Experimental Philosophy: The View from Social/Personality Psychology

Simine VAZIRE*

The emergence of experimental philosophy is fascinating to watch from the vantage point of social/personality psychology. Although experimental philosophy is quite broad, much of it tackles questions that are fundamentally about social psychology and personality. Understanding how social factors influence people's reasoning about morality, agency, and knowledge is, essentially, a social psychological question. Understanding how and why people differ from each other in their moral behavior and reasoning, their beliefs about free will, and their understanding of the world – these are questions about personality and individual differences. However, as philosophers discovered when they went in search of empirical evidence¹ on these topics, there are important gaps in the empirical literature on these topics. Experimental philosophy has done our field a great service by steering our attention towards some new, fundamental questions about human behavior and cognition. We have much to learn from the millennia of philosophical work on these questions.

Of course, there is much that experimental philosophers can learn from social/personality psychologists as well. Contrary to what one might think, however, my advice is not for experimental philosophers to do as we social/personality psychologists do. Rather, I implore experimental philosophers to learn from our mistakes and not make the same errors that we have.

As readers may be aware, there is a crisis sweeping across the sciences. This crisis is about the integrity of our research findings. Although concerns about fraud have helped fuel this crisis, the scope is much broader and raises concerns about very common research practices that scientists openly practiced, endorsed, and even taught their students. The concern is that many of the techniques considered standard operating procedure actually corrupt the evidence and increase the chances of false positives. The techniques are somewhat different in each discipline. The goal

* University of California, Davis

¹ Note the use of the word “empirical” rather than “experimental.” “Experimental philosophy” is actually the wrong term, it should be called “Empirical Philosophy”. Many of the studies conducted by experimental philosophers are not actually experiments (i.e., they are correlational studies), and that's great (I love correlational studies), but “experimental philosophy” is an unnecessarily narrow label for all these studies.

of this article is to highlight some of the common practices in psychology that, we are now realizing, 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. In my 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.

Overview of the problem

At the root of many of the problems is an overzealous pursuit of “ $p < .05$ ”. In Null Hypothesis Significance Testing (NHST), a result is “statistically significant” if you obtain a *p*-value less than .05. This is, in itself, not a terrible model (though I do think it’s pretty bad). The bigger problem is the single-minded pursuit of $p < .05$ (for definitions and common misconceptions, see Cumming, 2010). Sometimes it almost seems that psychologists have gotten so obsessed with *p*-values that we behave as if getting $p < .05$ is the goal of research, rather than finding (or approximating) the truth. In the service of getting the desired *p*-value, or “uncovering” it, as we sometimes tell ourselves we are doing, we torture our data until they tell us what we want to hear. Of course, this kind of motivated search for $p < .05$ (now you can see why it’s called “*p*-hacking”) leads to a greater proportion of false positives and unreplicable results.

Why do we need to torture our data to get $p < .05$, you might be wondering. Is it because we are bad armchair philosophers and come up with hypotheses that are actually wrong? Perhaps, but not necessarily. Even if our hypotheses are often right and we are looking for effects that are really there, we are not designing our studies in such a way that we are likely to detect them. What do I mean by this? Our likelihood of detecting an effect (i.e., concluding that it is “statistically significant” under NHST) is a function of a few things: how large the effect is, how good our instruments are (how little error there is in our measurement), and how many people (or observations) are in our study. We often can’t know a priori how large the effect we’re looking for is likely to be, but we can estimate that by looking at the distribution of effect sizes in the subdiscipline we are working in. In social and personality psychology, the average effect size is $r = .21$, which is equivalent to Cohen’s $d = .43$ (Richard, Bond, & Stokes-Zoota, 2003). Thus, we should probably design our studies so that, if the effect is there and is of typical magnitude, we would have a decent chance of detecting it. Happily, there are formulas, online calculators, and free software (<http://www.psych.uni-duesseldorf.de/abteilungen/aap/gpower3/>) that can calculate this probability for you (this probably is called “statistical power”). As it turns

out, to have an 80% chance of detecting an effect of $r = .21$ in a simple correlational study (or a two-independent-groups design), you need roughly 180 people (or 90 per group). If you think your effect might be on the small side (e.g., $r = .10$), and you want an 80% chance of detecting it (the commonly accepted minimum), you would need roughly 800 people (or 400 per group). In contrast, surveys of studies published in social and personality journals reveal that the median sample size is much lower. For example, in a survey of articles published in the field's most prestigious journal, the *Journal of Personality and Social Psychology*, the average sample size was 90 (Fraley & Vazire, 2014). This corresponds to a 48% chance of detecting a true effect of $r = .21$, and 16% chance of detecting a true effect of $r = .10$! No wonder we need to torture our data to get statistically significant results. Our hypotheses might be right, but we are not testing them in a way that can tell us that.

The “questionable” practices

If most of our studies are “under-powered” (do not have a large enough sample size to have a high chance of detecting typically-sized effects), but the vast majority of published studies find statistically significant results (Ioannidis, 2005), what is going on? What is going on is that many of us are manipulating our data so that they produce a very large effect – much larger than the effect likely is in the real world. Large effects reach statistical significance with much smaller samples (e.g., for $r = .50$ to be significant, you only need 16 people!). So if we can just get our data to show us a large effect, it will be statistically significant. How can we get our data to show us a large effect? There are several techniques that were developed – probably in good faith – that, as it turns out, lead to inflated effect sizes. Many (if not all) of these techniques are sometimes justified! What happens is that we deploy them without critically evaluating whether they are justified. And when we combine these techniques, the inflation of effect sizes is compounded. So without realizing it (and with some very easy rationalization), we can turn a dataset that initially told us that the effect was around $r = .18$ (not statistically significant with our measly $N = 80$) into one that tells us that r is $.40$ – statistically significant! For a more extensive review of the various practices researchers use, see Simmons, Nelson, and Simonsohn (2011) and John, Loewenstein, and Prelec, (2012).

Below I review some of the more common techniques that I think are likely to tempt experimental philosophers. What makes these techniques tempting? First, they may seem harmless, or even like prudent steps to take. Second, they are techniques that are easy to deploy when running simple experiments or correlational studies (e.g., studies with a few vignettes, with some self-report questionnaires, with mTurkers, etc.) because these designs often give researchers a great deal of flexibility. Flexibility can be great, but that flexibility can also be exploited to find what

you are looking for. Some people call this flexibility “researcher degrees of freedom” (Simmons, Nelson, & Simonsohn, 2011). The idea is that if you are only looking for confirmation of your hypothesis, and willing or able to ignore/rationalize disconfirmatory results, this flexibility is a liability. Of course, I would hope that philosophers would be less likely than others to fall prey to this unscientific thinking (philosophy of science 101: seek disconfirmation!), I worry that experimental philosophers will, reasonably enough, assume that we psychologists know what we’re doing and will follow the bad examples we have set. So do as we say, not as we do, please.

1. Peeking at your results as your data are coming in.

Let’s say you run a few participants, check your data to see how your results look, and if the result is not statistically significant, run a few more participants. Lather, rinse, repeat. Sounds pretty harmless, but in fact this seriously increases your chances of producing a false positive. How? It puts you at risk of capitalizing on chance. Imagine that the data coming in are, unbeknownst to you, being generated by a random number generator (i.e., the data come from a population where there is zero correlation, or zero difference between groups, because there are no systematic patterns in the source of the data). Now, if you collect 10 data points, examine the data for your hypothesized effect, and repeat this over and over again, once in a while you will get a statistically significant result just by chance (5% of the time, to be precise, because of the $p < .05$ threshold we use to determine statistical significance). If you stopped when you got that statistically significant effect, and reported that finding, it would be a false positive. The same goes for the data you collect from real participants. If the p -value occasionally dips below .05, but does not consistently stay there, it’s probably a random fluctuation and you should not believe it.

2. What if we try....

You’ve collected your data, run your analyses, and your p -value didn’t quite reach below the .05 threshold. Don’t despair! There are still techniques available to p-hack that result down to $< .05$. I call this category of techniques the “What if we try...” category, because they are techniques that are often perfectly justified when they were planned a priori, but when used post-hoc lead to more false positives.

A. Dropping outliers.

Check and see if maybe a couple of pesky people are ruining everything! Of

course it is often acceptable to drop outliers, but when it is done post-hoc (and especially if you try various combinations of outliers to drop), you increase the risk of producing a false positive. What do I mean by that? Imagine there is no effect in your data (zero correlation, or zero difference between groups). Presumably if you dropped exactly the right combination of people from your study, you could produce a dataset in which there is the effect you expected. Now, you'd probably have to go farther than just dropping a few outliers, but trying out different outliers to drop is a less extreme version of this. You are sculpting your data and you run the risk of turning what is actually a random distribution of scores into an ordered pattern that you want to see.

B. Adding covariates.

For the same reason that dropping outliers post-hoc increases the chance of false positive findings, adding covariates post-hoc (and especially trying out various covariates) does, too. Again, covariates are often justified and planned a priori, but trying out covariates after you have started analyzing your data exposes you to the same risks as dropping outliers post-hoc. So think twice before you decide to control for gender, ethnicity, major, SES, etc.

C. Dropping/adding measures.

Sometimes researchers include several different measures of a variable of interest. This is perfectly understandable. However, when it comes time to analyze the data, they try out the analyses separately with each measure (and if they used multiple measures for several different variables, you can imagine that the permutation of possible analyses gets big fast).

The solution(s)

1. Don't be a slave to the p-value

Criticisms of null hypothesis significance testing have been around for decades. However, there is finally some movement away from NHST, or at least towards complementing NHST with other approaches. Currently, the most popular approach is to focus on effect estimation and confidence intervals (Cumming, 2011). Meta-analysis is also gaining popularity, though there are important criticisms of the traditional approach to meta-analysis, suggesting that this approach does not adequately correct for publication bias (Simonsohn, Nelson, & Simmons, 2014).

2. Don't underpower all your studies

Perhaps the most important thing researchers can do to improve the validity of

their results is to use larger samples. The rules of thumb are constantly shifting, but a minimum of 100 participants per condition (or 200 participants total for a correlational study) is currently considered reasonable. Increasing sample size helps reduce both Type I (indirectly) and Type II error (Fraley & Vazire, 2014). Larger samples also make it harder to capitalize on chance – dropping outliers and other ‘questionable research practices’ have a smaller influence on a large sample. Finally, if all studies have large samples, it will be more difficult for researchers to ‘file drawer’ the studies with results they don’t like.

3. Don’t engage in QRPs

This one sounds easy. But what’s tricky about avoiding questionable research practices is that they are sometimes warranted. So knowing when to avoid them is difficult, and requires abstaining from wishful thinking. A researcher once joked that we should make researchers wear body cams while conducting data analyses. While this is obviously impractical, it’s a useful idea – analyze your data as if you were wearing a bodycam, and each decision you make can and will be scrutinized.

4. Don’t become obsessed with mediation

Psychologists have become obsessed with identifying the processes underlying their effects, and this is often done using a statistical technique called ‘mediation’ (which is little more than a series of regressions). But we have lost sight of several key points. First, the effects are often interesting in themselves, even if we don’t know the process. Second, conducting mediation analyses does not magically tell us about the process. The value of mediation analyses (especially with correlational data) comes from the strength of the theory about the causal pathway. If your theory is wrong, mediation won’t necessarily tell you that. This can be helped a bit by testing alternative models and ruling out confounds, but in the end, mediation analyses with correlational data cannot conclusively tell you about the causal process.

5. Replicate.

Another very important step researchers can take is to replicate their findings before submitting them for publication. If you run a follow-up study that includes an extension of your original result, include the measures necessary to also directly replicate your original result. If possible, pre-register your replication study (or even your original) study, for example on the Open Science Framework.

6. Report all results, including null results.

This is one is tricky because many journals still don’t want to see all of your failed studies. It’s also tricky because submitting a paper with some ‘successful’ studies (i.e., p -values below .05) and some ‘failed’ studies is questionable in itself –

you should provide evidence that adds to our knowledge, not a bunch of information that just confuses things further. So that means you should not submit a paper for publication until you have collected enough data that the results clearly point to one conclusion or another. This means making sure your studies have enough power (i.e., a large enough sample) to tell you something (i.e., small studies are more likely to lead to ambiguous results, because you have much less precision in your results), and doing enough studies and replications that you are confident in the result. However, it is good practice (and a sign of good science; Schimmack, 2012) to publish all studies even if every single study is not perfectly consistent with the overall conclusion.

7. Use multiple methods, not just mTurk.

Another way to bolster your own and others' confidence in your results is to test them using a variety of methods and designs. That means going beyond self-reports, going beyond mTurk, and even going beyond lab studies. To the extent possible, test your effects with non-self-report methods, in the real world.

Conclusion

Things are changing in psychology. These changes are for the better – we are finally heeding the calls to move away from the mindless pursuit of $p < .05$, and starting to use stronger methods, larger samples, and more replication. This is a great time to be learning psychological research methods, because you will be learning along with many of us who are re-training ourselves. There are also many great resources for keeping up with the latest changes, starting with the Center for Open Science (centerforopenscience.org; Open Science Collaboration, in press).

References

- Cumming, G. (2010). Understanding, teaching and using p values. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the 8th International Conference on Teaching Statistics*, Ljubljana, Slovenia. Voorburg, The Netherlands: International Statistics Institute.
- Cumming (2011). *Understanding the New Statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.
- Fraley, R. C., & Vazire, S. (2014). The N-Pact Factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLOS One*. DOI: 10.1371/journal.pone.0109019
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2, e124. doi: 10.1371/journal.pmed.0020124
- John, L. K., Loewenstein, G., & Prelec (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23(5), 524–532. doi: 10.1177/0956797611430953

- Open Science Collaboration (in press). Maximizing the reproducibility of your research. In S. O. Lilienfeld & I. D. Waldman (Eds.), *Psychological Science Under Scrutiny: Recent Challenges and Proposed Solutions*. New York, NY: Wiley.
- Richard, F. D., Bond Jr., C. F., Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, 7, 331–363.
- Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, 17, 551–566. doi: 10.1037/a0029487
- Simmons, J., Nelson, L., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as “significant”. *Psychological Science*, 22, 1359–1366. doi: 10.1177/0956797611417632
- Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve and effect size: Correcting for publication bias using only significant results. Manuscript in preparation.

(Received 2014.1.27; Revised 2014.12.31; Accepted 2015.1.5)