Birkbeck College
University of London
London WC1E 7HX, UK
f.steinberger@bbk.ac.uk

References

Dorst, K. forthcoming. Lockeans maximize expected accuracy. *Mind*.

Easwaran, K. 2016. Dr. Truthlove, or, how I learned to stop worrying and love Bayesian probability. *Noûs* 50: 816–53.

Easwaran, K. and B. Fitelson. 2015. Accuracy, coherence, and evidence. In *Oxford Studies in Epistemology*, vol. 5, eds. T. Szabó Gendler and J. Hawthorne. Oxford: Oxford University Press, 61–96.

Foley, R. 1979. Justified inconsistent beliefs. *American Philosophical Quarterly* 16: 247–57.

Friedman, J. 2013. Suspended judgment. *Philosophical Studies* 162: 165–81.

Meacham, C.J.G. forthcoming. Can all-accuracy accounts justify evidential norms? In *Epistemic Consequentialism*, eds. K. Ahlstrom-Vij and J. Dunn. Oxford: Oxford University Press.

Pettigrew, R. 2013. Accuracy and evidence. *Dialectica* 67: 579–96.

Pettigrew, R. 2016a. *Accuracy and the Laws of Credence*. Oxford: Oxford University Press.

Pettigrew, R. 2016b. Jamesian epistemology formalised: an explication of 'The will to believe'. *Episteme* 3: 253–68.

Pettigrew, R. 2017. Epistemic utility and the normativity of logic. *Logos and Episteme* 8: 455–92.

Shear, T. and B. Fitelson. forthcoming. Two approaches to belief revision. *Erkenntnis*.

Sylvan, K. 2018. Veritism unswamped. *Mind* 127: 381–435.

P-curving x-phi: Does experimental philosophy have evidential value?

MICHAEL T. STUART, DAVID COLAÇO AND EDOUARD MACHERY

1. Introduction

'Experimental philosophy' ('x-phi') refers to the use of experimental tools to investigate questions that bear on philosophical issues. This approach remains controversial on several grounds. One challenge is that x-phi research suffers from questionable research practices, including the selective reporting of statistically significant findings and p-hacking. If this were the case, many of the effects that experimental philosophers claim to have identified would

Analysis Vol 79 | Number 4 | October 2019 | pp. 669–684 doi:10.1093/analys/anz007

likely be false positives, and they could not justifiably be brought to bear on any philosophical issue.

In this article, we examine whether the corpus of x-phi suffers from selection bias and p-hacking by employing a p-curve (Simonsohn et al. 2014a). A p-curve plots the distribution of statistically significant p-values for a corpus of experimental studies. The way the p-curve deviates from the uniform distribution indicates whether the significant results likely resulted from selection bias and p-hacking (more on this below).

We developed a dataset of 365 published chapters and articles, which includes all x-phi studies written in English that we could locate, up to the year 2016. We then applied a number of p-curves to the corpus as a whole and to sections of this corpus. Our results suggest that x-phi findings as a whole do not result from selection bias or p-hacking and thus that, overall, the corpus has evidential value. We also find that p-hacking may have occurred in a few subsets of this corpus and that x-phi has methodologically improved.

Here is how we will proceed. In §1.1, we further motivate our project. In §1.2, we describe p-curves in more detail. We report our methods and results in §2 and §3 and discuss the significance of our findings in §4.

1.1 Motivation

Challenges to the relevance of x-phi findings can be divided into two categories: conceptual and empirical challenges. Conceptual challenges include arguments to the effect that, regardless of the quality of x-phi studies, their findings are not relevant to philosophical theorizing. This might be due to a misunderstanding of philosophical methodology or the nature of philosophical argumentation (Cappelen 2012, Deutsch 2015, Stuart 2015, Colaço and Machery 2017, Machery 2017). Empirical challenges, by contrast, relate to how x-phi studies are designed and their results analysed. These might take issue with experimental design (Cullen 2010, Woolfolk 2011, 2013, Scholl MS) or replication failures (Seyedsayamdost 2015a, 2015b, Kim and Yuan 2015, Adleberg et al. 2015, Machery et al. 2017). Woolfolk, for instance, asserts that 'the experiments conducted by experimental philosophers frequently fail to meet the methodological standards that are articulated by the experts on research design in those fields they would emulate' (2013: 80). Empirical challenges raise in practice, as opposed to in principle doubts about the relevance of the findings of x-phi to philosophical discussion.

While the methodological concerns one might have about x-phi are manifold, here are two specific reasons to worry that x-phi experimenters have inadvertently engaged in p-hacking. First, most x-phi studies use the methods of social psychology, a field that has recently faced allegations of p-hacking (Simmons et al. 2011), one likely source of its current 'replication crisis' (Open Science Collaboration 2015). Given that x-phi inherits many of its experimental and statistical practices from social psychology, one may worry

that it suffers from the same problems. In this spirit, Vazire (2015: 46) notes that some common practices in psychology 'increase the chances of producing false, unreplicable results. These practices have come to be called "questionable research practices" (QRPs), or "p-hacking", but it is important to note that most were not considered questionable by many until recently, and some are still taught in textbooks and research methods courses'. Thus, in her view, 'the most important thing experimental philosophers can learn from psychologists is to avoid these practices, follow the new "best practices", and spare themselves the decades of growing pains that psychology has experienced' (Vazire 2015: 46). Second, experimental philosophers are professionally trained as philosophers, and may therefore have limited experience as empirical researchers (Williamson 2010). This has led to speculation that experimental philosophers lack the requisite training in statistics and experimental design, leading them to engage in research practices they do not recognize as questionable (Woolfolk 2011, 2013).

Recent work assuages to some extent the concerns about the quality of methodological practices in experimental philosophy. The XPhi Replicability Project (Cova et al. 2018) set out to replicate 40 x-phi studies, some of which were chosen based on their citation number, while others were chosen at random. X-phi findings so far turn out to exhibit greater reproducibility than those in social psychology.

While the XPhi Replicability Project has numerous virtues, it is also limited in important respects. First, because few studies were targeted for replication, and because these were not entirely chosen at random, the targeted studies may not be representative. Our project has a much larger dataset, which is more representative of x-phi. Thus, we can determine the evidential status of the research corpus as a whole. Further, because of the small number of studies in Cova et al. (2018), claims about sections of x-phi cannot be made with confidence. In contrast, our large dataset allows us to investigate facets of this corpus in more detail, including changes in the quality of published studies over time, across research questions, and more.

Colombo and colleagues (2018) have examined the quality of the methods in x-phi from another angle. Inconsistencies in statistical reporting (discrepancies between the reported parameters of statistical tests and the reported p-values) are surprisingly common in psychology (Bakker and Wicherts 2011, Nuijten et al. 2016). The rate of inconsistencies in statistical reporting in 220 x-phi studies was found to be lower than in psychology.

While these findings are valuable, they are consistent with rampant p-hacking in x-phi, which only a p-curve can detect. Colombo and colleagues also p-curved their results, but their sample is smaller than ours and they did not p-curve subsections of their sample, as we do below.

1.2 P-curve analysis

A p-curve assesses the evidential value of a corpus of studies by determining if the distribution of the p-values of the main significant results indicates selection bias and p-hacking (Simonsohn et al. 2014a, 2014b, Simonsohn et al. 2015). P-curve analysis has already been used to identify several corpora in psychology and neuroscience that suffer from p-hacking (e.g., Segerstrom and Miller 2013, Simonsohn et al. 2014a, Vadillo et al. 2016, Medina and Cason 2017, Simmons and Simonsohn 2017) as well as to provide evidence that some corpora do have evidential value (e.g. Mahowald et al. 2016).

P-curve analysis works as follows. If all the null hypotheses tested by a set of studies are true, then the corresponding p-values should be uniformly distributed, and the p-curve for such a set of studies should look like Figure 1A. On the other hand, if the null hypotheses are false, then we expect the distribution of p-values to be right skewed: roughly, there should be fewer high p-values (near to .05) than low p-values. In this case, the p-curve distribution should look like Figure 1B. The more powerful the tests are, the more right skewed the distribution of p-values.

Furthermore, researchers who are p-hacking, for example, who are collecting data until they obtain statistical significance and then stop (a practice called 'optional stopping'), are trying to push p-values below the significance level, set by convention at .05. If researchers are p-hacking, we should then expect a larger number of p-values just at or below .05 than what would be expected by chance. P-hacking will thus produce a p-curve that is left skewed, as in Figure 1C.

A set of studies will only have the p-curve shape in Figure 1C if the null hypothesis is true and there is intense p-hacking. For any non-zero effect size, the p-curve will be right skewed, at least to some extent. Still, the shape of the p-curve can tell us how confident to be that a set of studies contains results that were p-hacked. For example, a U-shaped p-curve exhibits both a right and a left skew, indicating both a real effect *and* p-hacking (Simonsohn et al. 2014a, Head et al. 2015).

We can thus test whether a set of studies has evidential value by testing whether the distribution of significant p-values in a set of studies differs from the uniform distribution. If we can reject the null hypothesis that the p-curve is uniform and observe only a right skew, we can conclude that the set of studies under scrutiny has evidential value as a group.

One may object that a p-curve may be right skewed even if there is no evidential value among the p-curved findings. This could happen if there is a publication bias such that submitted articles are more likely to be accepted the lower the p-value for the relevant statistics. Given this possibility, we can at best conclude from a right-skewed p-curve that the significant findings in the p-curved literature are not due to p-hacking, not that this literature contains evidential value.

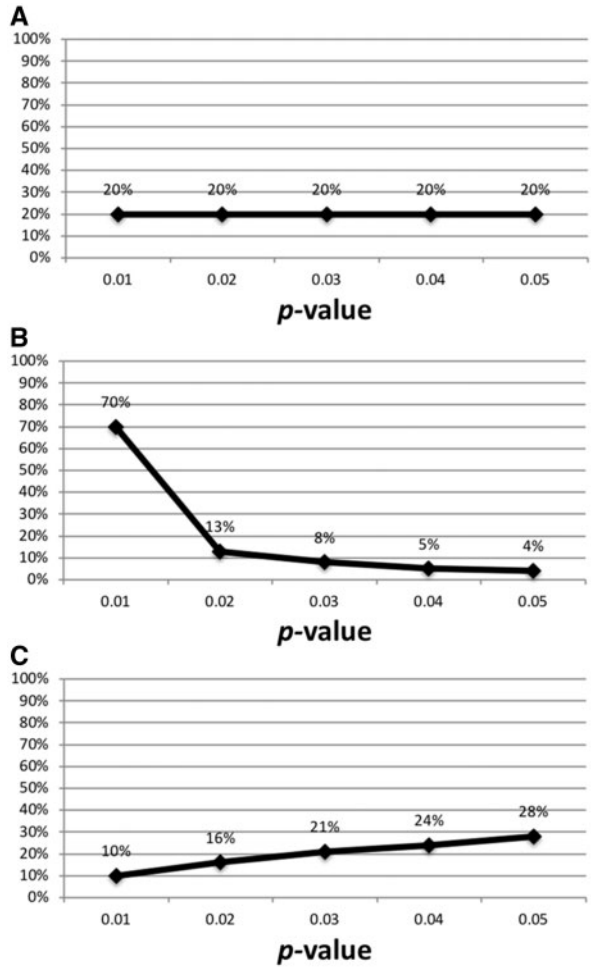


Figure 1. Different p-curves. (A) A p-curve with uniform distribution of p-values. (B) A 'right skewed' p-curve, displaying evidential significance. (C) A 'left skewed' p-curve, displaying evidence of p-hacking.

We respond to this concern in two ways. Simonsohn and colleagues (2014a) define the expression 'evidential value' as follows: 'We say that a set of significant findings contains evidential value when we can rule out selective reporting [p-hacking] as the sole explanation of those findings' (535). Our first response to the objection under discussion is then that we use this definition of evidential significance, that is, we take a right-skewed distribution of p-values to indicate that the p-curved literature has evidential value in the sense that this distribution is not best explained by the hypothesis of p-hacking (even though that set of studies might lack evidential value in the customary sense).

However, we can go further than this: a right-skewed p-curve indicates that the p-curved literature has evidential value in the sense that some of the tested null hypotheses in the p-curved literature must be false. Naturally, this inference is not deductive since, as discussed above, a p-curve can be right skewed even if there is no evidential value among the p-curved articles, but it is a good *inductive* inference all the same. The reason is that the publication bias discussed above is implausible: submitted articles are not much more likely to be accepted the lower the relevant p-values are because of the role of the significance level in publication decisions. Furthermore, we are aware of no other plausible defeating conditions of this inference.[1]

A p-curve analysis could fail to reject the null hypothesis that the p-curve is uniform for one of two reasons: the null hypotheses tested by a set of studies happen to be true or the power of the test is too low. To distinguish these two hypotheses, p-curve analysis introduces a distribution of p-values resulting from tests with a low power (0.33). If the p-curve is significantly flatter than this distribution, one concludes that the set of p-curved studies has no evidential value: either the null hypotheses are true, or the effects are too small to be measured (for a discussion of power analysis, see Machery 2012). If we cannot reject the null hypothesis that the p-curve is the distribution of p-values resulting from tests with a 0.33 power, we infer that our analysis has not enough power to conclude that a literature has no evidential value.

Finally, p-curve analysis can be applied to sets of studies investigating either a single null hypothesis or different null hypotheses. Thus, we are not limited to p-curving only the studies that investigate a hypothesis about a single effect.

2. *Materials and methods*

2.1 *Dataset coding*

The three authors of this article identified x-phi studies by searching philpapers.org (label 'Experimental Philosophy'), the Experimental Philosophy webpage at Yale (experimental-philosophy.yale.edu/ExperimentalPhilosophy.html), Google Scholar (using the names of all the experimental philosophers we know), the CVs and websites of experimental philosophers we know, online x-phi bibliographies (the 'Experimental Philosophy' and the 'Experimental Moral Philosophy' entries of the *Stanford Online Encyclopedia* and the 'Experimental Philosophy' entry of *Oxford Bibliographies*), and literature reviews in experimental philosophy (e.g. Knobe et al. 2012), and by examining all the issues of a long list of journals in philosophy and

1 Further, some x-phi works only report whether the p-value meets the significance level. While the actual value can be deduced from the reported statistics (as we have done in this study), it seems implausible that reviewers systematically compute p-values to the extent needed to result in this scenario.

psychology since 2001 (*Review of Philosophy and Psychology*, *Mind and Language*, *Philosophical Psychology*, *Analysis*, *Noûs*, *Philosophy and Phenomenological Research*, *Philosophers' Imprint*, *Synthese*, *Cognition*, *Cognitive Science*).

Three different coders were then trained to identify the date published (online first, when different dates existed), whether the paper was a replication or not, chapter or article, the number of authors, whether any of the authors was a non-philosopher, the number of studies in the paper and all the test statistics. The coders agreed on 92.6% of studies (excluding disagreements irrelevant to the p-curve). The authors of this article corrected entries that did not match.

2.2 Data collection

The corpus includes 365 works published between 1997 and 2017. Data collection was terminated in 2016, as our aim was to investigate three 5-year intervals between 2001 and 2016. Data from 2017 are not exhaustive; only works that were forthcoming in 2016 and published in 2017 are included. Data were collected from English-language publications only.

2.3 Inclusion criteria

To be included in our dataset, three criteria had to be met. First, the work had to include as author or co-author at least one individual with a PhD in philosophy or a cognate discipline (e.g. history and philosophy of science) and possess an affiliation to a philosophy department. This excludes works authored by psychologists or neuroscientists alone, even when philosophically relevant. Second, the work must either report novel data or analyse data previously collected by other philosophers in a new way. This excluded replication articles, or works where philosophers analysed data collected by non-philosophers. Third, the work must be explicitly framed as bearing on a philosophical issue. This excludes scientific works that happen to include collaborators that are philosophers (e.g. some work in the philosophy of science in practice).

2.4 Determination of main statistic(s) for a study

We recorded the results of both main and auxiliary statistical tests in each study. Main statistics were determined by relating them to the main hypotheses presented in each work. There were often more than one main statistic for a given study, and more than one study per paper. All other statistics were reported as 'other' in the dataset. P-curves are run on the main statistics only, because we are interested in the evidential status of the main conclusions (the main hypotheses tested). Because the p-values in the p-curve must be independent, we only used the statistic(s) coded as the main statistic(s) for our p-curves.

2.5 Designation of field of philosophy

We divided the corpus into eight categories in order to assign each of the 365 articles to a single field of philosophy. The categories were: metaphysics, epistemology, philosophy of action, philosophy of mind, ethics, philosophy of language, philosophy of science and 'other'. Papers in the 'other' category belong to fields that are less well-represented in x-phi, like philosophy of religion, race, aesthetics, medicine and metaphilosophy. Categorization was based on the topic of the paper. Categorization was not based on operational criteria, but on our own sense of the field in which the papers belong. Some decisions were made in advance of our categorization. For example, we had decided to categorize work on intentionality and free will into philosophy of action, given that this field is traditionally where these topics are investigated.

2.6 Negative versus positive programme

We divided the corpus into the so-called 'negative' and 'positive' programmes. The negative programme includes all papers meant only to undermine traditional philosophical methods such as the method of cases (O'Neill and Machery 2014, Machery 2017), the positive programme everything else. Categorization was not based on operational criteria, but on our own sense of the programme in which the papers belong.

2.7 P-Curve app

The p-curves were generated using version 4.06 of the p-curve app, created by Simonsohn, Nelson and Simmons. The app can be accessed here: http://www.p-curve.com/app4/, and its R-code here: http://p-curve.com/app4/pcurve_app4.06.r. For more details on the statistical underpinning of the app, or the logic supporting the statistics, see Simonsohn et al. 2014a, 2014b.

3. Results

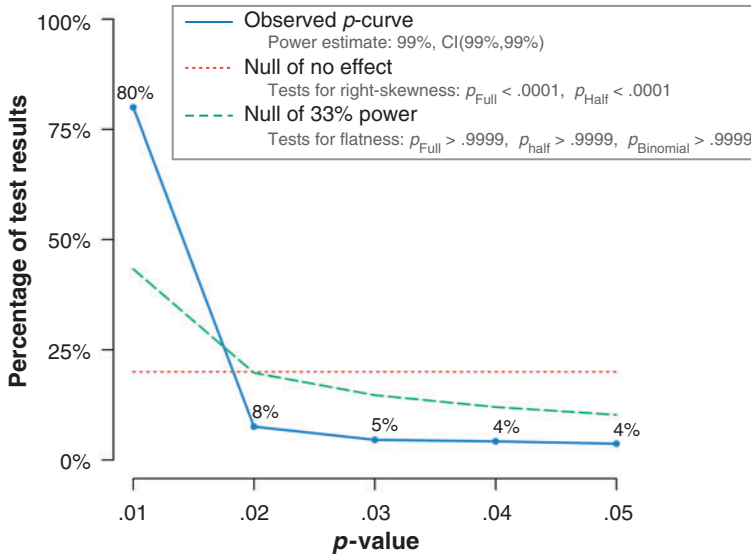
The skews of our p-curves provide insight into which areas of x-phi have findings that cannot be explained by selection bias or p-hacking. In each figure, the solid line is the observed p-curve, which is indicative of the corpus or relevant corpus subset. The hatched lines are the uniform distribution of p-values (narrow hatches) or the distribution of p-values resulting from a low-powered (0.33) test (wide hatches).

3.1 Overall

The overall p-curve for the corpus of x-phi is right skewed (Figure 2).

3.2 X-phi results over time

We then examined the corpus by time period in order to see whether the frequency of p-hacking changes over time. We have divided the corpus into



Note: The observed *p*-curve includes 569 statistically significant ($p < .05$) results, of which 516 are $p < .025$. There were 27 additional results entered but excluded from *p*-curve because they were $p > .05$.

Figure 2. P-curve for x-phi corpus as a whole.

three intervals: *early*, which includes all the articles and chapters published before 2006, *middle*, which includes those published from 2006 to 2010, and *later*, which includes those published in 2011 up to the most recent in our corpus. These intervals do not correspond to any transformation or discontinuity in x-phi, but they give us three groups of five or more year intervals, to determine how the field has changed over time, if at all.

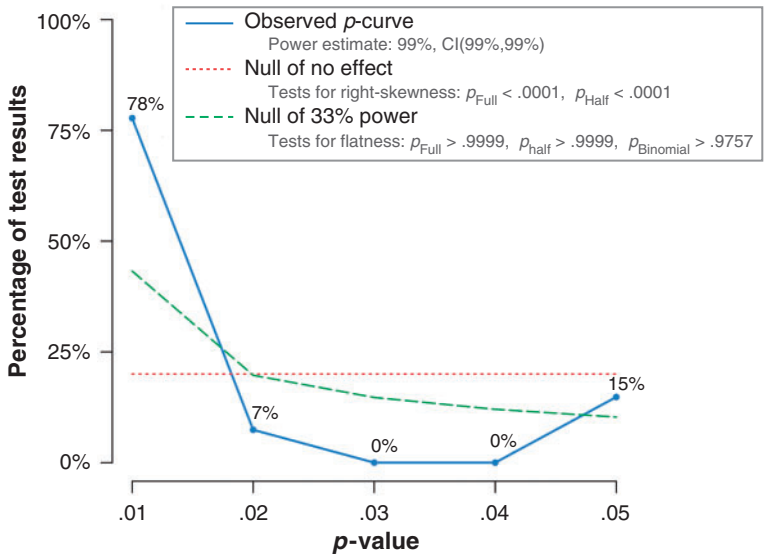
The p-curve for early x-phi has a U-shape, unlike that for the corpus as a whole (Figure 3). The right skew is more pronounced for middle x-phi (Figure 4), and even more so for later x-phi (Figure 5). Correspondingly, the frequency of high p-values is lower in the middle x-phi interval and even lower for the later x-phi interval.

3.3 Field of philosophy

The p-curve for each field of philosophy has a right skew (Supplementary Figure 1). However, the p-curves for ethics and epistemology have a tail of high p-values.

3.4 Collaborators

The p-curves for papers written by philosophers only as opposed to philosophers in collaboration with non-philosophers are both right skewed, and there are no important differences between them (Supplementary Figure 2).



Note: The observed *p*-curve includes 27 statistically significant ($p < .05$) results, of which 23 are $p < .025$. There were 15 additional results entered but excluded from *p*-curve because they were $p > .05$.

Figure 3. P-curve for studies published up to but not including 2006.

3.5 Small sample sizes

The p-curve for the studies with a sample size less than 20 is right skewed, but has also a large tail of high p-values (Figure 6).

For comparison, see Figure 7 for the p-curve for studies with a sample size between 500 and 1000.

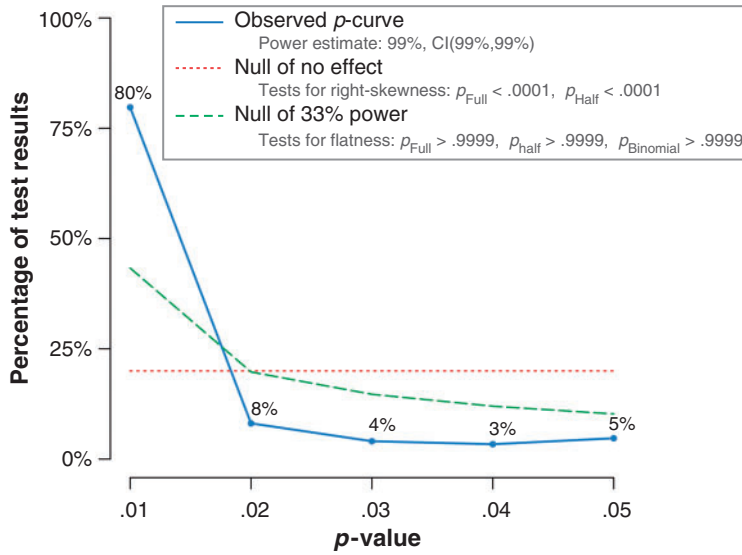
3.6 Negative versus positive x-phi

Negative and positive x-phi both have a right skew. Positive x-phi's p-curve has a high number of high p-values not present in the negative x-phi p-curve (Supplementary Figure 3). Also the proportion of low p-values is higher in the negative x-phi corpus.

In sum, all p-curves reported in this article are right skewed, although some p-curves have high p-value tails. Both of these can be taken into account when assessing the evidential value of the corpus, and the explanatory adequacy of selection bias and p-hacking for reported significant results.

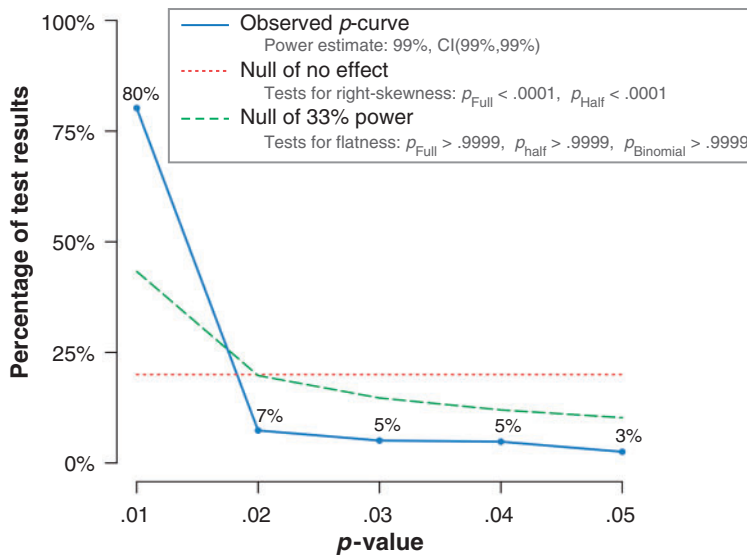
4. Discussion

Overall, the x-phi corpus fares well when p-curved. The right skew of p-values for the main statistics in the corpus (Figure 2) suggests that whatever selection bias and p-hacking might have occurred, it is not an adequate



Note: The observed *p*-curve includes 148 statistically significant ($p < .05$) results, of which 136 are $p < .025$. There were 4 additional results entered but excluded from *p*-curve because they were $p > .05$.

Figure 4. P-curve for the studies published in 2006 up to but not including 2011.



Note: The observed *p*-curve includes 394 statistically significant ($p < .05$) results, of which 357 are $p < .025$. There were 8 additional results entered but excluded from *p*-curve because they were $p > .05$.

Figure 5. P-curve for the studies published in 2011 up to the most recent studies in our dataset.

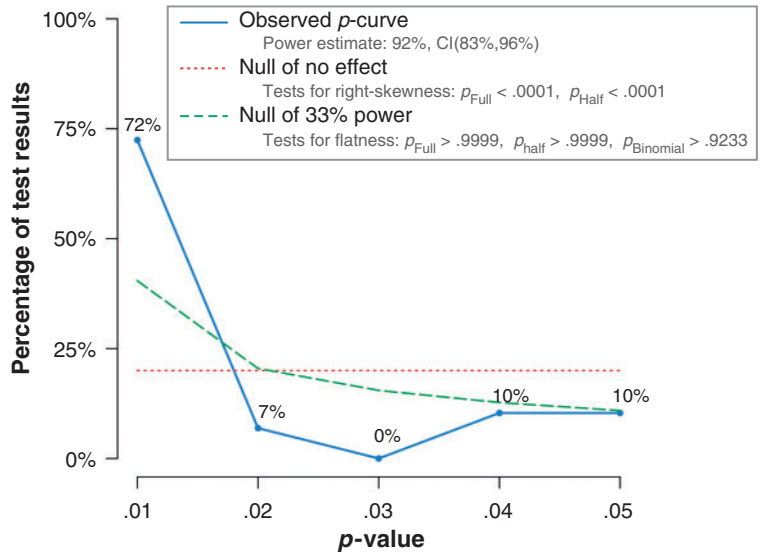
explanation for the results reported in the corpus as a whole. Thus, the corpus of x-phi results has evidential value (in the senses discussed in §1.2). Whatever other concerns one may have about the relevance of x-phi studies for philosophical problems, it cannot be dismissed by alleging that it suffers from the QRPs to which the p-curve is sensitive.

When our corpus is broken down into subsets, we see evidence of p-hacking in some of the resulting subsets. The p-curve for studies ran before 2006 (Figure 3) is somewhat left skewed, as the p-values of a large number of tests are just below 0.05. Thus, p-hacking probably explains some of the results in the first x-phi studies, although the high number of low p-values nevertheless suggests that many effects are real. The p-curve improves for 2006–10 (Figure 4), with a greater right skew and a lessened left skew, suggesting that p-hacking is less common. The p-curve for 2011–17 is also right skewed and shows no trace of p-hacking (Figure 5). We conclude that experimental philosophers have identified real effects in many of their studies, and that their empirical methodologically has improved: p-hacking appears to diminish with time. Experimental philosophers may have become more aware of the perils of QRPs, or their experiments may have become increasingly well-designed.

Turning now to the sub-disciplines within philosophy, metaphysics, language, mind and action all have evidential value (Supplementary Figure 1). Ethics and epistemology, though right skewed, have a slight tail of high p-values, suggesting that some work in experimental ethics and epistemology suffers from p-hacking (Supplementary Figure 1). We found no difference between the studies conducted solely by philosophers and those co-authored with non-philosophers (Supplementary Figure 2). This undermines the suggestion that experimental philosophers engage in QRPs because of their lack of training in psychology (Williamson 2010). Unsurprisingly, studies with fewer participants (fewer than 20) look to be p-hacked, although they still have some evidential value (Figure 6, and compare Figure 7). Negative x-phi comes out looking better than positive x-phi: both literatures have evidential significance, but the p-curve of negative x-phi is less indicative of p-hacking than that of positive x-phi (Supplementary Figure 3). This difference may be explained by the sample sizes of studies in negative and positive x-phi, rather than any difference in the nature of these studies, their targets or their implications.[2]

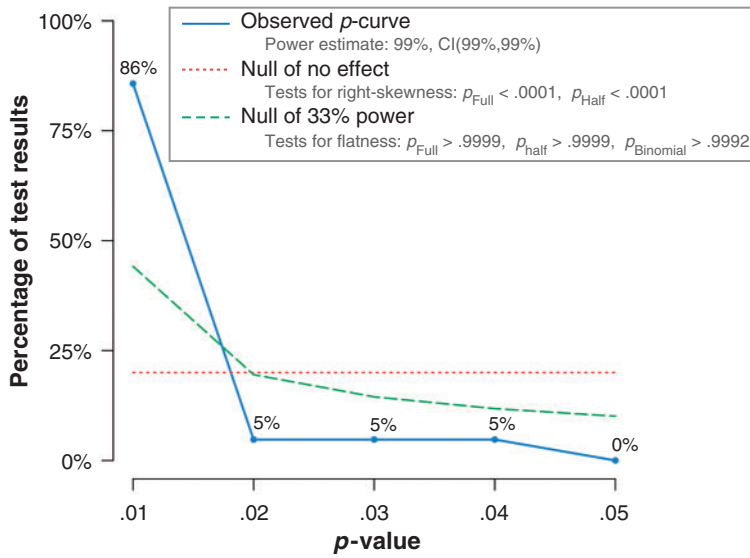
Currently p-curves are limited in two ways. First, they do not work with discrete test statistics (e.g. difference of proportions tests). We had 83 instances of such tests in our data set, although only 19 of these were main results (out of a total of 1030 main results). So this is not a serious issue. Second, the p-curve is naturally 'pessimistic', in the sense that it is more likely

2 We also found no difference in articles compared with book chapters (p-curves not reported for considerations of space).



Note: The observed *p*-curve includes 29 statistically significant ($p < .05$) results, of which 23 are $p < .025$. There were no non-significant results entered.

Figure 6. P-curve for the studies with a sample size less than 20.



Note: The observed *p*-curve includes 21 statistically significant ($p < .05$) results, of which 20 are $p < .025$. There were no non-significant results entered.

Figure 7. P-curve for studies with sample size between 500 and 1000.

to present corpora as lacking evidential value than having it (Simonsohn et al. 2014a: 546). Since each p-curve displayed evidential value, this pessimism has not been a concern.

5. Conclusion

To investigate the evidential value and presence of QRPs in x-phi, we developed a corpus of studies and performed p-curve analyses on them. Our findings indicate that, both as a whole and within all subsets, the corpus has evidential value. P-hacking has probably occurred in a few areas of x-phi, in particular, during its first years, even if the effects reported are, on the whole, genuine. We are optimistic about the trajectory of methodological progress we have identified, and hope to see it continue.[3]

Funding

This work was supported by the Center for Philosophy of Science at the University of Pittsburgh, and the Social Sciences and Humanities Research Council of Canada.

Supplementary data

Supplementary data is available at *ANALYSIS* online.

University of Geneva
Rue de Candolle 2
CH-1211 Genève 4
Switzerland
mike.stuart.post@gmail.com

Department of History and Philosophy of Science
University of Pittsburgh 1101 Cathedral of Learning
4200 Fifth Avenue
Pittsburgh, PA 15260
USA

Center for Philosophy of Science
University of Pittsburgh
1117 Cathedral of Learning, 4200 Fifth Avenue
Pittsburgh, PA 15260
USA

3 We are grateful to the two referees and an Associate Editor of *Analysis* for helpful remarks, as well as Matteo Colombo, Florian Cova and Uri Simonsohn.

References

Adleberg, T., M. Thompson and E. Nahmias. 2015. Do men and women have different philosophical intuitions? Further data. *Philosophical Psychology* 28: 615–41.

Bakker, M. and J.M. Wicherts. 2011. The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods* 43: 666–78.

Cappelen, H. 2012. *Philosophy without Intuitions*. Oxford: Oxford University Press.

Colaço, D. and E. Machery. 2017. The intuitive is a red herring. *Inquiry* 60: 403–19.

Colombo, M., G. Duev, M.B. Nuijten and J. Sprenger. 2018. Statistical reporting inconsistencies in experimental philosophy. *PLoS One* 13: e0194360.

Cova, F., B. Strickland, A. G. F. Abatista, A. Allard, J. Andow, M. Attie, J. Beebe, R. Berniūnas, J. Boudesseul, M. Colombo, F. Cushman, R. Díaz, V. Dranseika, B. D. Earp, T. Emile, I. R. Hannikainen, J. V. Hernández-Conde, W. Hu, F. Jaquet, K. Khalifa, H. Kim, M. Kneer, J. Knobe, M. Kurthy, A. Lantian, S. Liao, E. Machery, T. Moerenhout, C. Mott, M. Phelan, J. S. Phillips, N. Rambharose, K. Reuter, F. Romero, P. Sousa, J. Sprenger, K. Tobia, A. G. Torres, N. van Dongen, H. Viciana, D. A. Wilkenfeld, X. Zhou. 2018. *The XPhi Replicability Project*. doi: 10.17605/OSF.IO/DVKPR.

Cullen, S. 2010. Survey-driven romanticism. *Review of Philosophy and Psychology* 1: 275–96.

Deutsch, M.E. 2015. *The Myth of the Intuitive: Experimental Philosophy and Philosophical Method*. Cambridge: MIT Press.

Head, M.L., L. Holman, R. Lanfear, A.T. Kahn and M.D. Jennions. 2015. The extent and consequences of p-hacking in science. *PLOS Biology* 13: e1002106.

Kim, M. and Y. Yuan. 2015. No cross-cultural differences in the Gettier car case intuition: a replication study of Weinberg et al. 2001. *Episteme* 12: 355–61.

Knobe, J., W. Buckwalter, S. Nichols, P. Robbins, H. Sarkissian and T. Sommers. 2012. Experimental philosophy. *Annual Review of Psychology* 63: 81–99.

Machery, E. 2012. Power and negative results. *Philosophy of Science* 79: 808–20.

Machery, E. 2017. *Philosophy within Its Proper Bounds*. Oxford: Oxford University Press.

Machery, E., S.P. Stich, D. Rose, A. Chatterjee, K. Karasawa, N. Struchiner, S. Sirker, N. Usui and T. Hashimoto. 2017. Gettier across cultures. *Noûs* 51: 645–64.

Mahowald, K., A. James, R. Futrell and E. Gibson. 2016. A meta-analysis of syntactic priming in language production. *Journal of Memory and Language* 91: 5–27.

Medina, J. and S. Cason. 2017. No evidential value in samples of transcranial direct current stimulation (tDCS) studies of cognition and working memory in healthy populations. *Cortex* 94: 131–41.

Nuijten, M.B., C.H. Hartgerink, M.A. van Assen, S. Epskamp and J.M. Wicherts. 2016. The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods* 48: 1205–26.

O'Neill, E. and E. Machery. 2014. Experimental philosophy: what is it good for? In *Current Controversies in Experimental Philosophy*, eds. E. Machery and E. O'Neill, vii–xxix. New York: Routledge.

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349: aac4716.

Scholl, B. MS. Two kinds of experimental philosophy (and their experimental dangers).

Segerstrom, S.C. and G.E. Miller. 2013. Applying the p curve to PNI: are effects of stress on immunity real or incidental? *Brain, Behavior, and Immunity* 32: e5.

Seyedsayamdost, H. 2015a. On normativity and epistemic intuitions: failure of replication. *Episteme* 12: 95–116.

Seyedsayamdost, H. 2015b. On gender and philosophical intuition: failure of replication and other negative results. *Philosophical Psychology* 28: 642–73.

Simmons, J.P. and U. Simonsohn. 2017. Power posing: p-curving the evidence. *Psychological Science* 28: 687–93.

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

Simonsohn, U., L.D. Nelson and J.P. Simmons. 2014a. P-curve: a key to the file-drawer. *Journal of Experimental Psychology: General* 143: 534–47.

Simonsohn, U., L.D. Nelson and J.P. Simmons. 2014b. P-curve and effect size correcting for publication bias using only significant results. *Perspectives on Psychological Science* 9: 666–81.

Simonsohn, U., J.P. Simmons and L.D. Nelson. 2015. Better p-curves: making p-curve analysis more robust to errors, fraud, and ambitious p-hacking, a reply to Ulrich and Miller. *Journal of Experimental Psychology: General* 144: 1146–52.

Stuart, M. 2015. Philosophical conceptual analysis as an experimental method. In *Meaning, Frames and Conceptual Representation*, eds. T. Gamerschlag, D. Gerland, R. Osswald and W. Petersen, 267–92. Düsseldorf: Düsseldorf University Press.

Vadillo, M.A., N. Gold and M. Osman. 2016. The bitter truth about sugar and willpower: the limited evidential value of the glucose model of ego depletion. *Psychological Science* 27: 1207–14.

Vazire, S. 2015. Experimental philosophy: the view from social/personality psychology. *Annals of the Japan Association for Philosophy of Science* 23: 45–52.

Williamson, T. 2010. Philosophy vs. imitation psychology. *New York Times*, August 19. <http://www.nytimes.com/roomfordebate/2010/08/19/x-phis-new-take-on-old-problems/philosophy-vs-imitation-psychology?> last accessed 31 January 2019.

Woolfolk, R.L. 2011. Empirical tests of philosophical intuitions. *Consciousness and Cognition* 20: 415–6.

Woolfolk, R.L. 2013. Experimental philosophy: a methodological critique. *Metaphilosophy* 44: 79–87.

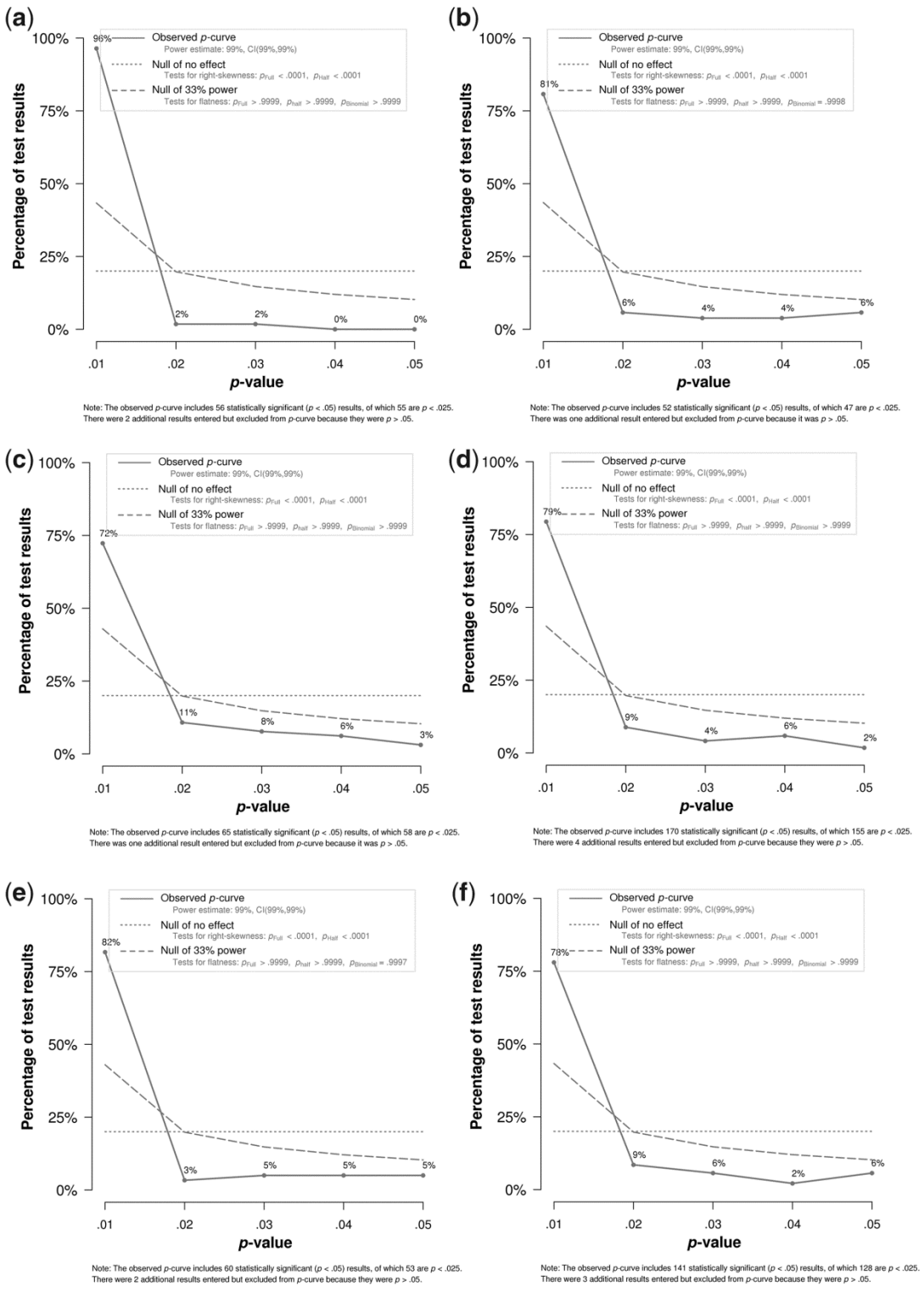
Reflective blindness, depression and unpleasant experiences

ELIZABETH VENTHAM

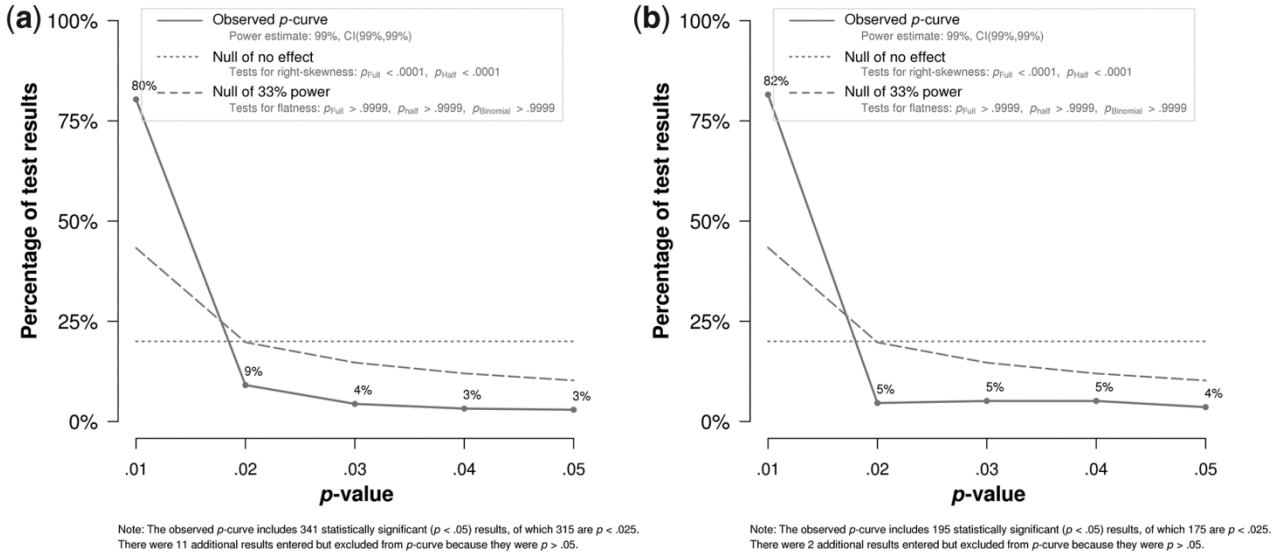
1. Introduction

This paper defends a desire-based understanding of pleasurable and unpleasant experiences. More specifically, the thesis is that what makes an

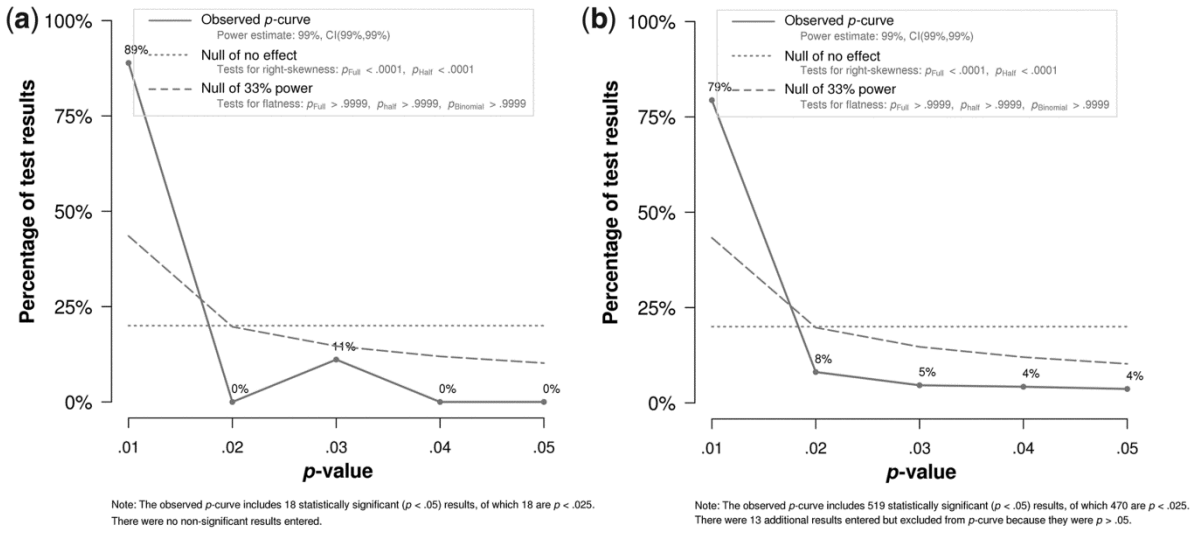
Analysis Vol 79 | Number 4 | October 2019 | pp. 684–693 doi:10.1093/analys/any093



Supplementary figure S1: P-curves for the fields of philosophy: (a) metaphysics, (b) epistemology, (c) language, (d) action, (e) mind and (f) ethics.



Supplementary figure S2: P-curve for the studies involving (a) only philosophers and (b) philosophers in collaboration with non-philosophers.



Supplementary figure S3: P-curve for (a) negative x-phi and (b) positive x-phi.