Type M error might explain Weisburd's Paradox*

Andrew Gelman†, Torbjørn Skardhamar‡, and Mikko Aaltonen§

5 Mar 2017

Abstract

Simple calculations seem to show that larger studies should have higher statistical power, but empirical meta-analyses of published work in criminology have found zero or weak correlations between sample size and estimated statistical power. This is "Weisburd's paradox" and has been attributed by Weisburd, Petrosino, and Mason (1993) to a difficulty in maintaining quality control as studies get larger, and attributed by Nelson, Wooditch, and Dario (2014) to a negative correlation between sample sizes and the underlying sizes of the effects being measured. We argue against the necessity of both these explanations, instead suggesting that the apparent Weisburd paradox might be explainable as an artifact of systematic overestimation inherent in post-hoc power calculations, a bias that is large with small N. Speaking more generally, we recommend abandoning the use of statistical power as a measure of the strength of a study, because implicit in the definition of power is the bad idea of statistical significance as a research goal.

1. Introduction

Criminologists have a strong interest in policy evaluations, with the goal of increasing the use of evidence-based criminal justice policies and practices. Experiments and quasi-experiments play a big role in this research and policy agenda, and criminological research findings have considerable potential to influence (for better or worse) citizens' lives, given the immense reach of the criminal justice system. About a third of all Americans are arrested by the age of 23 and almost a fifth are convicted by that age (Brame et al., 2014). Around 1% of American adults are incarcerated in jails or prisons.

The field of criminology is just now beginning to engage with the problems documented in other fields surrounding null hypothesis testing, forking paths, false positives, and publication bias; hence the special issue to which this article is being submitted. In the present paper we take a pair of meta-analyses by Weisburd, Petrosino, and Mason (1993) and Nelson, Wooditch, and Dario (2014) as a springboard to discuss the ways that our understanding of quantitative evidence can be distorted by biases in estimates of effect size and power.

2. Type M and Type S error

Statistical tests are typically understood based on type 1 error (the probability of falsely rejecting a null hypothesis of zero effect, if it is in fact true) and type 2 error (the probability of *not* rejecting a null hypothesis that is in fact false). But this paradigm does not match up well with much of social science, or science more generally.

*For the *Journal of Quantitative Criminology*. We thank Justin Pickett and Gary Sweeten for suggesting this topic, several reviewers for helpful comments, and the U.S. National Science Foundation, Institute of Education Sciences, Office of Naval Research, and Defense Advanced Research Projects Agency for partial support of this work.

†Department of Statistics and Department of Political Science, Columbia University, New York

‡Department of Sociology and Human Geography, University of Oslo

§Institute of Criminology and Legal Policy, University of Helsinki

Our first problem with type 1 and type 2 errors is that in many problems we do not think the null hypothesis can be true. For example, a change in sentencing guidelines will certainly produce *some* changes in behavior; the question is how these changes vary across persons and situations. The second issue is that, in practice, when a hypothesis test is rejected (that is, when a study is a success), researchers and practitioners go with the point estimate of the magnitude and sign of the underlying effect. So, in evaluating a statistical test, we should be interested in the properties of the associated effect-size estimate, conditional on it being statistically significantly different from zero.

The type 1 and 2 error framework is based on a deterministic approach to science that might be appropriate for some classic statistical problems such as agricultural experimentation (one of the domains in which significance testing was developed in the early part of the last century), but it is much less relevant in a modern social-science environment of highly variable treatment effects.

With these concerns in mind, Gelman and Tuerlinckx (2000) introduced the concept of type S ("sign") and type M ("magnitude") errors, both of which can occur when a researcher makes a *claim with confidence* (typically, a p-value of less than 0.05 or a confidence interval that excludes zero, but more generally any statement that is taken as strong evidence of a positive effect. A type S error occurs when the sign of the estimated effect is of the opposite direction as the true effect. A type M error occurs when the magnitude of the estimated effect is much different from the true effect. A statistical procedure can be characterized by its *type S error rate*—the probability of an estimate being of the opposite sign of the true effect, conditional on the estimate being statistically significant—and its *expected exaggeration factor*—the expected ratio of the magnitude of the estimated effect divided by the magnitude of the underlying effect.

As discussed by Gelman and Carlin (2014), when a statistical procedure is noisy, the type S error rate and the exaggeration factor can be large. They give an example where the type S error rate is 24% (that is, any statistically significant estimate has a one-quarter chance of being in the wrong direction) and the expected exaggeration factor is 9.7. We discuss that example further in the section 5.

In this paper we are particularly concerned with type M errors, or exaggeration factors, which can be understood in light of the "statistical significance filter." Consider any statistical estimate. For it to be statistically significant, it has to be at least 2 standard errors from zero: if an estimate has a standard error of S, any publishable estimate must be at least $2S$ in absolute value. (Some nuances to this argument are discussed in section 4). Thus, the larger the standard error, the higher the estimate *must* be, if it is to be published and taken as serious evidence. No matter how large or small the underlying effect, the minimum statistically significant effect size *estimate* has this lower bound. This selection bias induces type M error. And, as we shall see, it can explain a well-known paradox that has been reported in quantitative research.

3. Weisburd's Paradox

Weisburd, Petrosino, and Mason (1993) document a surprising pattern in criminal justice experiments:

> Using sample estimates as a guide, the very largest investigations are no more powerful than the very smallest. Indeed, we find little relationship in practice between sample size and statistical power.

As they note, this finding is counterintuitive because for any fixed design the standard error will

decline with increasing sample size. If effect size is constant and standard error declines, statistical power will increase.

Speaking generally (not just in criminology research), one can imagine three possibilities for a lack of relation between sample size and power:

1. Larger experiments could be employed when studying smaller effects.

2. Studies could tend to decline in quality when they are made larger. As sample size increases, the bias and variance of individual measurements increases, enough to overwhelm the gain in precision achieved by larger samples.

3. Statistical power from small samples is not as large as researchers have believed. Crude calculations based on sample estimates will systematically overestimate the power attainable from small studies. The apparent lack of relation between sample size and power is actually just an artifact of a biased power estimate.

All three of these can be happening at once, but we shall consider them one at a time.

Explanation 1 is possible; indeed, very large effects such as a hypothetical study of the effectiveness of the guillotine can be ascertained with $n = 1$, whereas a savvy researcher will know that a large experiment is needed when studying a small effect. Nelson, Wooditch, and Dario (2014) perform a meta-analysis of experiments in criminology and report a strong negative correlation between sample size and effect size, but for reasons discussed below we suspect this pattern is merely an artifact of how these effect sizes have been reported. The statistical significance filter artificially inflates the apparent average effect size in small sample studies, and this as a result, artificially enhances the apparent negative correlation between sample and effect size.

Overall we doubt that explanation 1 could be behind Weisburd's paradox. In general, we would expect the resources of large-N experiments to go toward studying effects that are believed to be large and important, with smaller sample sizes being allocated to more speculative topics.

However, a related argument is that some *types of programs* are typically evaluated with smaller samples, whereas others are evaluated with larger samples (Slavin and Smith, 2009). In other words; type of intervention correlates with sample size. The 66 studies reviewed by Nelson et al. were themselves gathered from earlier systematic reviews, and we have combined the information about sample size of each study with the meta-analysis it belonged to. Figure 1 displays median sample sizes for each review. The studies at the bottom of the chart, with the smallest sample sizes, are the ones with the largest number of individual experiments. Egli et al. (2009), the study at listed at the very bottom, reports results in sufficient detail for us to readily calculate standardized effect sizes, which is not easily done from all the reviews, nor from many of the original papers. For these drug studies, the average standardized effect sizes is 0.6. Considering Nelson et al.'s Figure 3, there are ten studies of moderate to large effect sizes (Cohen's $d > 0.5$), and compared to the results reported by Egli et al., six of these are drug-studies. Thus, explanation 1 might contribute to the Weisburd paradox for example if drug rehabilitation programs have stronger effects than, say, employment programs.

Explanation 2 will depend on context. Weisburd, Petrosino, and Mason (1993) offer examples where increasing the size of a study makes it more difficult to exercise the control required to achieve low bias and variance. Nelson, Wooditch, and Dario (2014) consider the possibility that large studies are more likely to involve heterogeneous populations and thus less stable effects, and that in larger, more diverse studies, measurement error and infidelity to treatment assignment can be more of a concern, but found no strong evidence for this idea. On the other hand, the opposite pattern can

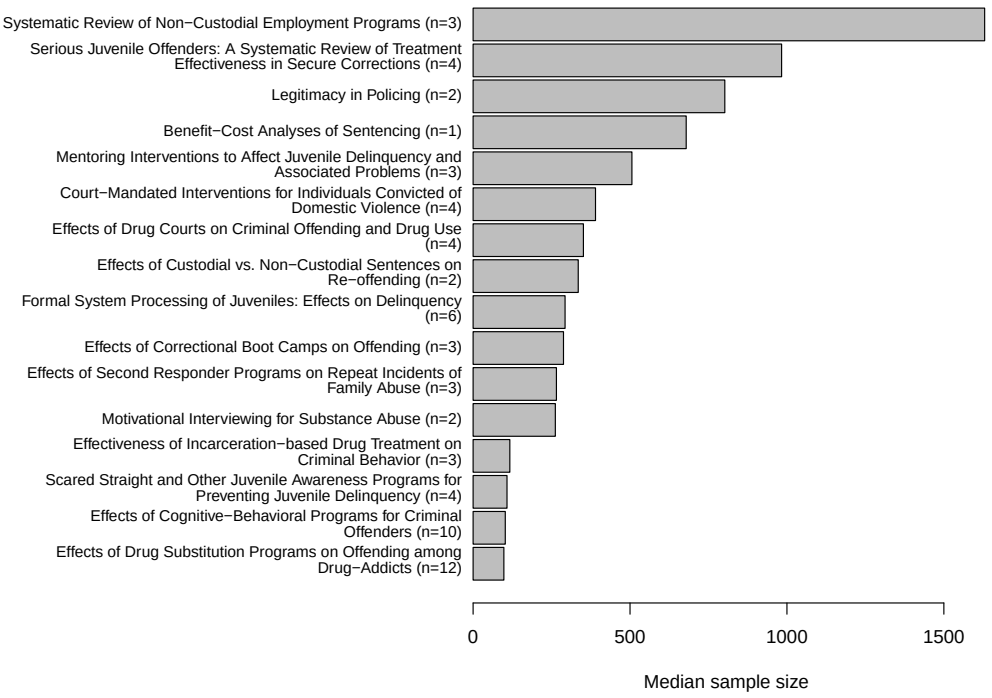

Figure 1: *Median sample sizes of studies included in a set of systematic reviews of criminology experiments reported by Nelson, Wooditch, and Dario (2014).*

also hold, in which large experiments are more careful and controlled, with smaller studies being more amateurish and more subject to error. Thus, we accept that some of the patterns identified by Weisburd et al. really happen in criminal justice experiments, but we think it is worth also considering other general explanations for the apparent lack of relationship between sample size and power.

This brings us to explanation 3. Published estimates are typically at least two standard errors from zero, as that is what it takes to be statistically significant at the 5% level. In every social and medical science field of which we are aware, the 5% level is taken as a threshold for a positive, and thus publishable, finding. For example, in a study of 2000 randomly selected abstracts in the Medline database, Ginsel et al. (2015) find that 82% of reported p-values were below 0.05. Jager and Leek (2014) find 75–80% of p-values below 0.05 in several leading journals in medicine and public health, Gerber and Malhotra (2008a,b) examine hypothesis tests published in three leading sociology journals and two leading political science journals and find well over half to be statistically significant. Indeed, in all this literature it is well accepted that the majority of published p-values are below 0.05, with the only controversy being the question of selection near this threshold (see, for example, Masicampo and Lalande, 2012).

And it is often not so difficult for researchers to find a statistically significant comparison in any dataset, whether via "p-hacking" using "multiple researcher degrees of freedom" (that is, trying out multiple data specifications and analyses until a statistically significant comparison is found; see Simmons, Nelson, and Simonsohn, 2011) or through the more innocuous "garden of forking paths" (making data coding and data analysis decisions after seeing the data, thus implicitly making data-dependent tests; see Gelman and Loken, 2014). Both p-hacking and forking paths invalidate p-values and power calculations by making it much more likely to attain statistical significance than would

be possible were analyses specified ahead of time (as is assumed by the statistical theory underlying standard calculations.

In the criminal justice context, Nelson, Wooditch, and Dario (2014) consider the possibility of publication bias. The implicit rule that statistical significance is required for publication, coupled with the ease of finding statistically significant comparisons, ensures that published estimates are systematically biased upward—and, as noted above, this statistical significance filter is a larger problem for smaller, noisier studies and can lead to large type M errors. And overestimate of effect sizes lead directly to overestimates of power (see, for example, Senn, 2002).

Here is what Weisburd, Petrosino, and Mason (1993) wrote:

> One solution used by others who have reviewed effect size across studies . . . is to take the mean of all of the outcome measures included by investigators. . . . Because we wanted some degree of consistency across the experiments reviewed, we developed an "average effect size" (AES) measure by taking the mean of all the effect coefficients at the follow-up period closest to one year. . . . we developed an additional measure—maximum effect size (MES)—that provides an upper range of effect for the experiments.

Both these definitions are subject to the statistical significance filter, and using these as effect size measures will overestimate power, with this bias being larger for small-N studies.

Thus, without making any judgments about the first two explanations listed above, it seems that explanation 3 can explain Weisburd's paradox all by itself. It is possible that explanations 1 and 2 also have validity, but the evidence provided by Weisburd et al. is not really enough to support such a claim, as their power calculations are based on published estimates, which automatically overestimate effect sizes, with this overestimation being particularly large for estimates with large standard errors and for non-preregistered studies.

One way this confusion can persist is that randomized experiments are celebrated for providing unbiased inference. But, even for randomized studies with unbiased measurements, the move to *statistically significant published estimates* leads to a bias that can be huge for small, noisy studies (recall Figure 1). So, clean design of individual studies does not guarantee freedom from bias for inferences such as those of Weisburd, Petrosino, and Mason (1993) which are based not on raw data but on published results.

In a more recent study, Nelson, Wooditch, and Dario (2014) replicated the meta-analysis of Weisburd et al. and again found no clear pattern between power and sample size. Nelson et al. report that "effect sizes decline as the sample size of the experiment increases," but this pattern can again be explained by the statistical significance filter. Nelson et al. report that "Publication bias, disclosure of fidelity issues, and year of publication do not account for a significant amount of variance in effect size," but this finding does not contradict our story in which all these effect size estimates are polluted by selection bias. Comparing effects in published vs. unpublished studies is not the same as significant vs. not significant, because the latter are often never even written up (Franco, Malhotra, and Simonovits, 2014).

4. Publication bias in experimental criminology?

The above reasoning implies the assumption that there is publication bias at work in criminology, or specifically in experimental criminology. If so, we would expect the majority of published findings to have $p < 0.05$. One meta-analysis of drug-court studies has concluded that there are indications of publication bias (Rothstein 2008). Relatedly, both the meta-analyses of Piquero et al. (2016)

and Wilson et al. (2011) found it reasonable to assume publication bias in criminology. However, Nelson, Wooditch, and Dario (2014) report that 68% of the studies in their review were not statistically significant, which might indicate that a "significance filter" is not a problem in experimental criminology.

The significance filter does not imply that every estimate in an article have to be statistical significant, but it is easier to build a story if one can report at least one estimate with $p < 0.05$, and that increases publication chances considerably. Sometimes, studies include additional subgroup analysis or effect estimates on alternative outcome measures. By increasing the number of estimates, the chances of a significant finding increase—and so do the publication chances even if the main result is not "statistically significant." Nelson et al. reported 402 effect estimates from 66 publications, implying the studies report 6 effect estimates on average. This leaves plenty of room for some of the estimates being statistically significant by chance alone. To get an indication of whether such concerns are warranted, we have taken a closer look at the studies included in the study by Nelson, Wooditch, and Dario (2014).

We wanted to reanalyze the dataset of Nelson et al. However, when we asked them for the data, they said they would only share the data if we were willing to include them as coauthors. We did not want to do so, and so we found the papers ourselves.[1] Of the initial 66 publications we were able to get a copy of 63; we do not doubt the existence of the other three papers but for our purposes it did not seem worth taking the effort to track them down. In total, 42 of these 63 papers reported at least one effect size where $p < 0.05$. Some of the studies include many outcome measures or treatment conditions, and several studies report more than 10 effect estimates. For example, Carroll et al. (2006) present 45 effect estimates of which 13 have $p < 0.05$, based on a sample of 136 marijuana users. That there are a large number of reported insignificant effect estimates in the literature is not contrary to a potential significance filter.

A non-significant finding might be easier to publish if it is a replication study that argues against earlier held beliefs or controversial treatments. For example, Carroll (2009) found no significant effect of motivational enhancement therapy for drug-offenders, arguing that such treatment does not work despite earlier beliefs. The same applies to evaluation of controversial policies such as "scared straight programs" (Lewis, 1983) or drug courts (Deschenes, 1995). That null findings are published in topics that are of considerable debate is not an argument against the significance filter, since they make valuable contributions to the literature precisely because they report negative results.

The statistical significance filter is ubiquitous in science, and based on the evidence one cannot dismiss its potential importance within the published in criminology literature. We have no good reason to suspect that criminology differs markedly from other fields of research in this regard. A more thorough examination of the literature would be needed to settle that issue, but in any case that the warnings in the following section hold more generally.

5. The problem with statistical power

The conventional view in applied statistics is that the reason for avoding low-power studies is that they are a waste of time and resources. For example, in the paper discussed above, Weisburd et al. write:

> Statistical power provides the most direct measure of whether a study has been designed to allow a fair test of its research hypothesis. When a study is underpowered it is

[1] Our summary is at `http://www.stat.columbia.edu/~gelman/documents/weisburd_table_of_studies.pdf`

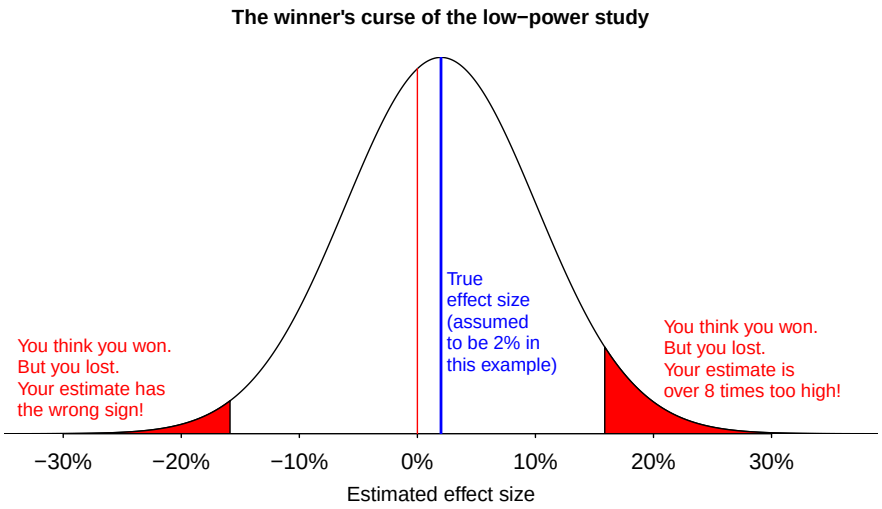

Figure 2: *When the standard error is small compared to the effect size, statistical power is low. In this diagram, the bell-shaped curve represents the distribution of possible estimates, and the red shaded zones correspond to estimates that are statistically significant (at least two standard errors away from zero). In this example, statistical significance is unlikely to be achieved, but in the rare cases where it does happen, it is highly misleading: there is a large chance the estimate has the wrong sign (a type S error) and, in any case, the magnitude of the effect size will be vastly overstated (a type M error) if it happens to be statistically significant. Thus, what would naively appear to be a "win" or a lucky draw—a statistically significant result from a low-power study—is, in the larger sense, a loss to science and to policy evaluation. From Gelman (2015).*

> unlikely to yield a statistically significant result even when a relatively large program or intervention effect is found.

This is correct but can mislead in that it too simply presents statistically significance as a goal.

To see the problem, suppose that a study is low power but can be performed for free, or for a cost that it is very low compared to the potential benefits that would arise from a research success. Then under the reasoning given in the above quote, a researcher might rationally conclude that a lower-power study is still worth doing, that it is a gamble worth undertaking.

The traditional threshold for power is 80%; funding agencies are reluctant to approve studies that are not deemed to have at least an 80% chance of obtaining a statistically significant result. But under a simple cost-benefit calculation, there would be cases where 50% power, or even 10% power, would suffice, for simple studies such as psychology experiments where human and dollar costs are minimal. Hence, when costs are low, researchers are often inclined to roll the dice, on the belief that a successful finding could potentially bring large benefits (to society as well as to the researcher's career).

The problem with this logic is that, in a low-power study, the seeming "win" of statistical significance can actually be a trap. Economists speak of a "winner's curse" in which the highest bidder in an auction will, on average, be overpaying. Research studies—even randomized experiments—suffer from a similar winner's curse, that by focusing on comparisons that are statistically significant, we (the scholarly community as well as individual researchers) get a systematically biased and overoptimistic picture of the world. When signal is low and noise is high, statistically significant patterns in data are likely to be wrong, in the sense that the results are unlikely to replicate.

To put it in technical terms, statistically significant results are subject to type S and type M errors, as described in Section 2. For example, Gelman and Carlin (2014) discuss a study where the true effect could not realistically be more than 2 percentage points and it is estimated with a standard error of 8.1 percentage points. Examining the statistical properties of the estimate using the normal distribution: conditional on it being statistically significant (that is, at least two standard errors from zero), it has at least a 24% probability of being in the wrong direction and is, by necessity, over 8 times larger than the true effect. Figure 2 demonstrates the winner's curse in the context of this example.

A study with these characteristics is dead on arrival: it has a power of about 6%, but it would be misleading to say it has even a 6% chance of success. In fact, arguably the 6% of the time that the study appears to succeed, represent the real failures, in that they correspond to ridiculous overestimates of treatment effects that are likely to be in the wrong direction as well. In such an experiment, to win is to lose.

Thus, the problem of a low-power study is not so much that it has a small chance of succeeding, but rather that an apparent success merely masks a larger failure. This point was made by Button et al. (2013) but does not seem to have yet been fully disseminated among research practitioners.

6. Discussion

In a meta-analysis of criminal justice experiments of various sizes and on various topics, Weisburd, Petrosino, and Mason (1993) found no empirical relation between sample size and estimated statistical power. They identified this as a paradox because, all else equal, power should increase with sample size. We resolve this paradox by refusing to accept the simple post-hoc power estimates which in turn are based on systematic overestimates of effect sizes. Especially for small, noisy studies, statistically significant effects can have high type M ("magnitude") errors, which can obscure an underlying relation between sample size and true statistical power.

Beyond this, we recommend following Nelson, Wooditch, and Dario (2014) and thinking of studies in terms of effect size and variation rather than statistical power. Power is the probability of obtaining a statistically significant difference in a fixed design, and we resist the framing of statistical designs using statistical significance as a goal. Noisy studies are not just a problem because they have "low power," thus unlikely to produce a successful result. No, the problem is worse: with sufficient noise, what seems like a success is actually a failure, in that statistically significant results are likely to be in the wrong direction and in any case will grossly overestimate effect sizes.

Hence we give the statistical advice to think less about power, both in evaluating patterns in past studies and in designing future experiments. Instead we recommend a greater acceptance of uncertainty, in particular, a willingness to publish non-statistically-significant results, and an understanding that results with p less than 0.05 can still fail to replicate in the general population.

That said, the mechanisms discussed by Weisburd et al. and Nelson et al.—explaining a hypothesized decline of effect sizes as studies get larger—are interesting and worthy of study in their own right. However, given the biases discussed above—selection in what studies are published and in what results are presented within any published study—will make it extremely difficult to study such patterns using the published record. Instead it could make more sense to study the relation between study size and effect size in a domain in which raw data were available.

Finally, this paper is also relevant to researchers who use power calculations in their own work. Suppose that, when planning an experiment or observational study, you determine the size of treatment and control groups based on a formal or informal power analysis determined by effect

sizes estimated from a literature review. If so, you should be aware that the magnitudes of published effect sizes are biased upward, both from the statistical significance fallacy and selection bias, hence power will be overestimated. This is a problem, wherever you as a researcher stand on null hypothesis significance testing.

References

Brame, R., Bushway, S., Paternoster, R., and Turner, M. (2014). Demographic patterns of cumulative arrest prevalence by ages 18 and 23. *Crime and Delinquency* **60**, 471–486.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., and Munafo, M. R. (2013). *Nature Reviews Neuroscience* **14**, 365–376.

Carroll, K. M., Easton, C. J., Nich, C., Hunkele, K. A., Neavins, T. M., Sinha, R., Ford, H. L., Vitolo, S. A., Doebrick, C. A., and Rounsaville, B. J. (2006). *Journal of Consulting and Clinical Psychology* **74**, 955–966.

Carroll, K. M., Easton, C.J., Nich, C., Hunkele, K. A., Neavins, T.M., Sinha, R., Ford, H. L., Vitolo, S.A., Doebrick, C. A and Rounsaville, B.J. (2009). The use of contingency management and motivational/skills-building therapy to treat young adults with marijuana dependence. *Journal of Consulting and Clinical Psychology* **77**, 993–999.

Deschenes E. P., Turner, S., and Greenwood, P. W. (1995). Drug court or probation?: An experimental evaluation of Maricopa County's drug court. *Justice System Journal* **18**, 55–73.

Egli, N., Pina, M., Christensen, P. S., Aebi, M., and Killias, M. (2009). Effects of drug substitution programs on offending among drug-addicts. *Campbell Systematic Reviews* 2009:3.

Franco, A., Malhotra, N., and Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science* **345**, 1502–1505.

Gerber, A. S., and Malhotra, N. (2008a). Publication bias in empirical sociological research: Do arbitrary significance levels distort published results? *Sociological Methods and Research* **37**, 3–30.

Gerber, A. S., and Malhotra, N. (2008b). Do statistical reporting standards affect what is published? Publication bias in two leading political science journals. *Quarterly Journal of Political Science* **3**, 313–326.

Gelman, A. (2015). Statistics and the crisis of scientific replication. *Significance* **12** (3), 23–25.

Gelman, A., and Carlin, J. B. (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science* **9**, 641–651.

Gelman, A., and Loken, E. (2014). The statistical crisis in science. *American Scientist* **102**, 460–465.

Gelman, A., and Tuerlinckx, F. (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics* **15**, 373–390.

Ginsel, B., Aggarwal, A., Xuan, W., and Harris, I. (2015). The distribution of probability values in medical abstracts: An observational study. *BMC Research Notes* **8**, 721.

Jager, L. R., and Leek, J. T. (2014). An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics* **15**, 1–12.

Lewis, R. V. (1983). Scared straight—California style. Evaluation of the San Quentin squires program. *Criminal Justice and Behavior* **10**, 209–226.

Masicampo, E. J., and Lalande, D. (2012). A peculiar prevalence of p values just below .05. *Quarterly Journal of Experimental Psychology* **65**, 2271–2279.

Nelson, M. S., Wooditch, A., and Dario, L. M. (2015). Sample size, effect size, and statistical power: A replication study of Weisburd's paradox. *Journal of Experimental Criminology* **11**, 141–163.

Patrick, S. and Marsh R. (2005). Juvenile diversion: Results from a 3-year experimental study. *Criminal Justice Policy Review* **16**, 59–73.

Senn, S. J. (2002). Power is indeed irrelevant in interpreting completed studies. *British Medical Journal* **325**, 1304.

Weisburd, D., Petrosino, A., and Mason, G. (1993). Design sensitivity in criminal justice experiments. *Crime and Justice* **17**, 337–379.