Bayesian Mixture Modeling of Significant P Values: A Meta-Analytic Method to Estimate the Degree of Contamination from $\mathcal{H}_0$

Quentin Frederik Gronau[1], Monique Duizer[1], Marjan Bakker[2], & Eric-Jan Wagenmakers[1]

[1]University of Amsterdam

[2]Tilburg University

Correspondence concerning this article should be addressed to:
Quentin Frederik Gronau
University of Amsterdam, Department of Psychology
Weesperplein 4
1018 XA Amsterdam, The Netherlands
E-mail may be sent to quentingronau@web.de

Abstract

Publication bias and questionable research practices have long been known to corrupt the published record. One method to assess the extent of this corruption is to examine the meta-analytic collection of significant p values, the so-called p-curve (Simonsohn, Nelson, & Simmons, 2014a). Inspired by statistical research on false-discovery rates, we propose a Bayesian mixture model analysis of the p-curve. Our mixture model assumes that significant p values arise either from the null-hypothesis $\mathcal{H}_0$ (when their distribution is uniform) or from the alternative hypothesis $\mathcal{H}_1$ (when their distribution is accounted for by a flexible nonparametric technique known as the Dirichlet process mixture). The model estimates the proportion of significant results that originate from $\mathcal{H}_0$, but it also estimates the probability that each specific p value originates from $\mathcal{H}_0$. We apply our model to two concrete examples from the published literature. Model code is provided in the online Supplemental Material.

Keywords: Publication bias, p-hacking, p-curve, Bayesian mixture model, Dirichlet process

Psychological science is experiencing a crisis of confidence (e.g., Pashler & Wagenmakers, 2012). In response to this crisis, psychologists have offered new guidelines for journals (e.g., Nosek et al., 2015), started large-scale replication initiatives (e.g., Open Science Collaboration, 2015), promoted preregistration (e.g., Chambers, 2013, 2015; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012), suggested different statistical reporting practices (e.g., Eich, 2014), and developed novel statistical techniques (e.g., Francis, 2013; Guan & Vandekerckhove, in press; Simonsohn et al., 2014a; van Assen, van Aert, & Wicherts, 2015).

Among the various newly developed statistical techniques, the p-curve procedure is of special interest (Simonsohn et al., 2014a; Simonsohn, Nelson, & Simmons, 2014b). This procedure considers a collection of significant p values and asks whether their distribution contains "evidential value". This question can be answered because of the fact that, under $\mathcal{H}_0$, the distribution of significant p values is uniform (Becker, 1991). Hence, if the observed distribution of significant p values is relatively flat, the most likely explanation for the findings is publication bias (e.g., Rosenthal, 1979; Sterling, 1959; Sterling, Rosenbaum, & Weinkam, 1995). In addition, when most observed p values are near .05 this indicates that the findings maybe have been the result of significance chasing (i.e., "p-hacking"; John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011). In the presence of a true effect, however, the distribution of p values is right-skewed such that low p values occur more often than high p values. The current p-curve analysis conducts a classical hypothesis test on the observed p values and concludes that their distribution contains "evidential value" when it is judged to be right-skewed.

The classical p-curve analysis is a promising tool to obtain an overall impression about the presence of true effects. Here we present a novel and complementary Bayesian analysis of the p-curve that approaches the problem from a slightly different angle. Similar to an analysis of false-discovery rates, our Bayesian method assumes that the observed significant p values may have originated from $\mathcal{H}_0$ or $\mathcal{H}_1$. The method then estimates the overall rate of contamination from $\mathcal{H}_0$; in addition, the method estimates the probabilities that each specific p value originates from $\mathcal{H}_0$. These estimates can help assess, on a continuous scale, the extent to which an empirical phenomenon is based on p values that are spurious. Below we first outline the method and then apply it to two concrete examples.

The Bayesian Mixture Model for Significant P Values

We start from the assumption that the observed p-curve is a mixture between two distributions: a uniform distribution associated with $\mathcal{H}_0$ and a right-skewed distribution

associated with $\mathcal{H}_1$. Thus,

$$p(p_i) = \phi\, f_{\mathcal{H}_0} + (1 - \phi)\, f_{\mathcal{H}_1},$$

where p_i denotes a specific observed p value, and $\phi \in [0, 1]$ represents a mixing parameter that can be regarded as an estimate of the proportion of studies originating from $\mathcal{H}_0$. Values of ϕ near 1 indicate that the collection of studies are heavily contaminated by $\mathcal{H}_0$.

As a first step, we probit-transform the p values in order to be able to use normal distributions (e.g., Efron, 2012; Tamhane & Shi, 2009). Under $\mathcal{H}_0$, the uniform distribution of raw p values corresponds to a standard normal distribution for the probit-transformed p values (i.e., $\Phi^{-1}(p) \mid \mathcal{H}_0 \sim N(\mu = 0, \sigma^2 = 1)$). Under $\mathcal{H}_1$, the exact distribution of p values is more complex and depends on several factors such as sample size and the values of population parameters that are relevant for the test statistic at hand (Becker, 1991). Furthermore, a given collection of observed p values will be comprised of studies with different sample sizes, different test statistics, and, potentially, different true effects; that is, there exists an unknown distribution of true effects such that the collection of observed p values is inherently heterogeneous.

Thus, in the second step, we need to address the fact that the distribution of p values under $\mathcal{H}_1$ is a combination of potentially many different distributions. One tempting method to deal with this complication is to ignore it, risk the possibility of model-misspecification, and proceed by using a simple parametric form for the probitized p values under $\mathcal{H}_1$ (e.g., $\Phi^{-1}(p) \mid \mathcal{H}_1 \sim N(\mu, \sigma^2)$). In this manuscript we explore a different method, one that respects the complex distribution of p values under $\mathcal{H}_1$ by employing a flexible nonparametric Bayesian formalization known as the Dirichlet process mixture. Details, code, and simulations can be found in the Supplemental Material available at `https://osf.io/mysbp/`.

The Dirichlet process (Freedman, 1963; Ferguson, 1973, 1974) is a prior for an infinite normal mixture model, such that the complexity of the model can grow flexibly with the data (Gershman & Blei, 2012; Müller, Quintana, Jara, & Hanson, 2015; Navarro, Griffiths, Steyvers, & Lee, 2006). Previous work has also used Dirichlet process mixtures to estimate false-discovery rates (e.g., Do, Müller, & Tang, 2005; Tang, Ghosal, & Roy, 2007); however, because these models consider all p values –not just the ones that are significant– they are not suitable for the analysis of p-curves.

In sum, our Bayesian model conceives of the distribution of significant p values as a two-component mixture, where one component corresponds to the uniform distribution of significant p values under $\mathcal{H}_0$ and the other component corresponds to the unknown distribution of significant p values under $\mathcal{H}_1$. The unknown distribution of significant p values under $\mathcal{H}_1$ is accounted for using a flexible nonparametric Bayesian procedure (i.e.,

the Dirichlet process mixture, see Supplemental Material for details). We now apply the model to two examples.

Example 1: 587 T-Test P Values

For our first example we apply the model to a set of p values from Wetzels et al. (2011); these authors collected the results from all 855 t-tests reported in the articles from the 2007 issues of *Psychonomic Bulletin & Review* and *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Here we focus on the subset of 587 p values that were significant. It should be noted that these significant p values are inherently heterogeneous: they come from a wide range of empirical fields, and they were not screened for relevance. Thus, it is important to keep in mind that many of these p values may correspond to manipulation checks, and only a subset corresponds to the test of the key research hypothesis. Because of their heterogeneous nature, this set of p values provides a good test case for our model.

The top-left panel of Figure 1 shows the distribution of the 587 significant p values. The same distribution was inspected by Johnson (2013), who argued that the significant p values

> "...presumably arise from two types of experiments: experiments in which a true effect was present and the alternative hypothesis was true, and experiments in which there was no effect present and the null hypothesis was true. For the latter experiments, the nominal distribution of P values is uniformly distributed on the range $(0.0, 0.05)$ (...) The P values displayed in this plot thus represent a mixture of a uniform distribution and some other distribution. Even without resorting to complicated statistical methods to fit this mixture, the appearance of this histogram suggests that many, if not most, of the P values falling above 0.01 are approximately uniformly distributed. That is, most of the significant P values that fell in the range $(0.01 - 0.05)$ probably represent P values that were computed from data in which the null hypothesis of no effect was true."

Nevertheless, the overall distribution of p values is clearly right-skewed, and many p values are relatively low. This impression is corroborated by the results of our Bayesian mixture model. Specifically, the top-right panel of Figure 1 shows the posterior distribution of ϕ, the $\mathcal{H}_0$ assignment rate. This contamination rate is estimated to be small, and a Bayesian 95% highest density interval ranges from 0.003 to 0.150. The Markov chain Monte Carlo chains for ϕ are shown in Figure 2, supporting the claim that the samples come from the posterior distribution.

In addition to the estimation of the overall contamination rate, the Bayesian mixture model also allows us to estimate the probability that each individual p value is assigned to

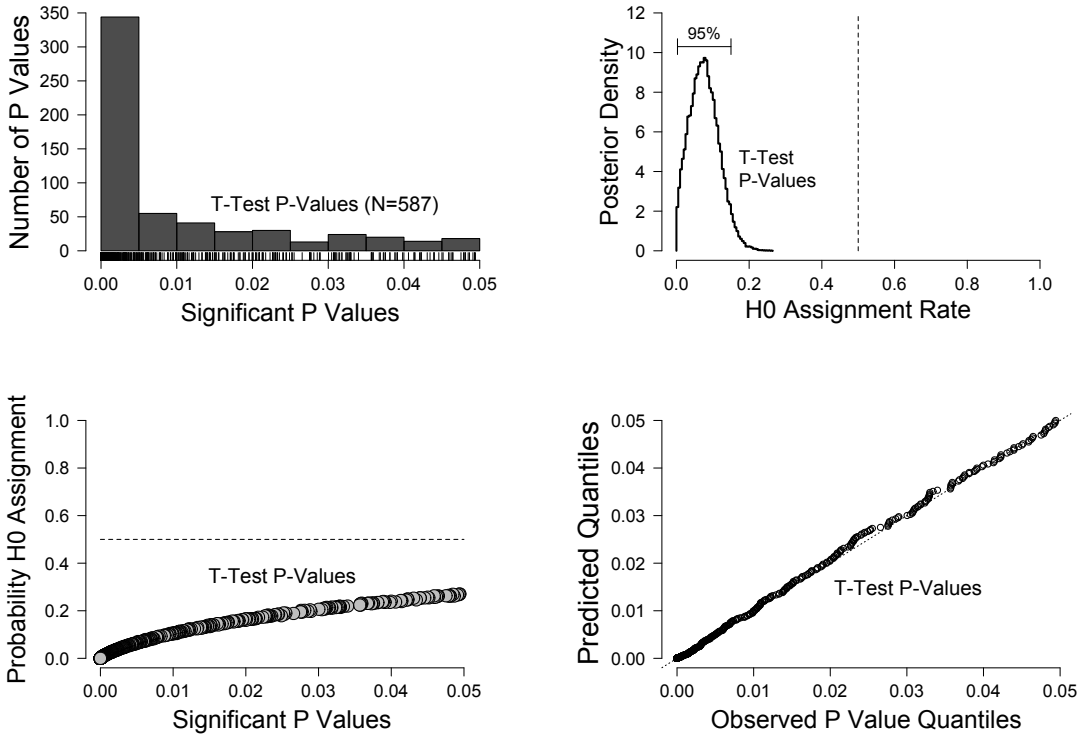

Figure 1. Application of the Bayesian mixture model to Example 1: 587 t-test p values. Upper-left panel: distribution of observed p values; upper-right panel: posterior distribution of the $\mathcal{H}_0$ assignment rate; lower-left panel: individual $\mathcal{H}_0$ assignment probabilities; lower-right panel: Q-Q plot for comparing the observed p value distribution to the posterior predictive distribution.

$\mathcal{H}_0$. These estimates are shown in the lower-left panel of Figure 1. The results indicate that none of the $\mathcal{H}_0$ assignment probabilities is larger than .5; this means that, starting from a position of equipoise, all observed significant p values are more likely to stem from $\mathcal{H}_1$ than from $\mathcal{H}_0$. This quantitative conclusion is somewhat more positive than the qualitative conclusion drawn by Johnson (2013).

Finally, the lower-right panel of Figure 1 shows the model fit by means of a Q-Q plot. The Q-Q plot allows a comparison between the distribution of observed p values and the distribution of posterior predictive p values, that is, the distribution of p values predicted by the model. Identical distributions yield a linear Q-Q plot with a slope of one. The present Q-Q plot suggests that the model fit is excellent.

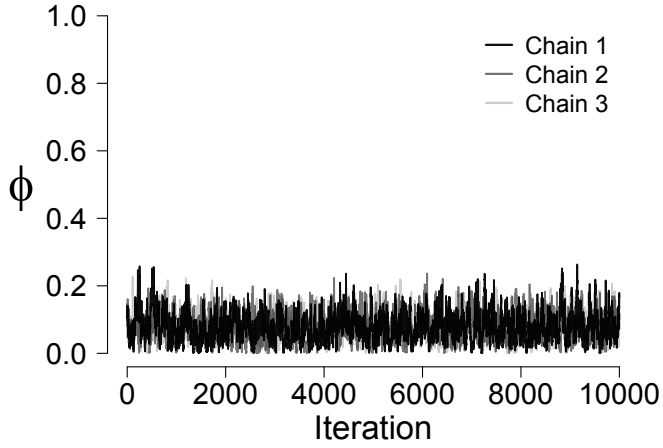

Figure 2. Markov chain Monte Carlo samples for the $\mathcal{H}_0$ assignment rate ϕ for Example 1: 587 t-test p values. The chains intermix, suggesting convergence to the posterior.

Example 2: Social Priming Studies and Yoked Controls

For our second example we apply the model to a set of p values from social priming studies (e.g., Kahneman, 2011) and a matched set of p values for yoked control studies. To obtain the p values for the social priming studies we collected a large set of articles published by a group of prominent researchers who study social priming. We used this selection method in order to obtain a relatively high-quality set of studies, thereby maximizing the probability of collecting p values that are compelling and relatively uncontaminated. We followed the p-curve instructions from Simonsohn et al. (2014a) and distilled a single significant p value from each experiment. Every p value was evaluated by three raters; occasional differences of opinion were readily resolved by discussion.

In addition, we sought to construct an appropriate comparison set of p values as a backdrop against which to evaluate the results for the social priming studies. This comparison set was constructed by selecting, for each social priming study under consideration, a yoked control study – that is, a study on a different topic and published in the same journal issue immediately after the social priming study. For each experiment in the yoked control studies, we distilled a single significant p value in the same manner as was done for the social priming studies.

This procedure yielded a total of 159 significant social priming p values and 130 significant yoked control p values. Further details regarding the studies that were included are available at `https://osf.io/98qsb/` (social priming studies) and `https://osf.io/2zhfy/`

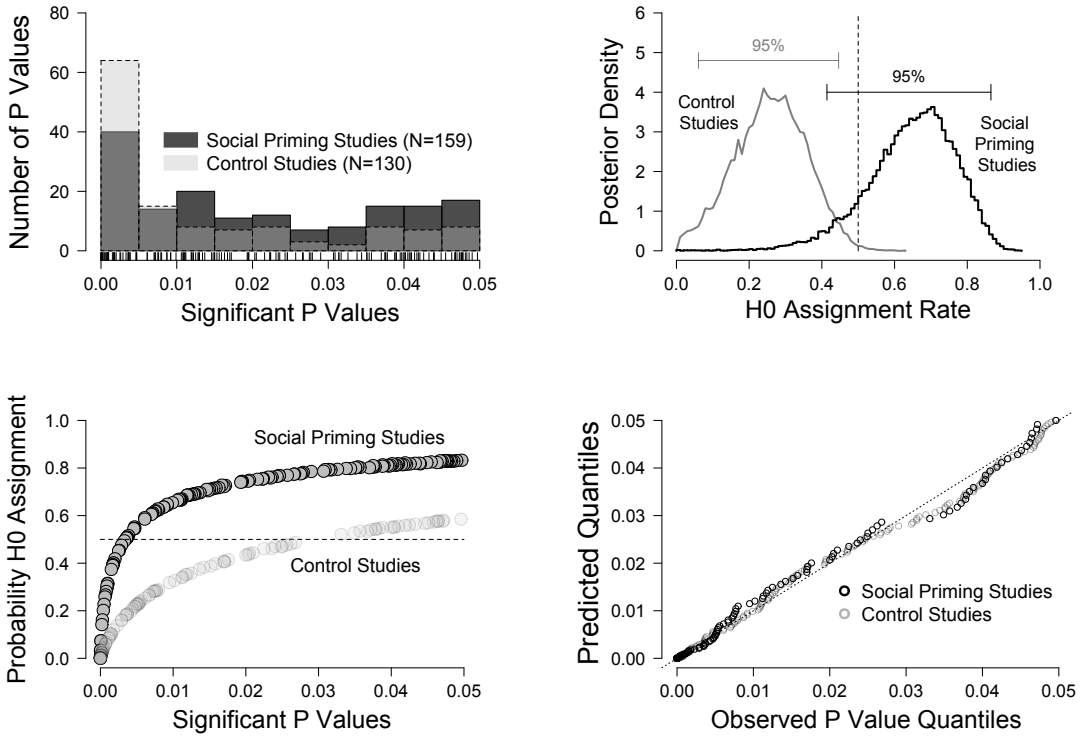

Figure 3. Application of the Bayesian mixture model to Example 2: social priming studies and yoked controls. Upper-left panel: distribution of observed p values; upper-right panel: posterior distribution of the $\mathcal{H}_0$ assignment rates; lower-left panel: individual $\mathcal{H}_0$ assignment probabilities; lower-right panel: Q-Q plot for comparing the observed p value distributions to the posterior predictive distributions.

(control studies).

Figure 3 summarizes the results from applying our Bayesian mixture model. The upper-left panel shows the distributions of p values for the social priming experiments and the yoked controls. Although both distributions are right-skewed, the extent of this skew is much less pronounced than for the t-test p values from Example 1. Furthermore, the distribution of p values for the social priming studies shows less skew than that for the yoked control studies. Both distributions look relatively flat from .05 to .01.

The upper-right panel of Figure 3 displays the posterior distributions of the $\mathcal{H}_0$ assignment rate ϕ. For both sets of p values, the degree of $\mathcal{H}_0$ contamination is substantial; the $\mathcal{H}_0$ assignment rate for the social priming p values has a 95% highest density interval that ranges from 0.414 to 0.865; for the yoked control p values, this interval ranges from 0.061 to 0.446. The social priming studies appear to suffer more from $\mathcal{H}_0$ contamination than do the yoked controls. The Markov chain Monte Carlo chains for ϕ are shown in

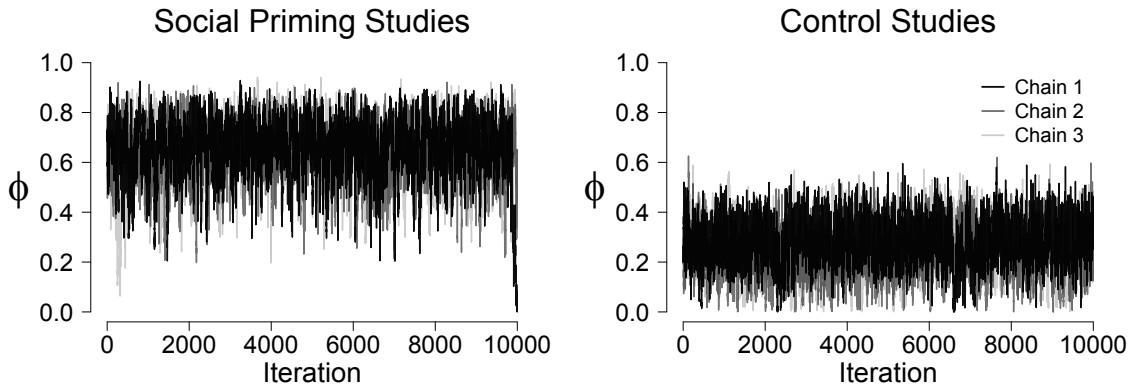

Figure 4. Markov chain Monte Carlo samples for the $\mathcal{H}_0$ assignment rate ϕ for Example 2: social priming studies and yoked controls. The chains intermix, suggesting convergence to the posterior. Left panel: social priming studies; right panel: yoked controls.

Figure 4, supporting the claim that the samples come from the posterior distributions.

The lower-left panel of Figure 3 shows the $\mathcal{H}_0$ assignment probabilities for the individual p values. These probabilities exceed .5 for 81% of the social priming p values and for 19% of the yoked control p values. The lower-right panel of Figure 3 shows the Q-Q plots; the fit for both sets of p values is perhaps subject to improvement, although bootstrap simulations indicate that the observed deviation from the identity line falls within an acceptable range.

Thus, any sweeping negative conclusions regarding the social priming studies need to be tempered by two insights. The first insight is that the posterior distribution of the $\mathcal{H}_0$ assignment rate is relatively wide, and it cannot be ruled out that the contamination rate it is as low as .4. The second insight is that the Q-Q plot reveals that, despite its flexible nonparametric nature, the mixture model may not have been able to fully account for the observed data pattern. For these reasons, the results of our analysis should be interpreted with a modicum of caution. Nevertheless, our analyses certainly do suggest that the level of $\mathcal{H}_0$ contamination in social priming studies warrants more scrutiny.

Concluding Comments

For studies that feature only a limited number of experiments, currently the sole arbiter of success is whether –for each experiment– the p value is lower than .05. This unfortunate state of affairs encourages publication bias, selective reporting, and questionable research practices (e.g., Barber, 1976). When studies are combined, however, the shape of the distribution of significant p values conveys additional information that allows one

to estimate the degree of the bias. To this aim, a classical "p-curve" analysis method was recently proposed by Simonsohn et al. (2014a). Here we presented an alternative Bayesian analysis of the p-curve. Our nonparametric Bayesian mixture model was inspired by a suggestion from Johnson (2013) and previous work on the control of false-discovery rates. The mixture model estimates the extent to which the overall results have been contaminated by $\mathcal{H}_0$; in addition, the method allows researchers to estimate how likely it is that a particular p value stems from $\mathcal{H}_0$.

Similar to the classical analysis method for p-curves, our model makes a number of assumptions. One assumption is that, under $\mathcal{H}_0$, the distribution of p values is uniform. In practice, this assumption may not hold; that is, particular forms of cherry-picking and questionable research practices may yield a p-curve that is right-skewed, thereby masquerading as the signature of a real effect. Consequently, the contamination rate estimated using our Bayesian model can be considered a lower bound on the true level of contamination from $\mathcal{H}_0$. Another assumption is that our analysis departs from a position of equipoise – the default prior on the contamination rate is uniform from 0 to 1 (see Supplemental Material). If needed, this default prior can be adjusted to incorporate existing knowledge; for example, when applied to set of p values for studies on extrasensory perception, a more appropriate prior on the contamination rate is a skewed beta distribution with a mode at $\phi = 1$. The analysis from equipoise allows one to assess the information in the data, but our final beliefs are always a combination from the information in the data and the prior information: in contrast to what current practice may suggest, statistical inference does not take place in a vacuum (Savage, 1954; Lindley, 2004).

We applied our mixture model to a set of significant p values from Wetzels et al. (2011) and to a set of significant p values from social priming and yoked control studies. The examples highlighted the added inferential value of our model. We have provided the model code in the online Supplemental Material, and we hope to incorporate the method in a future release of JASP (Love et al., 2015, `jasp-stats.org`). This way we hope to encourage other researchers to apply the model within their field of interest.

References

Barber, T. X. (1976). *Pitfalls in human research: Ten pivotal points.* New York: Pergamon Press Inc.

Becker, B. J. (1991). Small–sample accuracy of approximate distributions of functions of observed probabilities from t tests. *Journal of Educational Statistics*, *16*, 345–369.

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

Chambers, C. D. (2015). Ten reasons why journals must review manuscripts before results are known. *Addiction*, *110*, 10–11.

Do, K.-A., Müller, P., & Tang, F. (2005). A Bayesian mixture model for differential gene expression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *54*(3), 627–644.

Efron, B. (2012). Large-scale simultaneous hypothesis testing. *Journal of the American Statistical Association*, *99*, 96 – 104.

Eich, E. (2014). Business not as usual. *Psychological Science*, *25*, 3–6.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The annals of statistics*, *1*, 209–230.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *The annals of statistics*, *2*, 615–629.

Francis, G. (2013). Replication, statistical consistency, and publication bias. *Journal of Mathematical Psychology*, *57*, 153–169.

Freedman, D. A. (1963). On the asymptotic behavior of Bayes' estimates in the discrete case. *The Annals of Mathematical Statistics*, *34*, 1386–1403.

Gershman, S. J., & Blei, D. M. (2012). A tutorial on bayesian nonparametric models. *Journal of Mathematical Psychology*, *56*, 1–12.

Guan, M., & Vandekerckhove, J. (in press). A Bayesian approach to mitigation of publication bias. *Psychonomic Bulletin & Review*.

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.

Lindley, D. V. (2004). That wretched prior. *Significance*, *1*, 85–87.

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, J., ... Wagenmakers, E.-J. (2015). JASP (version 0.7.1.12). *Computer Software*.

Müller, P., Quintana, F. A., Jara, A., & Hanson, T. (2015). *Bayesian nonparametric data analysis*. Cham, Switzerland: Springer.

Navarro, D. J., Griffiths, T. L., Steyvers, M., & Lee, M. D. (2006). Modeling individual differences using Dirichlet processes. *Journal of Mathematical Psychology*, *50*, 101–122.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*, 1422–1425.

Open Science Collaboration, T. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716.

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*, 528–530.

Rosenthal, R. (1979). An introduction to the file drawer problem. *Psychological Bulletin*, *86*, 638–641.

Savage, L. J. (1954). *The foundations of statistics*. New York: John Wiley & Sons.

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.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014a). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*, 534-547.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014b). P-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science*, *9*, 666-681.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance–or vice versa. *Journal of the American Statistical Association*, *54*, 30–34.

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician*, *49*, 108–112.

Tamhane, A. C., & Shi, J. (2009). Parametric mixture models for estimating the proportion of true null hypotheses and adaptive control of FDR. *Lecture Notes-Monograph Series*, *57*, 304–325.

Tang, Y., Ghosal, S., & Roy, A. (2007). Nonparametric Bayesian estimation of positive false discovery rates. *Biometrics*, *63*(4), 1126–1134.

van Assen, M. A. L. M., van Aert, R. C. M., & Wicherts, J. M. (2015). Meta–analysis using effect size distributions of only statistically significant studies. *Psychological Methods*, *20*, 293–309.

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.