive experiment? *Science*, *123*, 17–18.

Price, G. R. (1972). Apology to Rhine and Soal. *Science*, *175*, 359.

Rhine, J. B. (1956). Comments on "Science and the supernatural". *Science*, *123*, 11–14.

Ritchie, S. J., Wiseman, R., & French, C. C. (2012). Failing the future: Three unsuccessful attempts to replicate Bem's 'retroactive facilitation of recall' effect. *PLoS ONE*, *7*, e33423.

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

Rouder, J. N., & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research*, *47*, 877–903.

Rouder, J. N., Morey, R. D., & Province, J. M. (2013). A Bayes-factor meta-analysis of recent ESP experiments: Comment on Storm, Tressoldi, and Di Risio (2010). *Psychological Bulletin*, *139*, 241–247.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*, 356–374.

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

Royall, R. M. (1997). *Statistical evidence: A likelihood paradigm.* London: Chapman & Hall.

Schwarzkopf, D. S. (2014). We should have seen this coming. *Frontiers in Human Neuroscience*, *8:332*.

Shanks, D. R., Newell, B. R., Lee, E. H., Balakrishnan, D., Ekelund, L., Cenac, Z., Kavvadia, F., & Moore, C. (2013). Priming intelligent behavior: An elusive phenomenon. *PLoS ONE*, *8*, e56515.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False–positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.

Soal, S. G. (1956). On "Science and the supernatural". *Science*, *123*, 9–11.

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

Stroebe, W., Postmes, T., & Spears, R. (2012). Scientific misconduct and the myth of self-correction in science. *Perspectives on Psychological Science*, *7*, 670-688.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*, 433-460.

Utts, J. (1991). Replication and meta–analysis in parapsychology (with discussion). *Statistical Science*, *6*, 363–403.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*, 779–804.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi. *Journal of Personality and Social Psychology, 100*, 426–432.

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*, 627–633.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science, 6*, 291–298.

Wetzels, R., & Wagenmakers, E.-J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review, 19*, 1057–1064.

Wolfe, J. M. (2013). Registered reports and replications in Attention, Perception, & Psychophysics. *Attention, Perception, & Psychophysics, 75*, 781–783.