Article type: Commentary

Empowering the underpowered? A comment on Nelson, Wooditch, and Dario (2015).

By Torbjørn Skardhamar and Mikko Aaltonen

In a review of randomized controlled studies in criminal justice, Weisburd, Petrosino, and Mason (1993) observed that studies with larger sample sizes had *lower* statistical power than smaller studies. Challenging the conventional view that larger samples increase statistical power, this phenomenon was later titled the "Weisburd paradox" (Sherman 2007: 305; Hinkle et al. 2012: 222). Nelson, Wooditch, and Dario (2015) (NWD, henceforth) have recently conducted a replication of Weisburd, Petrosino, and Mason (1993) analysis, and confirmed their conclusion on power, stating that: "No consistent relationship exists between sample size and statistical power in the real world" (p. 153). Since this observation runs counter to statistical theory, they advise that "further development of the Weisburd paradox literature base is critical" (p.156). They further suggest that the Weisburd paradox implies that it is preferable to conduct small studies with large effect sizes than larger studies with lower effect sizes even if the latter provide more reliable estimates (p. 152). In what follows, we take issue with this recommendation. We argue that (1) the conclusions reached by NWD are based on the misunderstanding that calculating "achieved power" is not formative on what can be learned from a study, (2) there is nothing paradoxical about "the Weisburd paradox", and (3) that small sample studies have important additional limitations that should be taken into account in evaluations of research designs.

Statistical power

In statistics, the concept of power refers to the probability of correctly rejecting the null hypothesis, and power calculations are typically performed when designing a study to ensure prospective data are sufficient to detect hypothesized effect(s) (Britt and Weisburd 2010). Statistical power relies on three parameters: the true effect size, the variance of the outcome, and the chosen level of significance. As standard errors decrease with sample size, power can be enhanced by increasing the sample size. The principle is straightforward: one should avoid collecting data or conducting an experiment unless the study has the potential to reliably estimate the parameter of interest with sufficient precision. The size of the experiment should be large enough to permit inferences with reasonable confidence about the parameter. It follows that if the variance is large and the true effect is small, one will not be able to tell if the effect estimate is statistically significant from zero. Simply put, you need larger samples to detect small effects, but you might detect large effects with smaller samples.

However, one typically does not know the true effect size, so one need to use your best guess which is often based on prior studies. One might also consider calculating the power for the minimum effect size of any substantive interest. The power calculations are then done based on one's best knowledge, and the only parameter left for manipulation is the sample size. Still, it is reasonable to design a study to get as large an effect as possible, which also enhances power. Gelman and Hill (2006) briefly discuss how medical studies might e.g. increase dosages rather than sample size to maximize power. Such a strategy can be more problematic in the social sciences as the "generalization to more realistic levels can be suspect" (Gelman and Hill 2007: 439).

Achieved power is a non-informative metric

The analysis of Weisburd, Petrosino, and Mason (1993) and NWD is based on post-hoc calculations of statistical power, and referred to as "actual achieved power". According to Britt and Weisburd (2010) there is little consensus about the appropriateness of retrospective power analysis, but it is "informative in the sense that the results will indicate to the researcher using these data sources what the achieved dataset can and cannot tell them about the statistical relationships they may be most interested in" (p. 323). Contrary to this position, our understanding of the statistical literature suggests clear consensus regarding the (in)appropriateness of retrospective power analysis. Analyses of achieved power are generally considered futile by statisticians (Hoenig and Heisey 2001; Gelman and Carlin 2014; O'Keefe 2007; Senn 2007; Button et al. 2013). The medical statistician Stephen Senn writes:

> A power calculation is used for planning trials and is effectively superseded once the data are in. . . . An analogy may be made. In determining to cross the Atlantic it is important to consider what size of boat it is prudent to employ. If one sets sail from Plymouth and several days later sees the Statue of Liberty and the Empire State Building, the fact that the boat employed was rather small is scarcely relevant to deciding whether the Atlantic was crossed. (Senn 2007: 209)

In a similar vein, O'Keefe (2007: 293) states – sarcastically, we assume – that the achieved power provides the answer to the question: "What chance was there of producing a statistically significant result, assuming that the population effect is exactly equal to the observed sample effect?" In most situations, this is not a question of interest. Without any reference to external information about true effect sizes, the analysis of achieved power does not bring anything new to the table (Gelman & Carlin 2014).

What paradox?

The apparent paradox is that achieved power is not consistently related to sample size, but what this means is that smaller studies tend to provide larger effect estimates. Weisburd, Petrosino, and Mason (1993) and NWD suggest that smaller studies report larger effect sizes because they tend to be qualitatively better than larger studies. This is so because increasing the sample might alter the experiment or the sample characteristics in unintended ways. For example, the experiment can be more difficult to implement with high integrity: managing attrition and ensuring correct dosages may be more challenging as studies get larger. Thus, the benefit gained from increased sample size is offset by decreased implementation quality and lower true effect size. Another possibility is that, in order to scale up, it may be necessary to relax eligibility criteria and to include participants for whom

the treatment is likely to have less impact. This would also reduce the true effect, resulting in lower statistical power of the study.

We think it is hardly paradoxical that studies with implementation problems have smaller effect sizes than well-implemented studies. We obviously agree with the recommendation that any study, small or large, should be implemented with high integrity, and that one should not compromise study quality in order to increase the sample size. Sometimes there are tradeoffs between integrity and sample size, and the consequences of such tradeoffs are worth investigating.

One of the underlying issues in NWD is the scaling of the trial, which we agree is important but for reasons unrelated to statistical power. If NWD and Weisburd, Petrosino, and Mason (1993) are correct to assume that scaling up compromises implementation, scaling down is not necessarily a good option. If the treatment under investigation is intended to be applied on a larger scale, it would be important to know if a realistic implementation on that scale works equally well. If a treatment is only successful in small settings, this implies that large scale implementations will be ineffective. If NWD intended to draw attention to the broader issue of scaling of experiments and implementing programs on a much larger scale than the initial study, we agree that this is an issue worth exploring further.

Publication bias, "p-hacking" and exaggerated effects

NWD discuss the possibility that the Weisburd paradox (smaller studies report larger effect sizes) might be a result of publication bias, where Null findings are less likely to be published. They argue that this is not a likely explanation based on their own analysis. First, they included both published and unpublished studies in their review, and argue that there is no notable publication bias in their sample. Considering their overview of studies included (Nelson, Wooditch, and Dario 2015: 156-160), it seems like by "unpublished" they mean that the article was not published in an international journal. It does not mean ending up in a file drawer or not being made publicly available at all. It is the file drawer which represents the main publication bias, and NWDs data do not include any studies taken from the file drawer.

Second, NWD point out that since the average "achieved power" is .32 in the reviewed studies, the average reviewed study consequently had a 68% chance of not finding a significant effect (Nelson et al. 2015: 156). This seems to suggest that criminological RCTs often get published even in the absence of significant results, which is counter to what would follow from publication bias. However, the review includes 402 outcome measures from a total of 66 studies, so if there is one effect estimate per outcome, there are on average about six estimates per study. It is not reported how many of the 66 studies do not report any significant effects, but we suspect that most of them report at least one significant estimate. It might be that any publication bias would primarily be related to the main estimate and non-significant additional outcomes do not affect publication chances. It is also possible that a non-significant main result is easier to get published if there are some significant effects in additional outcomes or for some sub-groups. In our opinion, there are no reasons to expect less publication bias in criminology – in both experimental and observational studies – than in other fields of research (Head et al. 2015).

If non-findings are harder to get published, estimates from small trials in published studies are likely to overestimate the underlying true effects (Gelman and Carlin 2014). This phenomenon has been

called “the winner’s curse” (Button et al. 2013: 373). As Button et al. write, it has been more widely recognized that underpowered studies might fail to find significant effects, while the opposite problem of underpowered studies producing inflated effects has received less attention. The “winner’s curse” describes the phenomenon of published studies with low power producing inflated estimates of true effect sizes. If the true effect is small and the sample size is small, the only result that can emerge as statistically significant is an overestimate of the true effect. Estimates from smaller samples will not be large enough (under reasonable levels of significance) to separate a small effect from a null effect. Thus, “winner’s curse” is essentially related to the use of statistical significance as a “screener” or threshold that the study needs to reach for it to be published and considered relevant.

As discussed by e.g. Head et al. (2015), the results can be affected by a wide range of practices such as recording many response variables and reporting only selected ones, deciding on whether or not to drop outliers, excluding, combining or splitting treatment groups, including or excluding covariates, and stopping data exploration when finding significant (or otherwise interesting) results. If the researcher runs many models and only reports the significant ones, this is sometimes referred to as “p-hacking” or “data snooping”.

While the term “p-hacking” implies intentional massaging of data, Gelman and Loken (2014) argue that similar results can arise unintentionally from multiple comparisons. Their argument is basically that in many studies there are many decisions to be made about definitions, modelling techniques and specifications, exclusion of some observations and so forth, which can all be reasonable, but imply multiple comparisons (Gelman and Loken 2014). Since small studies are likely to be more sensitive to minor adjustments of the analysis, this could affect the average effect sizes.

We cannot see that the effects of publication bias and p-hacking can be ruled out as potential explanations for the Weisburd paradox, and would rather encourage others to do a thorough analysis (for one example, see Head et al. 2015) of such bias in both experimental and observational studies in criminology. Pre-registration of randomized controlled trials could be one way forward, and some evidence from the medical sciences shows that the number of papers publishing null findings has increased after the establishment of clinical trial registry (Kaplan and Irvin 2015). However, the small - albeit increasing (Telep, Garner, and Visher 2015) - number of trials in our field might make a pre-registration system devoted only to criminology an unrealistic target. Preregistration would not, however, solve problems of p-hacking and multiple comparisons unless also the details of the modelling decisions are registered to a high degree in advance.

Conclusion

In our opinion, the Weisburd paradox is not paradoxical at all, but rather an expected empirical pattern on basis of statistical theory and publication practices. NWD argued against the assumption of “more people, more power” (p. 142). We would prefer to put it differently: All else equal, high-quality treatments are more likely to yield larger effect sizes. This applies to all studies regardless of the sample size. A more serious concern is that low-power studies are more likely to yield exaggerated effect sizes – even if the study is of good quality.

Calculating “achieved power” is motivated by informing the researcher what the dataset at hand can and cannot tell them about how much one should trust the results (Britt and Weisburd 2010: 323). A

far more promising approach to that end is the "design analysis" proposed by Gelman and Carlin (2014). To do so one has to rely on external information in a similar way that one would when conducting a prospective power analysis. They suggest to calculate what they call the "exaggeration ratio", the magnitude in which the estimated effect might be exaggerated given the current design, providing that a statistically significant result is discovered. They show that one should be concerned about exaggerated results if the *prospective* power is less than .50. In studies with even less power, one should also be concerned that the estimate might even have the wrong sign (Gelman and Carlin 2014).

The problem with calculating "achieved power" is that it might give the impression that underpowered studies are far more powerful than they really are. In fact, the "achieved power" is uninformative and does not provide more insight than what can be concluded from inspecting the estimates and standard errors. However, statistical power remains very important *when designing* the study (Britt and Weisburd 2010; Vuolo, Uggen, and Lageson 2015).

We certainly agree with both NWD and Weisburd, Petrosino, and Mason (1993) that ensuring implementation quality is an important part of any experiment that should not be sacrificed just to increase sample size. Clearly, one would expect a larger effect of an intervention that actually works as intended than an intervention that fails to do so. This is a crucial issue to consider when designing an experiment, and should indeed be considered when prospectively calculating statistical power. It is also important that the experiment is scaled according to substantive interest, so that programs intended to be implemented broadly should be scaled accordingly. There is little help in evaluation of an intervention that only works well on small samples if the goal is to implement the intervention more broadly.

In conclusion, the dangers of conducting underpowered studies are as follows: First, it is more likely to not find a significant treatment effect at all even though it exist. This is the conventional interpretation of low statistical power, and also the concern raised by NWD in their conclusion. Second, if a significant effect is found, it is likely to be much larger than the true effect. If (intentional or unintentional) p-hacking is prevalent in criminological literature, like in other disciplines (Head et al. 2015), then any meta-analysis of effect sizes is expected to find larger effects in smaller studies. While we find it plausible that implementation problems contribute to this empirical regularity as well, recent discussions about the shortcomings of small trials in the larger scientific community (Button et al. 2013, Gelman & Carlin 2014) should not be overlooked in criminology, particularly as the available evidence (Nelson, Wooditch, and Dario 2015) suggests that true effect sizes in many criminological RCT's are likely to be modest.

We advice to only calculate power prospectively and never use "achieved power" to justify too small studies. Generally, a study should be implemented and scaled to capture the effect of interest, not just to maximize effect size.

References

Britt, Chester L., and David Weisburd. 2010. Statistical power. In *Handbook of quantitative criminology*, edited by A. R. Piquero and D. Weisburd. New York: Springer-Verlag.

Button, Katherine S., John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson, and Marcus R. Munafò. 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nature* 14 (May):365-376.

Gelman, Andrew, and John Carlin. 2014. Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science* 9 (6):641-651.

Gelman, Andrew, and Jennifer Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Gelman, Andrew, and Eric Loken. 2014. The statistical crisis in science. *American scientist* 102:460-5.

Head, Megan L. , Luke Holman, Rob Lanfear, Andrew T. Kahn, and Michael D. Jennions. 2015. The Extent and Consequences of P-Hacking in Science. *PLoS Biology* 13 (3):e1002106.

Hinkle, Joshua C., David Weisburd, Christine Famega, and Justin Ready. 2012. The problem is not just sample size: The consequences of low base rates in policing experiments in smaller cities. *Evaluation Review* 37 (3-4):213-238.

Hoenig, John M., and Dennis M. Heisey. 2001. The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis. *American statistician* 55 (1):1-6.

Kaplan, Robert M., and Veronica L. Irvin. 2015. Likelihood of Null Effects of Large NHLBI Clinical Trials Has Increased over Time. *Plos One* 10 (8):e0132382.

Nelson, Matthew J., Alese Wooditch, and Lisa M. Dario. 2015. Sample size, effect size, and statistical power: a replication study of Weisburd's paradox. *Journal of Experimental Criminology* 11 (1):141-63.

O'Keefe, Daniel O. 2007. Post hoc power, observed power, a priori power, retrospective power, prospective power, achieved power: Sorting out appropriate uses of statistical power analyses. *Communication methods and measures* 4:291-99.

Senn, Stephen. 2007. *Statistical Issues in Drug Development, 2nd Edition*. Chichester, England: John Wiley & Sons.

Sherman, Lawrence W. 2007. The power few: experimental criminology and the reduction of harm. The 2006 Joan McCord Prize Lecture. *Journal of Experimental Criminology* 3 ( ):299-321.

Telep, Cody W., Joel H. Garner, and Christy A. Visher. 2015. The production of criminological experiments revisited: the nature and extent of federal support for experimental designs, 2001–2013. *Journal of Experimental Criminology*.

Vuolo, Mike, Christopher Uggen, and Sarah Lageson. 2015. Statistical power in experimental audit studies: Cautions and calculations for matched tests with nominal outcomes. *Sociological methods and research*.

Weisburd, David, Anthony Petrosino, and Gail Mason. 1993. Design Sensitivity in Criminal Justice Experiments. *Crime and Justice* 17:337-379.