

## Defending the Decimals: Why Foolishly False Precision Might Strengthen Social Science

Jeremy Freese


Northwestern University

**Abstract:** Social scientists often report regression coefficients using more significant figures than are meaningful given measurement precision and sample size. Common sense says we should not do this. Yet, as normative practice, eliminating these extra digits introduces a more serious scientific problem when accompanied by other ascendant reporting practices intended to reduce social science's long-standing emphasis on null hypothesis significance testing. Coefficient  $p$ -values can no longer be recovered to the degree of precision that  $p$ -values have been abundantly demonstrated to influence actual research practice. Developing methods for detecting and addressing systematically exaggerated effect sizes across collections of studies cannot be done effectively if  $p$ -values are hidden. Regarding what is preferable for scientific literature versus an individual study, the costs of false precision are therefore innocuous compared to alternatives that either encourage the continuation of practices known to exaggerate causal effects or thwart assessment of how much such exaggeration occurs.

**Keywords:** statistical reporting; publication bias; estimate precision; confidence intervals;  $p$ -values;  $p$ -hacking

**Editor(s):** Jesper Sørensen, Delia Baldassarri; **Received:** June 13, 2014; **Accepted:** July 3, 2014; **Published:** December 8, 2014

**Citation:** Freese, Jeremy. 2014. "Defending the Decimals: Why Foolishly False Precision Might Strengthen Social Science." *Sociological Science* 1: 532-541. DOI: 10.15195/v1.a29

**Copyright:** © 2014 Freese. This open-access article has been published and distributed under a Creative Commons Attribution License, which allows unrestricted use, distribution and reproduction, in any form, as long as the original author and source have been credited. 

SOCIAL science methodologists strongly advocate for thoughtful reporting practices that emphasize the substantive meaning of results. Increasingly despised are tables with vast seas of estimates that readers are left on their own to navigate, like ancient mariners, by way of the table's stars. Part of the movement for more mindful presentation is questioning the seemingly comedic precision with which social science results are sometimes reported.

As an example, *Sociological Science* presently offers the following as a guideline:

Authors are discouraged from implying too much certainty in their estimated results by offering coefficients that extend to three and four decimal places. Two or fewer decimal places are typically sufficient.

Recommending two digits has a strong methodological pedigree (e.g., Ehrenberg 1977; Wainer 1997). Given the typical limits of social science precision, it might also seem simply obvious good sense. If faulted for anything, perhaps the issue seems trivial, even persnickety.

Instead, I will here make an argument that is triply contrarian about how social science regression results ought to be reported. First, when considered in conjunction with other reporting practices, the number of significant figures used in reporting estimates is more important than one might think for the cumulative projects of social science. Second, to this end, two significant digits are typically not enough, and, if reporting only two digits were to become the norm, the resulting literature would have demonstrably weaker properties as science. Third, reporting additional digits might actually be most important

in those cases where that digit is most obviously substantively meaningless.

As I will explain, the culprit of all this mischief is social science's frenemy relationship with  $p$ -values. We attend to  $p$ -values too much but also want to discourage attending to  $p$ -values too much. This leads to a combination of ascendant reporting practices that individually make good sense but together may have negative consequences. Until this is put in order, false precision in reporting regression coefficients does social science more prospective good than harm.

## The Benefit of Brevity

Every experienced quantitative social scientist has seen drafts or manuscripts in which coefficients are reported with wild precision for small samples with coarse measures (e.g.,  $\hat{\beta} = 0.67423216$ ,  $N = 95$ ). Presumably the author has simply copied output from a statistical package. Of course, the problem is that so many digits might make the results appear to a naive reader to be more precise and scientific than they really are, when in fact nearly all the digits are literally meaningless because the variables involved were not measured to a corresponding level of precision. Even if variables are perfectly measured, it is likewise incoherent to pretend accurate estimation of population parameters that exceeds the order of magnitude of the sample size, such as providing a point estimate that 51.641 percent of the population is female because 472 out of 914 people in a sample are.

Coefficients with seven or more decimal places rarely make it into social science journals that anyone reads, but coefficients with three and four routinely do. Saying that two decimal places are typically sufficient correctly suggests that the accuracy of social science research rarely justifies more, or that, even when justified, these extra digits are typically inconsequential for any meaningful substantive interpretation (Ehrenberg 1977).<sup>1</sup> If one were to take a hard-nosed look at

<sup>1</sup>The key issue here regarding precision is significant figures, not decimal places as such (e.g. a coefficient of 0.5864 dollars has precisely the same amount of false precision if rescaled to 58.64 cents). For purposes of style guidelines, tables routinely report all estimates to the same number of decimal places, and inspection of a sample of sociology journal articles indicates the modal coefficient is

many social sciences literatures on these grounds, one might even question how often a second significant figure is defensible.

So it is fully understandable why providing more digits than accuracy warrants might feel more like science theater than actual science. Less obvious is the actual harm in it. For those scientists who have the appropriate statistical training, it is unclear why seeing reported coefficients of 0.5864 instead of 0.58 would provoke any illusions about the precision of social science. For those without this training, the difference between a regression coefficient of 0.58 and 0.5864 is smaller than the standard error in any case in which the coefficient and standard error are reported to the same number of decimal places, and typically vastly so. Consequently, for the uninitiated, *whatever misconception about precision is involved in providing point estimates and standard errors to more significant figures than the data warrant, it is less than the misconception fostered by providing point estimates in the first place.*

## The Cost of Concision

The problem with reporting coefficients to one or two significant figures is easily stated; the work is in explaining why the problem is worrisome and why countenancing false precision might be preferable to more direct solutions. If we only report coefficients and standard errors to one or two significant figures, then we cannot recover the  $p$ -values of coefficients with any precision. For example, Haskins (2014:151) uses three decimal places for reporting results, and the first result in her first table of coefficients has a value of 0.143 with a standard error of 0.070. We can calculate from this that the two-sided  $p$ -value is in the interval [0.039, 0.043] ( $2.021 < z < 2.065$ ). Had she only reported two decimals,  $b = 0.14$ ,  $se = 0.07$ , we would know only that the two-sided  $p$ -value was in the interval [0.026, 0.071] ( $1.800 < z < 2.231$ ), which is 10 times wider.<sup>2</sup>

$0.1 < \hat{\beta} < 1$ , so, at least in sociology, "significant figures" and "decimal places" are modally the same. Notably,  $\hat{\beta} < 0.1$  appears more common in practice than  $\hat{\beta} > 1$ , meaning that the number of significant figures is more often *less* than the number of decimal places than the other way around.

<sup>2</sup>Haskins uses a one-sided test, so the presence of an asterisk does not mean that we could have inferred from

One can easily find examples of key results reported to only one significant figure. Barnes and Jacobs (2013) estimate a gene environment interaction on violent criminal behavior, reporting a coefficient of 0.0002 and a standard error of 0.0001 (with an asterisk indicating  $p_{\text{one-sided}} < 0.05$ ). Here the two-sided  $p$ -value could be anywhere in (0.0000006, 0.1).

Granted, the inability to recover  $p$ -values might seem a good thing. Indeed, what creates the issue is the rise of reporting standard errors and confidence intervals instead of explicit reporting of  $z$ -statistics and precise  $p$ -values. The problems of null hypothesis significance testing (NHST) have been long and widely voiced (e.g., Cohen 1994; McCloskey and Ziliak 1996; Gill 1999; Kline 2004; Carver 1978). Commendably, social science methodologists have been counseling for decades not to focus too much on  $p$ -values and, in the absence of mass conversion to Bayesian statistics, to attend to substantive significance at least as strongly as statistical significance.

Regardless, however, it is unmistakable that social scientists still do in practice attend to statistical significance. Of 10 *Sociological Science* articles presenting regression results that I examined, all but one used starred coefficients to indicate statistically significant results, and this rate is fully consistent with how regression analyses are reported in other major journals. Using what are called *caliper tests*, more comprehensive analyses of  $p$ -values in sociology have shown a striking difference in the percentages of  $z$ -scores that imply  $p$ -values slightly below 0.05 compared to those slightly above 0.05 (Gerber et al. 2008; Auspurg and Hinz 2011). Such findings are not at all unique to sociology (e.g., Gerber and Malhotra [2008b] for political science; Masicampo and Lalande [2012] for psychology), and they provide definitive indication of a specifically  $p$ -oriented publication bias in the scientific literature of many fields (Dwan et al. 2008).

There is strong reason to suppose that  $p$ -value-oriented publication bias results from a mutually reinforcing set of sources that creates a difficult social dilemma (Auspurg and Hinz 2011). Some editors are more likely to decline publishing null results (or even sending them out for review). Yet, even if an editor had no such bias, the review-

two decimal places that the two-sided  $p$ -value was less than 0.05.

ers advising them are less likely to recommend publication of papers with nonsignificant findings (Emerson et al. 2010). Even if both the editors and reviewers at a particular outlet did not have this bias, all evidence indicates that authors are much less likely to pursue writing up and seeking to publish nonsignificant results, presumably because they anticipate low returns for doing so (Franco et al. 2014). And even if authors were confident that a particular outlet would be open to their substantively positive but statistically nonsignificant results, NHST is so widespread and ingrained that an author can reasonably expect statistical significance to influence the reception and citations an article receives, providing incentives to align findings with standards of statistical significance that go beyond simply being able to publish a paper.

Evidence indicates that preemptive author actions are more important for publication bias than actions of editors and reviewers (Franco et al. 2014). Authors have considerable agency over the  $p$ -values that appear in their oeuvre of submitted manuscripts. First, researchers may pursue research hypotheses to the point of being able to assess their prospects for producing statistically significant results in their favor or not, and only go to the trouble of writing up those that do. We might call this *study selection*, one manifestation of which is the “file drawer problem” of completed but unpublished studies (Rosenthal 1979). Second, researchers may use the discretion available when studying a particular hypothesis to either consciously or unconsciously make analytic decisions that lead to statistically significant results. Simonsohn et al. (forthcoming) call this *p-hacking*.<sup>3</sup>

The cumulative effect of these practices is chronic overestimation of effect sizes, and this has led to broad proposals that nearly all novel positive findings involve exaggerated effect sizes and most may simply be false (Ioannidis 2005, 2008). In sum, then, in sociology and elsewhere there is plain indication that researchers are attending to the nether decimals of their statistical output in terms of  $p$ -values. Despite this, sociol-

<sup>3</sup>As a somewhat different matter, researchers exploring data may happen upon unhypothesized results that are statistically significant, and then may write up these results as an a priori test of the hypothesis. This has been dubbed *HARKing* by Kerr (1998), for hypothesizing after results are known.

ogy papers often do not report  $p$ -values beyond indicating whether it is above or below some conventional threshold(s). In this case, when coefficients and standard errors are reported to only one or two significant figures, the  $p$ -value cannot be recovered with any precision.

We might particularly worry about overstated effect sizes and  $p$ -oriented publication biases for research based on smaller samples. Smaller samples require larger effect sizes to obtain statistical significance, and effect sizes may be more easily inflated by  $p$ -hacking practices that overfit the data. For this reason, being able to recover  $p$  with precision is likely most important for literatures replete with smaller samples. Accordingly, then, if  $p$ -values are only indirectly reported through starred coefficients and standard errors, then providing sufficient significant figures of these quantities to recover  $p$ -values is most important in smaller samples. This is so even though the smaller samples imply that the extra digits reported in these coefficients and standard errors are substantively meaningless.

In sum, it is clear that  $p$ -values influence researcher practice and that the cumulative consequence of this influence are literatures that, on average, exaggerate the effect sizes they purport to estimate. What may be less clear from the foregoing is what knowing the  $p$ -values of coefficients with precision can contribute to evaluating or addressing this problem. I explain why next.

## Patterns in $p$

What makes systematic exaggeration of effect sizes so pernicious is that it can be difficult or impossible to detect in individual studies. Every practice within a given study may be completely defensible, and yet the cumulative properties of a set of such studies provide unmistakable indication of exaggerated effect sizes. For example, funnel plots may be used to show that published effect sizes on a topic are routinely larger on average with small samples than large samples, which would not happen in a set of studies that were all providing unbiased estimates of the same parameter (e.g., Duval and Tweedie 2000).

The ability to do better meta-analysis and diagnostics like funnel plots serves itself as an argument for reporting coefficients with greater

precision than may seem warranted by only considering each individual study in isolation. That said, meta-analysis is much rarer in fields like sociology than in other fields, like epidemiology or psychology, because of the far slower accumulation of a body of studies on a given topic that might all be treated as pursuing the same parameter (see Branigan et al. [2013] as one exception). However, what is quickly becoming better appreciated is that  $p$ -values of a set of studies are informative in ways the effect sizes themselves are not. The preceding caliper examples provide one compelling example, but it is also becoming clearer that looking simply for discontinuities around conventional thresholds is likely only the beginning.

Although various methods for drawing inferences based on accumulated  $p$ -values are under development (Ioannidis and Trikalinos 2007; Gadbury and Allison 2012), particularly accessible and compelling to sociologists may be the  $p$ -curve analysis being pioneered by Simonsohn et al. (forthcoming). The logic of  $p$ -curve analysis can be readily understood by imagining a literature in which researchers were doggedly and accurately estimating parameters that were actually 0, but the literature comprised only positive and significant findings because, otherwise, results went unpublished. Because the distribution of  $p$  is uniform for a set of studies in which  $\beta = 0$  (regardless of sample size), the distribution of  $p$ -values in this scenario should be roughly equal for any equal-sized intervals between 0 and 0.05. That is, the number of published results in which  $p$  is in the interval (0.01,0.03) should be as frequent as the number in the interval (0.03,0.05). Conversely, if  $p$ -hacking and related practices are involved, we might expect investigators to be more likely to stop making arbitrary decisions that moved past a conventional threshold rather than to keep trying to reduce  $p$ -values beyond that threshold. In such cases, studies with  $p$ -values of (0.03,0.05) would be more common than  $p$ -values of (0.01, 0.03). Faith that literatures are not simply the result of  $p$ -hacking and fortuitous noise may be found when the  $p$ -values are skewed the other direction, with  $p = (0.01, 0.03)$  being more common than  $p = (0.03, 0.05)$ . Importantly, an appropriately right-skewed distribution of  $p$ -values can still be observed if a literature systematically exaggerates coefficients owing to bad study designs,

but it does suggest less epistemic pollution from motivated practices to nudge  $p$  under a threshold.

Simonsohn et al. (forthcoming) present the  $p$ -curve for a set of experiments that employ a particular technique that they hypothesized to be a red flag for  $p$ -hacking and compare this to the curve for a control set of experiments from the same journal that did not contain any red flags. They found 15 of 20 statistically significant ( $p < 0.05$ ) key results in the suspected group reported  $p > 0.03$ , compared to only 4 of 22 in their control group. In a different example, Nelson et al. (2014) compare  $p$ -curves from many replications of contrasting experiments that have been used to conclude that, though under some conditions, having more choices is positive (which is intuitive), under other conditions having more choices is negative (which is counterintuitive and has been the subject of much popular attention [e.g., Schwartz 2004]). Their analyses show clear evidence of an evidentiary  $p$ -curve in the experiments finding statistically significant results in the direction that more choice is good, while finding no such pattern among the statistically significant results in the direction that more choice is bad.

A very rough demonstration of the potential values of looking at the distribution of  $p$  in assessing the sociological literature can be found by combining data from the histograms of  $z$ -scores presented in previously mentioned studies that used caliper tests (Gerber and Malhotra 2008; Auspurg and Hinz 2011) and converting these to  $p$ -values.<sup>4</sup> The demonstration is rough because the histograms prohibit recovering  $p$  to a desirably precise interval although it is still more precise than what would be available if two decimal places was the norm by which sociological results were reported (and no  $z$ -scores or  $p$ -values were directly reported). If, for simplicity, we assume  $z$ -scores within a given bin are uniformly distributed, we can present the distribution of

<sup>4</sup>The aforementioned introduction to  $p$ -curves also makes clear one problem with the caliper test, which is that, in the absence of reporting biases, the numbers of studies above and below a caliper expressed in  $p$  are only expected to be equal if all tested hypotheses are false, whereas the far more likely case in practice is that at least some hypotheses are true. The assumptions about true positives involved in the NHST for a caliper test involving  $z$  are even less clear.

$p$ -values for 0.01 intervals from 0.01 to 0.11.<sup>5</sup> We can see a monotonic decreasing curve, which is consistent with the idea that many hypothesis tests in sociology do indeed estimate parameters that are not zero, although again we cannot say how much of this is due to the hopeful scenario of many hypotheses being true versus the gloomier one of many studies being compromised by design artifacts (Figure 1).

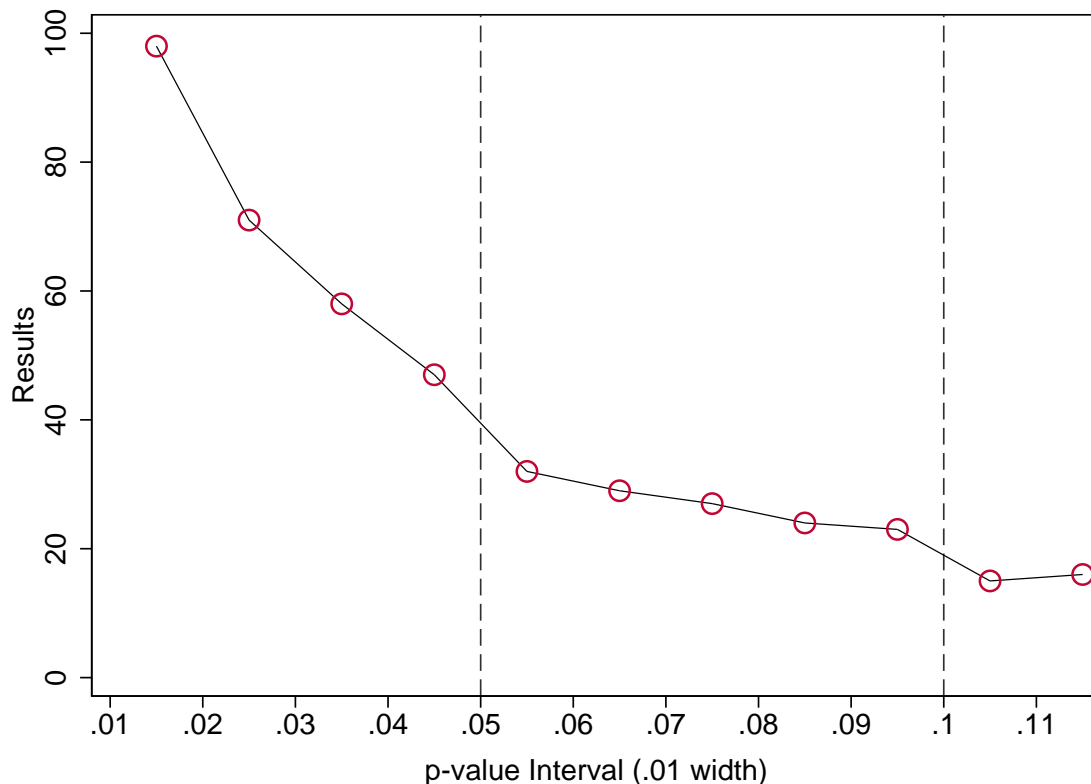
We can also see discontinuities in the shape of the curve at the  $p < 0.05$  and  $p < 0.10$  thresholds that are indicative of some analytic or publication decisions being influenced by  $p$  thresholds. In truth, the cumulative result here is not as bad as one might expect from the very clear evidence of orientation toward  $p$ -thresholds depicted in the caliper test by Gerber and Malhotra (2008). A large part of this is due to their caliper tests separating one-tailed and two-tailed tests, which was valuable in their study because it demonstrated clearly that one-tailed tests are often not invoked for any philosophical reason but because of their strategic consequences for  $p$ . However, combining results permits the intriguing conjecture that one-tailed significance testing might be a primary apparatus by which  $0.05 < p_{\text{two-tailed}} < 0.1$  results of tested hypotheses make it into the sociological literature. In other words, researchers might  $p$ -hack less because they can simply switch to one-tailed tests for marginal results. Although strategic invocation of one-tailed tests might seem bad practice from the standpoint of an individual study, if we think instead about the probative value of a literature, a system in which one-tailed tests were openly used to obtain publishable results is preferable to one that provides incentives for stealthy  $p$ -hacking that exaggerates effect sizes to obtain a result that passes a two-tailed significance test.<sup>6</sup>

Although this example is admittedly coarse and non comparative, one can imagine substan-

<sup>5</sup>As would be hoped for a literature that is not simply pursuing noise, the interval (0.00, 0.01) contains too many values to make the difference among the rest visible in a plot.

<sup>6</sup>This is not to deny concerns about the use of one-tailed tests. The obvious one is if they are used to overstate how confident we really are that an estimated coefficient reflects a true parameter in the same direction. Also, the caliper tests referenced and the  $p$ -curve shown here indicate the strong possibility that some  $p$ -hacking is done to get hypothesized results below the  $p_{\text{two-tailed}} < 0.1$  threshold.





**Figure 1:** Estimated  $p$ -values of study results reported in two studies of publication bias in sociology.

tively intriguing comparisons: hypotheses tested by main effects versus interaction effects; hypotheses tested using surveys versus experiments; articles published by one author versus multiple authors (see Auspurg and Hinz 2011); results for key coefficients as presented in bivariate results versus multivariate results. The pace of recent development of methods of probing  $p$ -values gives hope of the possibility of considerable further elaboration of techniques. But *what can be assessed in the future will be conditioned by the information that is preserved now*. If social scientists fail to present results in ways compatible with cumulative assessment methods, those methods will be unavailable to assess and probe their sciences.<sup>7</sup>

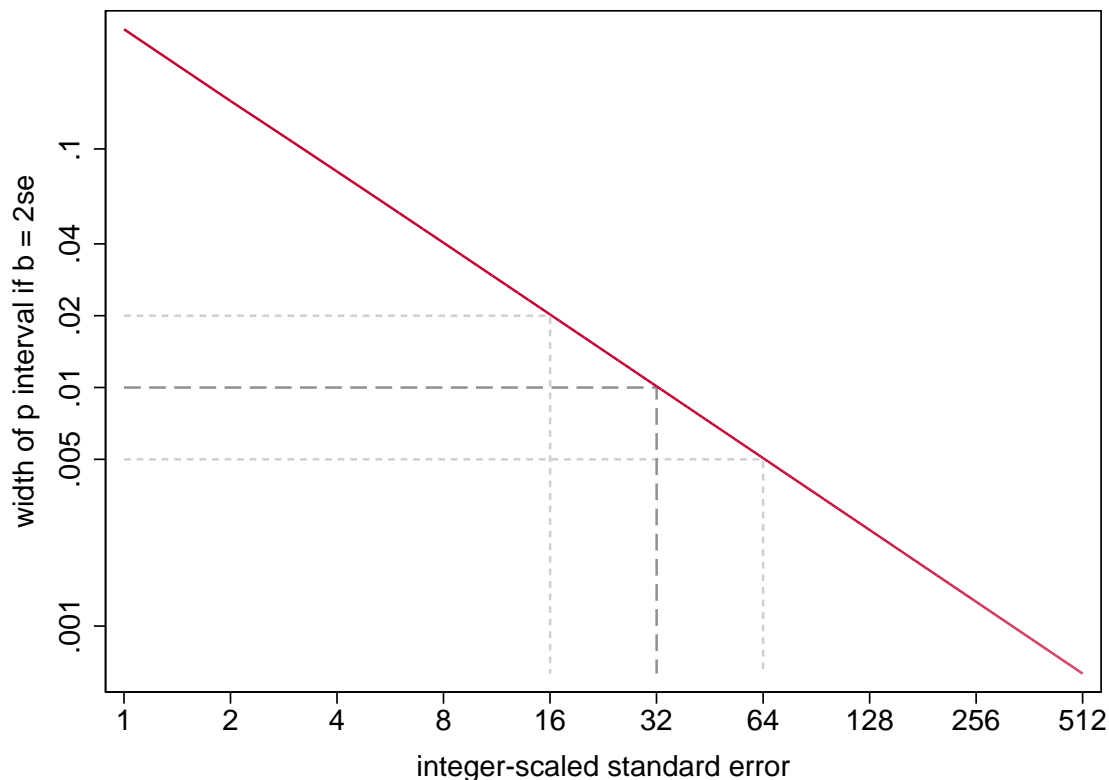
<sup>7</sup>As a further example, *Sociological Science* has launched with guidelines and protocols that provide a much clearer nudge away from mindless fealty to NHST than other generalist sociological journals. This inspires a hopeful hypothesis that the literature of *Sociological Science* may be less influenced by  $p$ -mischief than elsewhere

## Identifying the Ideal

If being able to recover a  $p$ -value from a coefficient and standard error is deemed important, what does it mean for how coefficients and standard errors should be reported? The answer is straightforward: if precise  $p$ -values are otherwise not reported, *coefficients should be reported to three significant figures*.

To show why, note first the importance of 0.05 to conventional significance testing, such that the prototypic case is that in which the coefficient is twice the standard error. For a given standard error,  $p$  can be recovered with greater precision as the coefficient increases, so any guideline that

and so that published effect sizes in this journal may be less exaggerated by publication bias. If true, this would be a wonderful demonstration that *Sociological Science* is improving sociological science. But of course, this hypothesis will only ever be potentially testable if reported results permit recovery of  $p$ -values.



**Figure 2:** The integer-scaled standard error and the width of the interval of possible  $p$ -values.

affords a particular level of precision when  $p$  is approximately 0.05 only improves as  $p$  decreases.

The key quantity for the precision with which we may recover  $p$  in this case is not the significant figures *per se*, but the *integer-scaled standard error*. This is just the standard error rescaled to whatever yields precision in whole units (e.g., if  $se_{\text{reported}} = 0.044$ ,  $se_{\text{integer}} = 44$ ). When  $p$ -values are based on  $z$ -scores and  $\beta = 2se_{\text{integer}}$ , then the width of the interval ( $p_{\text{min}}, p_{\text{max}}$ ) has a simple log-log relationship with  $se_{\text{integer}}$ , which is shown in Figure 2.

If we consider being able to recover the  $p$ -value within a 0.01-wide interval to approximate the minimum precision that might still prove useful, then the integer-scaled standard error needs to be at least 32. For significant results, this corresponds to an integer-scaled coefficient of at least 64. Given that many articles have multiple coefficients and standard errors as key results, a minimum of 32 for the standard error would seem

to imply in practice some integer-scaled standard errors of at least 50 and hence some integer-scaled coefficients of at least 100 if  $p$  is about 0.05. An integer-scaled coefficient of at least 100 is exactly what reporting a coefficient to three significant figures means. Also, happily, three significant figures are just enough to allow recovery of  $p$ -value within .01 when  $p \approx 0.10$ , so this guideline remains appropriate even with many results reported as significant using a one-tailed  $p < 0.05$  test.<sup>8</sup>

## Conclusion

I have argued that social scientists should report regression coefficients to at least three significant figures because otherwise one cannot use these to reconstruct the  $p$ -value with any precision, and

<sup>8</sup>Since  $|z| = 1.64$  when  $p = 0.1$ , the minimal case is when  $\beta = 100$  and  $se_{\text{integer}} = 100/1.64 = 61$ , for which the interval of  $|z|$  is (1.61, 1.66) and  $p$  is (0.097, 0.106).

that reconstructing the  $p$ -values of a set of studies has evidentiary significance for the cumulative evaluation of scientific literatures. Of course, a far more direct approach would be to just present  $p$ -values along with coefficients in tables. The downside to doing so is that it might be understood as focusing too much attention on  $p$ -values and encouraging continued naïvete in the face of the many accumulated critiques of NHST.

Reporting the test statistic to two decimal places instead of the standard error might be a useful alternative, as with a  $z$ -score this allows recovery of the  $p$ -value within at least an interval of 0.0005 for a  $p < 0.05$  result. Given that two decimal places for a  $z$ -score implies at least three significant figures so long as  $p < 0.32$  ( $|z| > 1$ ), the consequences for recovering standard errors and constructing confidence intervals are small, especially given that many standard errors in social science are now reported to only one or two significant figures. The normative and aesthetic question is whether  $z$ -scores still connote too much credulity toward statistical versus substantive significance. When reported with enough significant figures, coefficients and standard errors allow  $p$ -values to be recovered with good precision without any explicit practice that might be interpreted as endorsing close attention to  $p$ -values.

Overall, the foregoing issues in the reporting of results imply a trade off among different ways that a literature might be said to be naive. Overreliance on  $p$ -values leads to distorted research practices that are increasingly recognized as exaggerating effect sizes within a given literature, often dramatically. Deemphasizing  $p$ -values in publications, so long they are still influencing research practices, might *eventually* have a salutary effect of reducing their influence on practice, but at the expense of *definitely* making it harder to detect and demonstrate any distortions that  $p$ -oriented publication practices induce. In contrast to consequences of either promoting exaggerated effect sizes or keeping these exaggerations from being detectable, the actual scientific downsides of false precision in presenting coefficients are downright minor. Researchers might look a little silly—like they don't know any better—and coefficients may be imbued with unwarranted certainty by readers who have no understanding of standard errors. If we consider social science

to be a collective and aspirationally cumulative project, then scientific standards for presenting results should not be based on what is most logical or optimal for any individual study taken on its own; rather, we should focus on what is optimal for the cumulative accuracy of a field and its availability for cumulative assessment.

## References

- Auspurg, Katrin and Thomas Hinz. 2011. "What Fuels Publication Bias?" *Jahrbücher für Nationalökonomie & Statistik* 231:630–660.
- Barnes, J. C. and Bruce A. Jacobs. 2013. "Genetic Risk for Violent Behavior and Environmental Exposure to Disadvantage and Violent Crime: The Case for Gene Environment Interaction." *Journal of Interpersonal Violence* 18:92–120. <http://dx.doi.org/10.1177/0886260512448847>.
- Branigan, Amelia R., Kenneth J. McCallum, and Jeremy Freese. 2013. "Variation in the Heritability of Educational Attainment: An International Meta-analysis." *Social Forces* 92:109–40. <http://dx.doi.org/10.1093/sf/sot076>.
- Carver, Ronald P. 1978. "The Case against Statistical Significance Testing." *Harvard Educational Review* 48:378–99.
- Cohen, Jacob. 1994. "The Earth is Round ( $p < .05$ )." *American Psychologist* 49:997–1003. <http://dx.doi.org/10.1037/0003-066X.49.12.997>.
- Duval, Sue and Richard Tweedie. 2000. "Trim and Fill: A Simple Funnel-plot-Based Method of Testing and Adjusting for Publication Bias in Meta-analysis." *Biometrics* 56:455–63. <http://dx.doi.org/10.1111/j.0006-341X.2000.00455.x/>
- Dwan, Kerry, Douglas G. Altman, Juan A. Arnaiz, Jill Bloom, An-Wen Chan, Eugenia Cronin, Evelyne Decullier, Philippa J. Eastbrook, Erik Von Elm, Carrol Gamble, et al. 2008. "Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias." *PLoS one* 3:e3081. <http://dx.doi.org/10.1371/journal.pone.0003081>.



- Ehrenberg, A. S. C. 1977. "Rudiments of Numeracy." *Journal of the Royal Statistical Society, Series A* 140:277–97. <http://dx.doi.org/10.2307/2344922>.
- Emerson, Gwendolyn B., Winston J. Warme, Fredric M. Wolf, James D. Heckman, Richard A. Brand, and Seth S. Leopold. 2010. "Testing for the Presence of Positive-Outcome Bias in Peer Review: a Randomized Controlled Trial." *Archives of Internal Medicine* 170:1934–39. <http://dx.doi.org/10.1001/archinternmed.2010.406>.
- Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. "Publication Bias in the Social Sciences: Unlocking the File Drawer." In *Paper presented at the Annual meeting of the Midwest Political Science Association*, Chicago, IL.
- Gadbury, Gary L. and David B. Allison. 2012. "Inappropriate Fiddling with Statistical Analyses to Obtain a Desirable *p*-value: Tests to Detect its Presence in Published Literature." *PloS one* 7:e46363. <http://dx.doi.org/10.1371/journal.pone.0046363>.
- Gerber, Alan, Neil Malhotra, et al. 2008. "Do Statistical Reporting Standards Affect What Is Published? Publication Bias in Two Leading Political Science Journals." *Quarterly Journal of Political Science* 3:313–26. <http://dx.doi.org/10.1561/100.00008024>.
- Gerber, Alan S and Neil Malhotra. 2008. "Publication Bias in Empirical Sociological Research: Do Arbitrary Significance Levels Distort Published Results?" *Sociological Methods & Research* 37:3–30.
- Gill, Jeff. 1999. "The Insignificance of Null Hypothesis Significance Testing." *Political Research Quarterly* 52:647–74. <http://dx.doi.org/10.1177/106591299905200309>.
- Haskins, Anna R. 2014. "Unintended consequences: Effects of Paternal Incarceration on Child School Readiness and Later Special Education Placement." *Sociological Science* 1:141–58. <http://dx.doi.org/10.15195/v1.a11>.
- Ioannidis, John P. A. 2005. "Why Most Published Research Findings are False." *PLoS medicine* 2:e124. <http://dx.doi.org/10.1371/journal.pmed.0020124>.
- Ioannidis, John P. A. and Thomas A. Trikalinos. 2007. "An Exploratory Test for an Excess of Significant Findings." *Clinical Trials* 4: 245-253. <http://dx.doi.org/10.1177/1740774507079441>.
- Ioannidis, John P. A. 2008. "Why Most Discovered True Associations are Inflated." *Epidemiology* 19:640–48. <http://dx.doi.org/10.1097/EDE.0b013e31818131e7>.
- Kerr, Norbert L. 1998. "HARKing: Hypothesizing after the Results Are Known." *Personality and Social Psychology Review* 2:196–217. [http://dx.doi.org/10.1207/s15327957pspr0203\\_4](http://dx.doi.org/10.1207/s15327957pspr0203_4).
- Kline, Rex B., American Psychological Association, et al. 2004. "Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research." Washington D.C.: American Psychological Association.
- Masicampo, E. J. and Daniel R. Lalande. 2012. "A Peculiar Prevalence of *p* Values Just Below .05." *The Quarterly Journal of Experimental Psychology* 65:2271–79. <http://dx.doi.org/10.1080/17470218.2012.711335>.
- McCloskey, Deirdre N. and Stephen T. Ziliak. 1996. "The Standard Error of Regressions." *Journal of Economic Literature* 34:97–114.
- Nelson, Leif, Uri Simonsohn, and Joseph P. Simmons. 2014. "P-curve and Effect Size: Correcting for Publication Bias Using Only Significant Results." Available at SSRN: <http://ssrn.com/abstract=2377290>.
- Rosenthal, Robert. 1979. "The File Drawer Problem and Tolerance for Null Results." *Psychological Bulletin* 86:638. <http://dx.doi.org/10.1037/0033-2909.86.3.638>.
- Schwartz, Barry. 2004. *The Paradox of Choice: Why Less is More*. New York: Ecco.
- Simonsohn, Uri, Leif D. Nelson, and Joseph P. Simmons. 2014. "P-Curve: A Key to the File Drawer." *Journal of Experimental Psychology: General* 143:534–47. <http://dx.doi.org/10.1037/a0033242>.

Wainer, Howard. 1997. "Improving Tabular Displays, with NAEP Tables as Examples and Inspirations." *Journal of Educational and Behavioral Statistics* 22:1–30. <http://dx.doi.org/10.3102/10769986022001001>.

**Acknowledgements:** The author is grateful to Rebecca McDonald and members of the editorial team at *Sociological Science* for comments that have improved this article.

**Jeremy Freese:** Northwestern University.  
E-mail: [jfreese@northwestern.edu](mailto:jfreese@northwestern.edu).