Commentary

Psychological Science
2017, Vol. 28(5) 687–693
© The Author(s) 2017
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797616658563
www.psychologicalscience.org/PS

Power Posing: *P*-Curving the Evidence

Joseph P. Simmons and Uri Simonsohn
University of Pennsylvania

Received 7/6/15; Revision accepted 6/16/16

In a well-known article, Carney, Cuddy, and Yap (2010) documented the benefits of *power posing*. In their study, participants ($N = 42$) who were randomly assigned to briefly adopt expansive, powerful postures sought more risk, had higher testosterone levels, and had lower cortisol levels than those randomly assigned to adopt contractive, powerless postures. This result has led some individuals to recommend power posing as a way to improve performance and life outcomes (e.g., Blodget, 2013; Cuddy, 2012).

Despite the attention Carney et al.'s (2010) study has received, there had until recently been no attempts to closely replicate its methods. Ranehill et al. (2015), using a larger sample ($N = 200$) and similar but not identical procedures, found that although adopting powerful postures led to self-reported increases in feelings of power (thus verifying the effectiveness of Carney et al.'s manipulation), it did not affect participants' behavior or hormonal levels.[1]

In their response to the failed replication, Carney, Cuddy, and Yap (2015) reviewed 33 successful studies investigating the effects of expansive versus contractive posing, focusing on differences between these studies and the failed replication to identify possible moderators that future studies could explore. But before spending valuable resources on that, it is useful to establish whether the literature that Carney et al. (2015) cited actually suggests that power posing is effective.

It may seem that the existence of 33 supportive published studies is enough to conclude that there is an effect of expansive versus contractive posture on psychological outcomes. However, one needs to account for selective reporting. If results get published only when they show an effect, the fact that all the published evidence shows an effect is not diagnostic (see, e.g., Pashler & Harris, 2012).

In this Commentary, we rely on *p*-curve analysis to answer the following question: Does the literature reviewed by Carney et al. (2015) suggest the existence of an effect once one accounts for selective reporting? We conclude that it does not. The distribution of *p* values from those 33 studies is indistinguishable from what would be expected if (a) the average effect size were zero and (b) selective reporting (of studies or analyses) were solely responsible for the significant effects that were published.

Our results do not imply, nor could they imply, that the effect size examined in these studies is exactly zero. It is possible that it is undetectably small in the predicted direction, say $r = .03$, or in the unpredicted direction, say $r = -.03$. But *p*-curve's estimates are precise enough to allow one to reject effects that would have been detectable in the power-posing studies cited by Carney et al. (2015). Thus, what the results do imply is that direct replications of these studies would not be expected to succeed.

The next three sections give an overview of selective reporting and *p*-curve analyses. Readers familiar with these topics may safely skip ahead to the Results section.

Selective Reporting

Statistically significant results are more likely to be published than results that are not significant (Greenwald, 1975; Rosenthal, 1979; Sterling, 1959; Sterling, Rosenbaum, & Weinkam, 1995). Selective reporting comes in at least two forms. One form, *file-drawering* (Rosenthal, 1979), involves the selective reporting of individual studies that are statistically significant. For example, a researcher may run five studies investigating the same effect but then only report the one study that achieved statistical significance, keeping the remaining four in the file drawer. (Or equivalently, five researchers may each run one study, but only the researcher who obtains a $p < .05$ publishes it.)

The other form of selective reporting is known as *p-hacking* (Simonsohn, Nelson, & Simmons, 2014a), which consists of conducting alternative analyses on the same data set and then selectively reporting those that provide statistically significant support for a publishable

Corresponding Author:
Joseph P. Simmons, The Wharton School, University of Pennsylvania, 500 Huntsman Hall, Philadelphia, PA 19104
E-mail: jsimmo@wharton.upenn.edu

claim. For example, researchers may attempt to control for different variables, to exclude participants they had previously included, to log-transform the dependent variable, to analyze a few more (or fewer) participants than planned, etc., until reaching a $p < .05$ (Simmons, Nelson, & Simonsohn, 2011).

Both forms of selective reporting threaten the validity of the published literature by hiding from view unsupportive (nonsignificant) results. This leads one to mistakenly conclude that an effect is larger than it actually is, or even that an effect is real when it actually is not.

A variety of statistical techniques exist to determine whether selective reporting is present in a literature (Egger, Smith, Schneider, & Minder, 1997; Ioannidis & Trikalinos, 2007; Rothstein, Sutton, & Borenstein, 2005). These tools can be used to answer the question, are there some studies or results we are not observing in this literature? (Francis, 2014; Ioannidis, 2011; Schimmack, 2012). However, they cannot be used to answer what is, in our view, the more important question: Once one accounts for selective reporting, do the observed results suggest that the effect is real? Answering *this* question requires correcting for selective reporting rather than just diagnosing its existence.

The most common approach to correcting for selective reporting is the *trim-and-fill* procedure (Duval & Tweedie, 2000). Unfortunately, it performs very poorly, often leaving estimates nearly as biased as the uncorrected estimates were. For example, Simonsohn, Nelson, and Simmons (2014b, Fig. 2) showed that when a nonexistent effect (Cohen's $d = 0$) is studied with predetermined per-cell sample sizes between 5 and 35 (and there is no *p*-hacking), the average statistically significant estimate is $\hat{d} = 0.72$. The trim-and-fill procedure lowers that estimate only to $\hat{d} = 0.70$. A less well-known method is PET-PEESE (precision-effect test and precision-effect estimate with standard error; Stanley & Doucouliagos, 2014). It too performs poorly. For example, Gervais (2015) simulated a literature in which half the studies investigated a true effect size of $d = 0.40$ and half investigated a true effect size of $d = 0.80$. PET-PEESE estimated the true effect to be zero.[2] In our view, the use of these methods should be discontinued.

P-Curve Analysis

In Simonsohn et al. (2014a), we introduced *p*-curve analysis, a statistical tool that tests whether a set of findings contains *evidential value*. A set of findings contains evidential value if one can statistically rule out that selective reporting was solely responsible for the statistically significant results that have been observed. *P*-curve analysis can also be used to obtain a selective-reporting-corrected estimate of the average statistical power of a set of studies (Simonsohn et al., 2014b).

P-curve is the observed distribution of statistically significant *p* values testing the hypothesis of interest from a set of studies (i.e., $ps \leq .05$). Its shape is diagnostic of evidential value.

In the absence of *p*-hacking, we expect studies investigating a nonexistent (i.e., zero) effect to result in a flat (uniform) *p*-curve. To understand why, consider that when the null hypothesis is true, there is a 5% chance of observing a $p < .05$, a 4% chance of observing a $p < .04$, a 3% chance of observing a $p < .03$, and so on. This means there is a 1% chance of observing a $p < .01$, a 1% chance of observing a *p* value between .01 and .02, a 1% chance of observing a *p* value between .02 and .03, and so on.

This is what would be expected if the effect were zero in all studies and if *p*-hacking were absent from all studies. When *p*-curve analysis includes some effects that exist (i.e., some nonzero effects), *p*-curve is expected to be right-skewed, with more low significant *p* values (e.g., .01s) than high significant *p* values (e.g., .04s). Thus, if at least some of the studies in a literature are actually investigating a true effect, then more of the critical *p* values will be very significant (e.g., .01s) rather than barely significant (e.g., .04s). For example, if one conducts *p*-curve analysis on a literature in which half of the studies with statistically significant findings investigated truly existent effects (studied with 80% power), and the other half investigated truly nonexistent effects, the resulting *p*-curve would be expected to have about four times as many *p* values below .01 as between .04 and .05 (also see Cumming, 2008; Hung, O'Neill, Bauer, & Kohne, 1997; Wallis, 1942).[3]

Some kinds of *p*-hacking, the selective reporting of *analyses* conducted on the same data set, are analogous to file-drawering, to selectively reporting *studies* (e.g., reporting results only for men or only for women). Thus, when a studied effect does not exist, these kinds of *p*-hacking are equally likely to result in low significant *p* values (e.g., .01s) and high significant *p* values (e.g., .04s). In contrast, other kinds of *p*-hacking are disproportionately more likely to result in high significant *p* values (e.g., .04s) than in low significant *p* values (e.g., .01s). Thus, *p*-hacking generally makes *p*-curves flatter (i.e., less right-skewed) and possibly left-skewed.[4]

When it comes to concluding that a literature lacks evidential value, *p*-curve analysis is conservative; it occasionally results in right-skewed *p*-curves even in the absence of an effect. As discussed in Simonsohn, Simmons, and Nelson (2015), this can occur if the findings are misreported or fraudulent, or if researchers choose the smallest possible *p* value from a large set of analyses (Ulrich & Miller, 2015). Simonsohn et al., (2015) recently revised the *p*-curve procedure to be more robust to these circumstances.

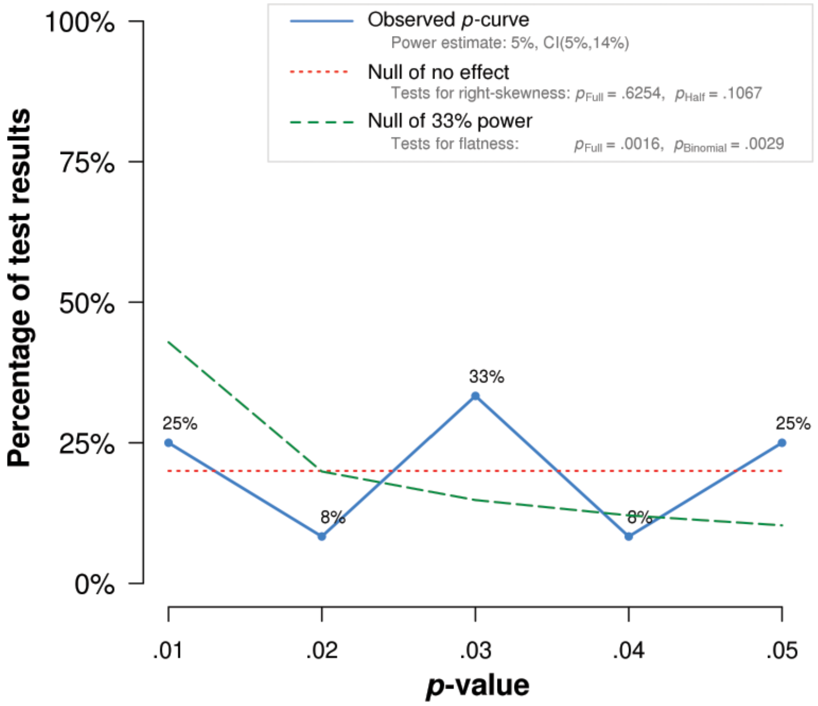
Inferences From Observed *P*-Curves

P-curve analysis involves two tests, one examining whether *p*-curve's shape is significantly right-skewed and one examining whether *p*-curve is significantly flat. The second test requires some explanation. In the same way that statistical inference cannot establish that two population means are exactly the same, one cannot establish that a distribution is exactly flat (i.e., that the "population" frequency of *p*s = .01 is exactly the same as the frequency of *p*s = .04). To circumvent this problem, one can rely on the fact that how right-skewed a *p*-curve is expected to be depends on the statistical power of the underlying studies. Studies with greater power yield steeper right-skewed *p*-curves (see Simonsohn et al., 2014a, 2014b). To test whether *p*-curve is flat, *p*-curve analysis tests whether *p*-curve is significantly less right-skewed than one would expect if the studies were so underpowered as to be able to detect a true effect only 33% of the time.[5] Thus, although one cannot establish whether *p*-curve is flat, one can establish whether it is significantly flatter than would be expected if the studies had 33% power.

This test provides protection against underpowered *p*-curves. When too few studies are used for *p*-curve analysis, the results will be inconclusive, neither significantly right-skewed nor significantly flat.

Results

Using the online *p*-curve app (http://www.p-curve.com), we analyzed the 33 studies that Carney et al. (2015) cited as evidence for the effectiveness of power posing (visit https://osf.io/ujpyn for our *p*-curve disclosure table and archived copy of R code used by the app). We had to exclude two studies because they investigated only feelings of power (the manipulation check) rather than downstream effects of the postural manipulations. We excluded two further studies because the critical test statistics were unreported. In addition, our *p*-curve analysis necessarily (and automatically) excluded seven *p* values because they were nonsignificant. Studies in which 2 × 2 reversing interactions were hypothesized require researchers using *p*-curve analysis to enter *p* values from each simple effect and thus to include two *p* values rather than one. For two studies in this sample, *p*-curve analysis automatically excluded one simple effect (because it was nonsignificant) but retained the other. Thus, we ultimately excluded 11 *p* values from 9 studies from the analysis, giving us a final sample size of 24 *p* values from 24 studies (33 – 9 = 24). The resulting *p*-curve is shown in Figure 1.



Note: The observed *p*-curve includes 24 statistically significant ($p < .05$) results, of which 10 are $p < .025$. There were 7 additional results entered but excluded from *p*-curve because they were $p > .05$.

Fig. 1. *P*-curve of the 33 studies cited by Carney, Cuddy, and Yap (2015) as evidence for the effects of power posing on downstream outcomes. The solid line shows the distribution of critical *p* values. It shows, for example, that 25% of the statistically significant *p* values were between .04 and .05, and that 33% were between .02 and .03. The key gives the 90% confidence interval (CI) for the estimate of the average power of the studies graphed. The dotted line shows the expected distribution of *p* values if there were truly no effect, and the dashed line shows the expected distribution of *p* values if the effect existed and the existing studies were powered at 33%. This figure was generated by *p*-curve app 4.05.

We first determined whether evidential value was present. As explained in Simonsohn et al. (2015), we conclude that a literature contains evidential value if either the half *p*-curve (which analyzes only critical *p* values below .025) is significantly right-skewed at the 5% level, or if both the half and full *p*-curve are significantly right skewed at the 10% level. Neither condition was met here (half: $z = -1.24$, $p = .11$; full: $z = 0.32$, $p = .63$).

We then compared the observed *p*-curve with what would be expected when studies have an average power of only 33%. One can conclude that there is an absence of evidential value if the full *p*-curve is significantly flatter than the 33%-power *p*-curve at $p < .05$.[6] This condition was met (full: $z = -2.95$, $p = .0016$), which allowed us to conclusively reject the null hypothesis that the sample of existing studies examines a detectable effect.

Finally, one can use *p*-curve analysis to estimate the average power of these studies. It is only 5%, which is the "power" we expect when the true effect size is zero and the significance threshold is .05 (since 5% of null effects will be significant at a threshold of .05). The 90% confidence interval around this estimate is narrow, excluding levels of average power greater than 14%. If the same studies were run again, it is unlikely that more than 14% of them would replicate, and our best guess is that 5% of them would be significant (in any direction).

Additional Analyses

Simonsohn et al. (2014a) provided detailed guidelines for selecting test results from studies. Because we followed those guidelines here, there was minimal ambiguity as to which test to select from each study. Moreover, we conducted a "robustness" *p*-curve that included 12 valid alternative *p*-value selections. The results from this analysis were very similar to the results reported in the previous section: The test for evidential value was nonsignificant (full: $p = .60$, half: $p = .53$), and the *p*-curve was significantly flatter than if the studies were powered at 33% on average (full: $p = .0031$); the estimate of average power was still 5%, with a 90% confidence interval excluding values greater than 17% (rather than 14%). Because Ranehill et al.'s replication obtained a significant effect of power posing on the manipulation check, self-reported power, we constructed a separate *p*-curve including only the seven manipulation-check results. The resulting *p*-curve was directionally right-skewed (full: $p = .051$, half: $p = .184$). Our *p*-curve disclosure table (http://osf.io/2fq9c) includes all *p*-value selections (and justifications), as well as everything the reader needs to easily evaluate and reproduce our analyses.

Power of *P*-Curve

The conclusion that this literature lacks evidential value cannot be explained (or explained *away*) by our *p*-curve analysis's lack of power. With 24 *p* values, our *p*-curve analysis has vastly more power than the underlying studies do. For example, if the 24 studies investigating expansive versus contractive posing had 33% power on average, then the resulting *p*-curve would have an 89% chance to detect evidential value. If the 24 studies had 50% power on average, then the resulting *p*-curve would have 99% power to detect evidential value. If 14 studies examined null effects, and 10 examined real effects, a *p*-curve based on all 24 would have more power than those 10 studies do on average (R code for these calibrations can be found at https://osf.io/sdgkq/). Moreover, Figure 2 shows that the results do not at all hinge on a few extreme observations.

Set of Studies

Like all statistical analyses, *p*-curve analyses provide information only about the sampled populations. The sample of studies we analyzed consists of what Carney et al. (2015) described as "all published tests (to our knowledge) of expansive (vs. contractive) posture on psychological outcomes" (p. 657). Thus, our conclusions apply only to all studies on the effects of expansive versus contractive posing that were known to Carney et al. in 2015. One reviewer criticized our focus on this set of studies, believing it to be arbitrary and subjective. Thus, it seems worthwhile to explain it.

Carney et al.'s (2015) response to the failed replication of their work was to say that 33 other studies provided evidence for their effects. Our goal in this commentary was to examine whether that sample of studies contains evidential value.

Given this objective, our set of studies was chosen *for* us, not *by* us. Moreover, given that this sample was not selected by Carney et al. (2015) for the purpose of conducting a *p*-curve analysis, it seems implausible that the selection of studies was guided, either implicitly or explicitly, by how large or small the critical *p* values were. Thus, this sample is both valid—it is by definition the population of interest to us—and unbiased—it was not selected by researchers interested in using *p*-curves to draw a particular conclusion. It is difficult to imagine a less arbitrary or subjective way to choose a sample of studies to analyze.[7]

Conclusion

Taken together, the results from Ranehill et al.'s (2015) replication and from our *p*-curve analysis suggest that the behavioral and physiological effects of expansive versus contractive postures ought to be treated as hypotheses currently lacking in empirical support. Although more highly powered future research may find replicable evidence for those benefits (or unexpected detriments), the existing evidence is too weak to justify a search for

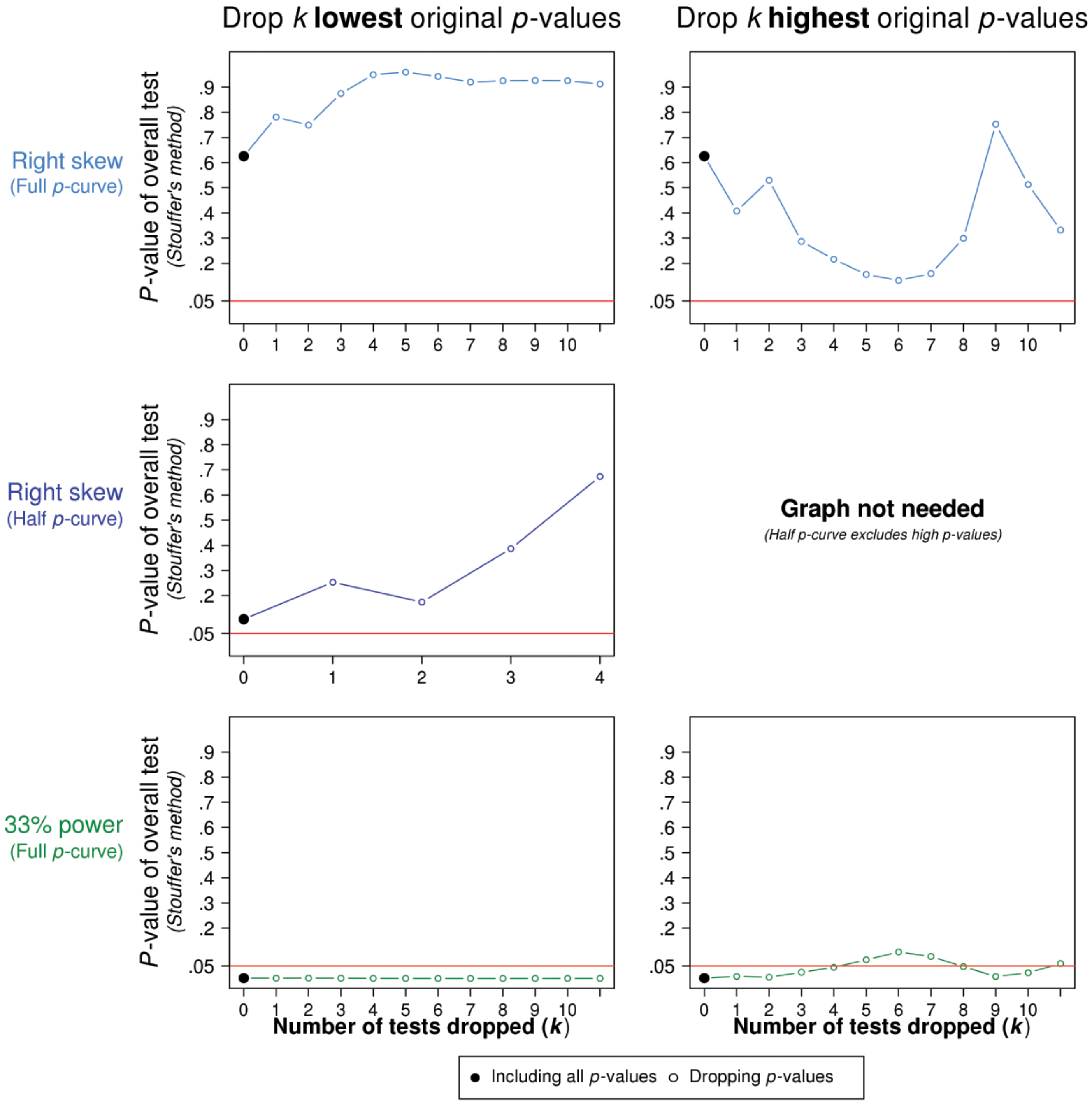


Fig. 2. Effect of dropping the lowest and highest *p* values on the significance of the full *p*-curve test for right-skewness (top row), the half *p*-curve test for right-skewness (middle row), and the test for flatness relative to 33% power (bottom row). Within each graph, the red horizontal line demarcates the significance threshold (*p* = .05), and the filled marker is the result reported in the text. This figure was generated by *p*-curve app 4.05.

moderators or to advocate for people to engage in power posing to better their lives.

Action Editor

D. Stephen Lindsay served as action editor for this article.

Author Contributions

J. P. Simmons and U. Simonsohn constructed the *p*-curve disclosure table, conducted the *p*-curve analysis, and wrote the manuscript. The author order is arbitrary.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Open Practices

All data have been made publicly available via the Open Science Framework and can be accessed at https://osf.io/ujpyn.

The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797616658563. This article has received the badge for Open Data. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

Notes

[1] In a more recent article (Cuddy, Wilmuth, Yap, & Carney, 2015), the original authors state that they consider self-reported feelings of power to be a manipulation check rather than an outcome, writing that "as a manipulation check, participants reported how dominant, in control, in charge, powerful, and like a leader they felt on a 5-point scale" (p. 1289). Moreover, the effects of postural manipulations on self-reported feelings of power are susceptible to demand effects. For example, if an experimenter asks participants to slouch for 2 min and then to rate how powerful they feel, participants may assume that the experimenter expects them to feel relatively powerless or may instead answer the question, "How powerful is the pose you just assumed?"

[2] Our own simulations show that, in general, PET-PEESE estimates are virtually nondiagnostic of true effect size.

[3] When studies have 80% power to detect an effect, about 72% of significant results are expected to have a $p < .01$ and only 4% to have a $p > .04$ (see Fig. 1 in Simonsohn et al., 2014a). Averaging each of these percentages with 20%, which is what is expected under the null hypothesis, one sees that 47% of significant *p* values would be expected to be below .01 and that 12% would be expected to be between .04 and .05.

[4] The effect of *p*-hacking on *p*-curve's shape hinges on whether the *p*-hacked analyses are correlated with each other. When the analyses are *un*correlated with each other, then *p*-hacking will do the same thing to *p*-curve as file-drawering does (i.e., it will make *p*-curve flat under the null hypothesis). When the analyses are correlated with each other, then *p*-hacking is more likely to result in significant *p* values that are closer to .05 than to .01 (Simonsohn et al., 2014a, 2014b). See Supplement 3, "Modeling *p*-hacking," in Simonsohn et al. (2014a) for a formal analysis of this distinction.

[5] Like all cutoffs, the 33%-power cutoff is necessarily arbitrary. Simonsohn et al. (2014a) chose it because it is a very low level of power, as a study with 33% power would be twice as likely to fail as to succeed. Because cutoffs are arbitrary, they should be used as reference points rather than as meaningful categorical divides. In the case of *p*-curve analysis, the more strongly one rejects the null hypothesis that the study has 33% power, the more inconsistent the evidence is with the existence of the hypothesized effect.

[6] This test is also significant if both the binomial and the full *p*-curve are flatter at $p < .10$.

[7] The reviewer identified seven additional studies that Carney et al. (2015) did not include in their review. The editor suggested we update our analysis by including the studies that were published since Carney et al.'s review. Only two of the seven studies mentioned by the reviewer potentially met this criterion, and neither one of them actually investigated the effects of expansive versus contractive postures. Leitan, Williams, and Murray (2015) manipulated whether people tilted their heads up and down (a manipulation Carney et al. explicitly chose to exclude; see the note in their Table 1). Michalak, Rohde, and Troje (2015) manipulated whether people walked on a treadmill in a happy versus depressed pattern. This is a good opportunity to emphasize a critical point about the use of *p*-curves: The rule guiding the selection of studies must be set in advance and be disclosed to protect against the cherry-picking of studies. The reviewer not only suggested studies that do not belong in the analysis, but also did not disclose an a priori study-selection rule.

References

Blodget, H. (2013, May). This simple 'power pose' can change your life and career. *Business Insider*. Retrieved from http://web.archive.org/web/20151220171940/http://www.businessinsider.com/power-pose-2013-5

Carney, D. R., Cuddy, A. J. C., & Yap, A. J. (2010). Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychological Science*, *21*, 1363–1368.

Carney, D. R., Cuddy, A. J. C., & Yap, A. J. (2015). Review and summary of research on the embodied effects of expansive (vs. contractive) nonverbal displays. *Psychological Science*, *26*, 657–663.

Cuddy, A. J. C. (Producer). (2012). Amy Cuddy: Your body language shapes who you are. Retrieved from http://web.archive.org/web/20160204004017/https://www.ted.com/talks/amy_cuddy_your_body_language_shapes_who_you_are?language=en#

Cuddy, A. J. C., Wilmuth, C. A., Yap, A. J., & Carney, D. R. (2015). Preparatory power posing affects nonverbal presence and job interview performance. *Journal of Applied Psychology*, *100*, 1286–1295.

Cumming, G. (2008). Replication and *p* intervals: *p* values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, *3*, 286–300.

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*, 455–463.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *The British Medical Journal*, *315*, 629–634.

Francis, G. (2014). The frequency of excess success for articles in *Psychological Science*. *Psychonomic Bulletin & Review*, *21*, 1180–1187.

Gervais, W. (2015). *Putting PET-PEESE to the test* [Web log post]. Retrieved from http://web.archive.org/web/20160120140336/http://willgervais.com/blog/2015/6/25/putting-pet-peese-to-the-test-1

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*, 1–20.

Hung, H. M. J., O'Neill, R. T., Bauer, P., & Kohne, K. (1997). The behavior of the *p*-value when the alternative hypothesis is true. *Biometrics*, *53*, 11–22.

Ioannidis, J. P. A. (2011). Excess significance bias in the literature on brain volume abnormalities. *Archives of General Psychiatry*, *68*, 773–780.

Ioannidis, J. P. A., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials*, *4*, 245–253.

Leitan, N. D., Williams, B., & Murray, G. (2015). Look up for healing: Embodiment of the *heal* concept in looking upward. *PLoS ONE*, *10*(7), Article e0132427. doi:10.1371/journal.pone.0132427

Michalak, J., Rohde, K., & Troje, N. F. (2015). How we walk affects what we remember: Gait modifications through biofeedback change negative affective memory bias. *Journal of Behavior Therapy and Experimental Psychiatry*, *46*, 121–125.

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

Ranehill, E., Dreber, A., Johannesson, M., Leiberg, S., Sul, S., & Weber, R. A. (2015). Assessing the robustness of power posing: No effect on hormones and risk tolerance in a large sample of men and women. *Psychological Science*, *26*, 653–656.

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

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. West Sussex, England: John Wiley & Sons.

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, *17*, 551–566.

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.

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Better *p*-curves: Making *p*-curve more robust to errors, fraud, and ambitious *p*-hacking, a reply to Ulrich and Miller (2015). *Journal of Experimental Psychology: General*, *144*, 1146–1152.

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, *5*, 60–78.

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.

Ulrich, R., & Miller, J. (2015). *p*-hacking by post hoc selection with multiple opportunities: Detectability by skewness test? Comment on Simonsohn, Nelson, and Simmons (2014). *Journal of Experimental Psychology: General*, *144*, 1137–1145.

Wallis, W. A. (1942). Compounding probabilities from independent significance tests. *Econometrica*, *10*, 229–248.