ONLINE APPENDICES

Was the Pope to blame? Statistical powerlessness and the predictive policing of micro-scale randomized control trials

Ralph B. Taylor & Jerry H. Ratcliffe

Department of Criminal Justice

Center for Security and Crime Science

Temple University

Online at:

www.rbtaylor.net/supplemental/pub_cpp_2020_app.pdf

APPENDIX A: Power estimation, approaches, and parameter choices

This section outlines the exact combination of power estimation levels and parameter choices that might help some readers understand the rest of the article. It also provides some more technical background on statistical power. Of course, these issues are well explained elsewhere. Cohen (1988) provides the standard introduction and Hinkle et al. (2013), as described in the main article, consider power issues in the context of hot spots policing experiments with low base rates for the outcome indicators. Next, more information is provided on the two different ways that power was calculated here. The standard approach (using a Stata routine called *pc_twoproportions*) used only mean proportions and did not interrogate the actual data. This contrasts with the second approach that relied on a program called *pc_simulate,* part of a Stata add-on package called *pcpanel* created by Louis Preonas. This module relied on

information gained from the experiment itself, the control and awareness conditions, to fine-tune the analysis. This simulation approach, like the standard approach, also could be used for a priori power calculations with baseline data. The point here was not to see which approach was better. The two were just different. Rather the question was: do the two approaches generally agree on the severity of the statistical powerlessness problem under different scenarios, and the conditions likely to produce acceptable levels of power?

First an aside on statistical power. What is it?

> The statistical power of a significance test is the long-term probability, given the population ES [effect size], α [alpha level including whether one or two tailed], and *N* [sample size] of rejecting H_0 [null hypothesis]. When the ES is not equal to zero, H_0 is false, the failure to reject it also incurs an error. This is a Type II error, and for any given ES, α, and *N*, its probability of occurring is β. Power is thus 1 - β, the probability of rejecting a false H_0 (Cohen, 1992: 156).

Stated more simply, if you had a magical instrument that let you know when there *really* was a difference on the outcome, of a specified size, between control and treatment conditions, what chances would your statistical analysis have of labeling this difference as statistically significant rather than just noise? Again, your magical instrument tells you this difference is really there and is sizable.

What level of statistical power does the researcher need? “A convention proposed for general use” is power of 80 percent or .80. (Cohen, 1992: 156). “A materially smaller value than .80 would incur too great a risk of a Type II error. A materially larger value would result in a

demand for *N* that is likely to exceed the investigator's resources. Taken with the conventional α = .05, power of .80 results in a $\beta:\alpha$ ratio of 4:1 (.20 to .05) of the two kinds of risks" (Cohen, 1992: 156). [1]

The key contrast in findings from the predictive policing experiment (Ratcliffe et al., 2020) was between the control condition and the marked patrol condition. Therefore, in the standard power estimation approach (*power_twoproportions* in Stata v. 15) the focus was on just those two conditions when using the standard statistical power estimation approach. To keep the analysis straightforward and explicitly aligned with the experimental design, here results from *just* the mission areas, the three predicted 500' x 500' grids each day in each district, were used. The focus is just on crime occurrence vs. nonoccurrence during the treatment shift, not crime counts. In effect then, these are reported property crime prevalence rates. These prevalence rates appeared in Table 6 in the final project report (Ratcliffe, Taylor, Askey, Fisher, & Koehnlein, 2019). These results are different from those reported in the journal article describing key quantitative results (Ratcliffe et al., 2020) which used buffered mission areas.

Stata program '*power_twoproportions*', accommodates cluster-randomized designs. It required some parameter decisions, which we describe here to aid future replication.

The number of units (clusters) at the level of randomization are important. In the program, parameter *k1* reflects the number of clusters, districts in our case, randomly assigned to

---

[1] Cohen, as a psychologist, was primarily concerned with academic psychologists planning how many subjects they will need for their experiments. Other types of social scientists may desire higher levels of statistical power, or, when working with extremely large data sets, may opt to use much stricter significance levels like $p < .001$ so that in their analyses they are not overwhelmed with significant findings (e.g., Baldassare (2000)).

That said, "although there are risks associated with having too much statistical power, the pressing need for most social science research is to have much more of it" (Ellis, 2010: 81).

the control group, and the parameter *k2* is the number assigned to the marked car treatment condition. In the property crime experiment, these were 5 and 5.

The program also requests the size of each cluster, *m1* and *m2*, respectively, for the control and treatment conditions. Here, cluster size reflected the number of shifts, or days, studied. In the property crime phase of the experiment, these were 90 and 90.

In a cluster-randomized trial, power is shaped by the intracluster correlation coefficient, the fraction of outcome variation associated with the level of randomization. In our case, the randomization level was the police district.[2] Including this in analyses estimating needed sample sizes takes care of the needed "inflating by the design effect (DE), which is a simple function of the intracluster correlation (ICC)" (Hemming & Marsh, 2013: 114). A mixed effects logit model analyzing just the control and marked car districts yielded ICC = .1346. [3] This parameter was retained in all power estimations with the standard power approach. Further, throughout, just one tailed testing was used at the default alpha level, $p < .05$.

The power estimations provided by *power_twoproportions*, where the power estimates did not depend on records in specific data files, allowed exploring each alternate reality on its own. It also permitted examining combined scenarios to see if certain alternate realities, generating insufficient statistical power on their own, could generate sufficient statistical power when paired with other scenarios. In other words, with the standard approach, one could explore combinations of alternate realities. It was not possible to explore these combinations with the

---

[2] The "intracluster correlation coefficient (ICC) … is a measure of the relatedness of clustered data … mathematically it is the between-cluster variability divided by the sum of the within-cluster and between-cluster variabilities" (Killip, Mahfoud, & Pearce, 2004: 206)

[3] Obtained with the post estimation command *estat icc* which yielded the intraclass correlation coefficient, which in this instance is equivalent.

simulation approach based on real data because different alternative realities sometimes relied on data structured in different ways.

Sample Stata do files can be found online at www.rbtaylor.net/pub_cpp_2020_sample_do.pdf

APPENDIX B: Predictive Policing Statistical Powerlessness and Hot Spots Results: Thoughts on Discrepancies

One reviewer posed the following excellent question: “There are hot spot studies that have found statistically significant results for specific types of crime that do occur more rarely. How is that consistent with their argument?”

The short answer, as noted in the main paper, had several parts. (1) Some of these hot spots studies have used calls for service, which occur more frequently than reported Part I serious crime incidents. The latter are the preferred outcome in predictive policing studies. (2) Even when using reported Part I serious crime incidents, not all studies have looked at outcomes by crime type. At least one study looked at just total Part I incidents. (3) Not all hot spot studies looking at specific Part I serious crime category counts for outcomes have used randomized control trial experimental designs. More rigorous designs may make it less likely to observe significant treatment effects. (4) Further, not all rigorously designed hot spot studies looking at serious crime outcomes have observed statistically significant treatment impacts in the anticipated direction. And finally, (5) some rigorously designed hot spots studies with serious Part I reported crime outcomes have used treatment areas that were larger, when comparisons could be made, than typically used in predictive policing studies. If the outcome is crime

prevalence rates, crime occurrence/non-occurrence, rates go up with larger areas. The details behind this short answer appear below.

Starting point: A Recent systematic review

Fortunately, Braga and colleagues (2019) published a recent systematic review of hot spots studies at micro places. “All studies using units of analysis smaller than a neighborhood or community were considered” (Braga et al., 2019: 5/88). Tables 2 & 3 in that publication provide key details. Going to published articles themselves provided additional crucial details. We started sub-setting the studies included there as follows.

Subset 1: Rigorous experimental design

The first cut was on experimental design “Twenty‐seven eligible studies used randomized controlled trials” (Braga et al., 2019: 9/88).

Subset 2: Serious (Part I) reported crime outcome

Among these 27, the next cut was on the outcome. Did the study include a Part I reported crime outcome, or grouping of Part I reported crime types? Excluding three studies that did not appear in refereed journals (In Braga et al. (2019: Table 2): Sherman et al., 2014; Atterman, 2017; Blattman et al., 2017), left 13 studies that included at least one Part I crime outcome based on crime incident reports. In the order listed in Table 2 these were: (1) Braga et al. (1999); (2) Taylor et al. (2011); (3) Ratcliffe et al. (2011); (4) Lum et al. (2011); (5) Rosenfeld et al. (2014); (6) Telep et al. (2014); (7) Groff et al. (2015); (8) Koper et al. (2015); (9) Weisburd et al. (2015); (10) Ariel et al. (2016); (11) Santos & Santos (2016); (12) Ariel & Partridge (2017); (13) Ratcliffe et al. (2019).

Subset 3: Statistically significant crime impact in expected direction

Among these thirteen studies, some reported non-significant treatment impacts for a serious reported crime outcome. For example Ariel et al. (2016: 297) reported "the effect size for reported crime data yields an effect size of d =-.189 (95 % CI -.653, .27)." Since the confidence interval crosses zero, this is a non-significant finding. Santos & Santos (2016: 391) reporting on their Part I outcome indicated: "The direction of the intervention variables in both models shows that the treatment areas had lower counts of residential burglary and theft from vehicle crimes and that more intervention dosage was related to fewer crimes; however, the predictors were not significant in either model." At least one study reported an unexpected outcome. Ariel & Partridge (2017: 813) found that "victim generated crimes—the primary outcome measured in nearly all hot spots policing experiments—increased."

Subset 4: Rigorous design, crime outcome, significant results in expected direction: Areal comparison

So narrowing further, and focusing just on the rigorous studies that found significant treatment impacts with a serious crime outcome in the expected direction, were the treatment areas of comparable size to those used in predictive policing? In the current study, three 500' x 500' grids were generated for each shift. In toto, this represented 750,000 square feet or 2.69 percent of a square mile.

At least one hot spots study had treatment areas appreciably larger than used in this predictive policing experiment. Groff et al. (2015: 28) reported that their hot spots averaged ".044 square miles." Their average hot spot was therefore 63 percent larger in area than the three mission grids per shift per district used in the current study.

For numerous studies, however, it was not possible to compare the square footage of the treatment areas because many hot spot studies reported geographic area in non-comparable ways. Studies often reported numbers of intersections or streetblocks or miles of streets. Some examples. Braga et al. (Braga, Weisburd, Waring, Lorraine Green, & et al., 1999: 549) described their hot spots as intersection areas where "an intersection area is the intersection and its four adjoining street segments." Rosenfeld et al. (2014: 433) reported "the final 32 hot spots contained 258 street segments, with an average of eight segments per area (standard deviation [SD] = 2.84, minimum = 3, maximum = 14)."

Finally, as mentioned earlier, at least one study finding a significant treatment effect on reported crime used total Part I crime incidents, rather than individual Part I crime categories (Telep, Mitchell, & Weisburd, 2014: Table 3)

Takeaway

Hot spot studies showing statistically significant impacts of a policing intervention on specific Part I reported crime categories do appear in the literature. Our questioning reviewer is absolutely correct, of course, on that point. But it is not clear at this time what specific fraction of those studies proved comparable to the current predictive policing study on *all* of the following attributes: it was a randomized controlled trial experiment, appearing in a refereed journal, considering specific Part I reported crime incident outcomes, using comparably sized treatment areas of 750,000 square feet. Because *this* subset of hot spot studies cannot be identified, we therefore cannot know the fraction of this subset that found a statistically significant ($p < .05$) impact in the expected direction. In short, it is not at all clear, given current available information whether the statistical powerlessness problem highlighted here aligns or fails to align with what we know from hot spot studies of policing interventions.

APPENDIX C: Do file excerpts

Estimates using power twoproportions (Standard approach)

```
* keeping just the two key conditions
****************************
*************************
/* NOTE - HERE WE *** ARE *** KEEPING CONTROL AND MARKED
* THIS IS CORRECT
* THIS IS DIFFERENT FROM WHAT WE DID WITH THE PC SIMULATE
*/
keep if(expcond==1)|(expcond==3)
tab during01 expcond, col

melogit during01 ||district:
estat icc
/*
 estat icc

Intraclass correlation

------------------------------------------------------------------------------
                       Level |        ICC   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
                    district |   .1345848   .1262944      .0182303     .565679
------------------------------------------------------------------------------

*/

* as conducted
power twoproportions .0333 .0133, k1(5) k2(5) m1(90) m2(90) rho(0.1346) onesided

* AR-1 - double time of experiment
power twoproportions .0333 .0133, k1(5) k2(5) m1(180) m2(180) rho(0.1346) onesided

* AR - 2 - ALL EGGS IN ONE BASKET
power twoproportions .0333 .0133, k1(5) k2(15) m1(90) m2(90) rho(0.1346) onesided

* AR - 3 - Philadelphia is London - 23 districts each
power twoproportions .0333 .0133, k1(23) k2(23) m1(90) m2(90) rho(0.1346) onesided

* AR - 4
display .0333*5
display .01333*5
/*
. display .0333*5
.1665

. display .01333*5
.06665
*/

* 5 times higher
power twoproportions .1665 .06665, k1(5) k2(5) m1(90) m2(90) rho(0.1346) onesided

* ten times higher
power twoproportions .333 .133, k1(5) k2(5) m1(90) m2(90) rho(0.1346) onesided

* 15 times higher
display .0333*15
display .01333*15

/*
. display .0333*15
.4995

. display .01333*15
.19995
*/
power twoproportions .4995 .1995, k1(5) k2(5) m1(90) m2(90) rho(0.1346) onesided
```

```
* Ten times larger and  Philadelphia as London (Ar-4-10x / AR-3)
power twoproportions .333 .133, k1(23) k2(23) m1(90) m2(90) rho(0.1346) onesided

* 5 times larger and philadelphia as london
power twoproportions .1665 .06665, k1(23) k2(23) m1(90) m2(90) rho(0.1346) onesided

* Ten times larger and all eggs in one basket (AR-4-10X / AR-2)
power twoproportions .333 .133, k1(5) k2(15) m1(90) m2(90) rho(0.1346) onesided

* FIVE times larger and all eggs in one basket (AR-4-5X / AR-2)
power twoproportions .1665 .06665, k1(5) k2(15) m1(90) m2(90) rho(0.1346) onesided

* Ten times larger and no Papal visit (AR-4-10x / AR-1)
power twoproportions .333 .133, k1(5) k2(5) m1(180) m2(180) rho(0.1346) onesided

* FIVE times larger and no Papal visit (AR-4-10x / AR-1)
power twoproportions .1665 .06665, k1(5) k2(5) m1(180) m2(180) rho(0.1346) onesided


* No Papal visit and all eggs in one basket (AR-1 / AR-2)
power twoproportions .0333 .0133, k1(5) k2(15) m1(180) m2(180) rho(0.1346) onesided


* No Papal visit and Philadelphia as London (AR-1 / AR-3)
power twoproportions .0333 .0133, k1(23) k2(23) m1(180) m2(180) rho(0.1346) onesided

* All eggs in one basket and Philadelphia as London (AR-2 / AR-3)
power twoproportions .0333 .0133, k1(11) k2(35) m1(90) m2(90) rho(0.1346) onesided
power twoproportions .0333 .0133, k1(12) k2(34) m1(90) m2(90) rho(0.1346) onesided
* TAKE THE AVERAGE OF THE ABOVE 2 RESULTS
display ((.347+.3439)/2)
clear all
log close
exit
```

Simulations using pcpanel

Estimating power for experiment as conducted – not absorbing fixed effects for time

```
* USING ONLY CONTROL AND AWARENESS
* THESE ARE STANDING IN AS A BROADER CONTROL CONDITION
pc_simulate during01 , ///
        model(POST) ///
        post(90) ///
         mde(-.024) ///
        i(district) ///
        t(edate) ///
        onesided ///
        nsim(1000) ///
        outfile(asc_037no_collapse_pc_sim_actual_no_absorb) replace
* absorb
pc_simulate during01 , ///
        model(POST) ///
        post(90) ///
         mde(-.024) ///
        i(district) ///
        t(edate) ///
        onesided ///
        nsim(1000) ///
        /// collapse ///
        absorb( edate) ///
        outfile(asc_037no_collapse_pc_sim_actual_yes_absorb) replace
```

Estimating power for AR-1 – not absorbing fixed effects for time

```
* NO VISIT FROM POPE
* USE DOUBLED DATA SET
pc_simulate during01 , ///
        model(POST) ///
        post(180) ///
         mde(-.024) ///
        i(district) ///
        t(edate) ///
```

```
        onesided ///
        nsim(1000) ///
        /// collapse ///
        n(10) ///
        outfile(asc_037no_collapse_pc_SIM_AR_1_NO_POPE_180_NO_ABSORB) replace
```

Estimating power for AR-2 – not absorbing fixed effects for time

```
* AR - 2 - ALL EGGS IN ONE BASKET
* NEED TO PULL IN 20 DISTRICT FILE
* treatment ratio = .75
clear all
use property_merged_exposed_byday_d20_20190815_xtset_fixed.dta
* set it so 3/4 = treatment
pc_simulate during01 , ///
        model(POST) ///
        post(90) ///
         mde(-.024) ///
        i(district) ///
        t(edate) ///
        onesided ///
        nsim(1000) ///
        /// collapse ///
        p(.75) ///
        outfile(asc_037no_collapse_pc_sim_ar_2_all_eggs_no_absorb) replace
```

Estimating power for AR-3 – not absorbing fixed effects for time

```
* AR-3 - Philadelphia as London
* more districts - draw bootstrap samples of 46
pc_simulate during01 , ///
        model(POST) ///
        post(90) ///
         mde(-.024) ///
        i(district) ///
        t(edate) ///
        onesided ///
        nsim(1000) ///
        /// collapse ///
        n(46) /// LONDON SCENARIO
        bootstrap ///
        outfile(asc_037no_collapse_pc_sim_ar_3_london_no_absorb) replace
```

Estimating power for AR-4 – five times expansion - not absorbing fixed effects for time

```
* bring in 10 district file
clear all
use ASC_2019_037_CONTROL_AWARE_ONLY.dta
tabstat during01, by(expcond)
clonevar during01x5=during01
* WANT TO GET FROM .04 TO .04 X 5 = .2
* the above proportion would be the new prevalence rate
* recoding a random fraction of the control condition that previously had no
crime
replace during01x5=1 if(rando<=.17)&(expcond==1)&(during01==0)
* recoding a random fraction of the awareness condition that previously had no
crime
replace during01x5=1 if(rando<=.15)&(expcond==2)&(during01==0)
tabstat during01x5, by(expcond)
/* GETS PRETTY CLOSE
 expcond |      mean
---------+----------
       1 |  .2111111
       2 |  .1911111
---------+----------
   Total |  .2011111
--------------------
/*
```

```
THINKING THROUGH HOW MUCH THE MDE WOULD CHANGE
Started with a 60% reduction with a prevalence rate of .04
Now the prevalence rate is 5 x .04 = .20
40% of this would = .2 x .4 =.08
  Below goes from an mde smaller than this to an mde bigger than this
*/
pc_simulate during01x5 , ///
       model(POST) ///
       post(90) ///
        mde(-.02(-.01)-.16) ///
       i(district) ///
       t(edate) ///
       onesided ///
       nsim(1000) ///
       outfile(asc_037no_collapse_ar_4_5x__no_absorb) replace
```

References

Ariel, B., & Partridge, H. (2017). Preidctable policing: Measuring the crime control benefits of hotspots policing at bus stops. *Journal of Quantitative Criminology, 33*, 809-833.

Ariel, B., Weinborn, C., & Sherman, L. W. (2016). “Soft” policing at hot spots—do police community support officers work? A randomized controlled trial. *Journal of Experimental Criminology, 12*(3), 277-317. doi:10.1007/s11292-016-9260-4

Baldassare, M. (2000). *California in the next milennium*. Berkeley: University of California Press.

Braga, A. A., Turchan, B., Papachristos, A. V., & Hureau, D. M. (2019). Hot spots policing of small geographic areas effects of crime. *Campbell Systematic Reviews, 15*, e1046. doi:https://doi.org/10.1002/cl2.1046

Braga, A. A., Weisburd, D. L., Waring, E. J., Lorraine Green, M., & et al. (1999). Problem-oriented policing in violent crime places: A randomized controlled experiment. *Criminology, 37*(3), 541-580. doi:http://dx.doi.org/10.1111/j.1745-9125.1999.tb00496.x

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Earlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Ellis, P. D. (2010). *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis and the Interpretation of Research Results*. Cambridge, UK: Cambridge University Press.

Groff, E. R., Ratcliffe, J. H., Haberman, C. P., Sorg, E. T., Joyce, N. M., & Taylor, R. B. (2015). Does what police do at hot spots matter? The Philadelphia policing tactics experiment. *Criminology, 53*(1), 23-53. doi:10.1111/1745-9125.12055

Hemming, K., & Marsh, J. (2013). A menu-driven facility for sample-size calculations in cluster randomized controlled trials. *The Stata Journal, 13*(1), 114-135.

Hinkle, J. C., Weisburd, D., Famega, C., & Ready, J. (2013). 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. doi:10.1177/0193841x13519799

Killip, S., Mahfoud, Z., & Pearce, K. (2004). What Is an Intracluster Correlation Coefficient? Crucial Concepts for Primary Care Researchers. *The Annals of Family Medicine, 2*(3), 204-208. doi:10.1370/afm.141

Ratcliffe, J. H., Taylor, R. B., Askey, A. P., Fisher, R., & Koehnlein, J. M. (2019). *The Philadelphia Predictive Policing Experiment: Final Report submitted to the National Institute of Justice (grant 2014-R2-CX-0002). [ONLINE: https://bit.ly/376RuYf ; accessed June 9, 2020]*. Retrieved from Philadelphia:

Ratcliffe, J. H., Taylor, R. B., Askey, A. P., Thomas, K., Grasso, J., Bethel, K. J., . . . Koehnlein, J. (2020). The Philadelphia predictive policing experiment. *Journal of Experimental Criminology*. doi:10.1007/s11292-019-09400-2

Rosenfeld, R., Deckard, M., & Blackburn, E. (2014). The effects of directed patrol and self-initiated enforcement on firearm violence: A Randomized controlled study of hot spot policing. *Criminology, 52*(3), 428-449. doi:10.1111/1745-9125.12043

Santos, R. B., & Santos, R. G. (2016). Offender-focused police intervention in residential burglary and theft from vehicle hot spots: a partially blocked randomized control trial. *Journal of Experimental Criminology, 12*(3), 373-402. doi:10.1007/s11292-016-9268-9

Telep, C. W., Mitchell, R. J., & Weisburd, D. (2014). How m uch time should the police spend at crime hot spots? Answers from a police agency directed randomized field trial in Sacramento, California. *Justice Quarterly, 31*(5), 905-933.