

**How *Powerful* is the Evidence in Criminology?
On Whether We Should Fear a Coming Crisis of Confidence**

J.C. Barnes, Ph.D.
School of Criminal Justice
University of Cincinnati
Cincinnati, OH 45221

Michael F. TenEyck, Ph.D.
Department of Criminology and Criminal Justice
University of Texas at Arlington
Arlington, TX 76019

Travis C. Pratt, Ph.D.
Corrections Institute
University of Cincinnati
Cincinnati, OH 45221

Francis T. Cullen, Ph.D.
School of Criminal Justice
University of Cincinnati
Cincinnati, OH 45221

*The authors would like to thank David P. Farrington, Cory P. Haberman, Jean M. McGloin, David C. Pyrooz, Kyle J. Thomas, Jillian J. Turanovic, Jacob T.N. Young, four anonymous reviewers, and the Editor, Megan C. Kurlychek, for their helpful comments on earlier versions of this paper.

**How *Powerful* is the Evidence in Criminology?
On Whether We Should Fear a Coming Crisis of Confidence**

Abstract

A crisis of confidence has struck the behavioral and social sciences. A key factor driving the crisis is the low levels of statistical power in many studies. Low power is problematic because it leads to increased rates of false-negative results, inflated false-discovery rates, and over-estimates of effect sizes. To determine whether these issues impact criminology, we computed estimates of statistical power by drawing 322 mean effect sizes and 271 average sample sizes from 81 meta-analyses. Results indicated criminological studies, on average, have a moderate level of power (mean = 0.605), but there is variability. This variability is observed across general studies as well as those designed to test interventions. Studies using macro-level data tend to have lower power than studies using individual-level data. To avoid a crisis of confidence, criminologists must not ignore statistical power and should be skeptical of large effects found in studies with small samples.

Keywords: crisis of confidence; *P* values; null hypothesis significance tests; statistical power

A crisis of confidence is gripping the behavioral and social sciences, causing scholars to revisit debates that were long thought settled and to view scientific “breakthroughs” with healthy skepticism (Baker, 2016; Earp & Trafimow, 2015; Gelman & Loken, 2014b; National Academy of Science, Engineering, & Medicine, 2016). Scientists—and the public—have been bombarded by findings that are often contradictory to prevailing wisdom, suggesting confidence in any given result is weak. It is reasonable to be confused, for example, when science suggests coffee is good for your health and that it also might kill you (Higdon & Frei, 2006; Woodward & Tunstall-Pedoe, 1999); that you should eat eggs, but maybe not (Lajous et al., 2015; Nakamura et al., 2006); and that spanking your kids will do long-term damage (Lansford et al., 2014) or it might do nothing at all (Ferguson, 2013).

Although much of the current crisis seems to be playing out in other fields like psychology (see Schmidt & Oh, 2016), there are signs that these same concerns impact criminological¹ research. Indeed, there are recent examples of uncertainty over research evidence in our discipline. Take, for example, the correctional intervention known as Project HOPE. Nearly a decade ago, a single evaluation (one with a relatively low level of statistical power) of a correctional program in Hawaii found that threatening offenders on probation with a short but immediate stint in jail for even minor technical violations substantially reduced recidivism (Hawken & Kleiman, 2009). On the coattails of this single small study, HOPE-like programs spread rapidly to over a hundred locations in dozens of states, and have even gone international (Bartels, 2017). But then replications of programs based on the HOPE model started to trickle in—studies with the kind of statistical power the original evaluation lacked (see, e.g., Hamilton et al., 2016)—and the results were unequivocal: HOPE programs had no appreciable effect on

¹ We recognize the conceptual distinctions between the terms *criminology* and *criminal justice*. For simplicity, we will use *criminology* to refer to both in this article.

recidivism (Lattimore et al., 2016; O'Connell, Brent, & Visher, 2016). That thousands of offenders, victims, and community members have been subjected to the consequences of what now appears to be a failed program underscores the importance of exercising caution when studies reveal startling results (Cullen, Pratt, & Turanovic, 2016).

For this and other reasons spelled out below, it is our position that criminologists must take the crisis of confidence seriously and that we should consider whether it affects our own evidence base. This raises a fundamental question: how would we know if criminology were headed for a crisis of confidence? We see at least two ways to forecast that outcome. One is to survey scientists and ask them whether they trust the evidence base (see, generally, American Academy of Arts & Science, 2018). No such information exists for criminologists, so we must rely on the second approach, which is to look for indicators that trouble is brewing. One of the most obvious indicators is that studies begin to fail tests for direct replication (see Clemens [2017], Duvendack et al. [2017], and Pridemore et al. [2018], for discussions of the definitions of replications). When this happens, scientists question which findings are true-positives and which are false-positives. This all came to the forefront recently in psychology when a large-scale effort (Open Science Collaboration, 2015) suggested only about one-third of psychology experiments could be replicated (Nosek et al., 2012; Pashler & Wagenmakers, 2012; but see Gilbert et al., 2016; Johnson et al., 2017; Maxwell, Lau, & Howard, 2015; Stroebe & Strack, 2014). In the aftermath of that study, a survey of scientists revealed that 90% of researchers agree there is a reproducibility crisis in science (Baker, 2016).

Criminology has not experienced anything like the reproducibility crisis that is currently afflicting psychology, so there has been comparatively less discussion of the possibility of a crisis of confidence in our discipline. This may incline some to conclude that criminology does

not suffer the same fate as psychology. But we believe this conclusion is premature if for no other reason than criminologists do not place a high priority on replication (McNeely & Warner, 2015) and, therefore, we have little idea what proportion of our evidence base would replicate if an attempt were made. As evidence, Pridemore and colleagues (2018) reported that less than 1% of nearly 40,000 criminology articles published up to 2014 were replication efforts.

Criminologists must, therefore, look for other indicators to determine whether we are headed for a crisis of confidence. So what are those other indicators? The recent survey conducted by *Nature* (Baker, 2016) identified a short list of problems that scientists believe contributed to the current reproducibility crisis in other disciplines. A large majority of the survey participants indicated that selective reporting of statistically significant results—a practice the public finds morally objectionable (Pickett & Roche, 2018)—and pressures to publish novel findings as a result of the ever-changing structure of science contributed to the problem (see also, Hagan, 1973; King, 1995; Merton, 1957; Simmons et al., 2011; Vostal, 2016). Related concerns stem from publication bias (i.e., reviewers and editors prefer statistically significant results), *P*-hacking (i.e., researchers run lots of tests but only report the ones that result in statistically significant *P* values), and researcher degrees of freedom (i.e., researchers have wide latitude when deciding how to conduct a study and can, therefore, end up publishing a result that is unlikely to hold up) (Bailystok et al., 2015; Carter & McCullough, 2014; Ferguson & Heene, 2012; Franco et al., 2014; Gelman & Loken, 2014b; Schmidt & Oh, 2016).²

But there is another indicator that can reliably forecast a crisis of confidence in an area of research: the level of statistical power that prevails in that area (Button et al., 2013; Ioannidis,

² These concerns are, practically speaking, impossible to eliminate and one might argue that it is undesirable to eliminate them altogether. Instead, finding small ways to incentivize transparent and accountable research practices is likely the best way forward. As an anonymous reviewer noted, the badge systems used at top psychology journals (e.g., *Psychological Science*) appear to have had the intended effects.

2005). Statistical power is an important indicator because it is the fulcrum upon which applied statistical analysis is balanced. Statistical power provides an estimate of the probability that a researcher will correctly reject the null hypothesis when the null is in fact false.³ Moreover, statistical power plays a central role in other computations such as the false-discovery rate, which is the probability that a statistically significant effect is not real (meaning the “true” effect is zero) (see Colquhoun, 2014; Ioannidis, 2005). When statistical power is low, study findings become less trustworthy, leading to a loss of confidence. When an entire discipline is built on *underpowered* research, that entire body of evidence becomes suspect. For context, consider that more than 80% of the researchers surveyed by *Nature* indicated that low statistical power or poor analysis contributed to the reproducibility crisis in other disciplines (Baker, 2016).

Statistical power played a central role in the arguments developed by Ioannidis (2005) in his now classic article, “Why Most Published Research Findings are False.” Ioannidis showed that the positive predictive value for a study—which is the complement to the false-discovery rate, meaning it is the probability that a statistically significant result reflects a true effect—is directly tied to the level of statistical power in that study. All else being equal, as the statistical power for a study goes up, so too does its probability of reaching an accurate conclusion (i.e., the positive predictive value goes up). The inverse is also true: as statistical power goes up, the false-discovery rate goes down. Building on these observations, Ioannidis (2005) was able to show that low levels of statistical power: a) lowers the probability of discovering true effects; b) lowers the probability that a statistically significant effect reflects a true effect; and c) makes it more likely that statistically significant effect size estimates will be inflated. This latter point

³ This statement presupposes that the null hypothesis is either true or false. The validity of this perspective has been drawn into question by leading statisticians (see, for example, Gelman & Loken, 2014a). We will briefly return to this point in the Discussion section, but interested readers are directed to introductory Bayesian texts for more thorough considerations (e.g., Gill, 2014; Lee, 2012).

came to be known as the “winner’s curse” because, “...the ‘lucky’ scientist who makes the discovery in a small study is cursed by finding an inflated effect” (Button et al., 2013:367).

The winner’s curse recently garnered attention in criminological circles after Gelman and colleagues (2018) suggested it provides a plausible explanation for the counterintuitive finding that effect sizes tend to decrease as sample sizes increase in experimental criminology. Prior to Gelman et al.’s study, this relationship was widely known as Weisburd’s paradox—which suggested that effect sizes drop as studies get bigger because researchers are unable to maintain quality control (see also Nelson et al. [2015]). This is all to say that statistical power is an important indicator for criminologists to follow given its ability to explain counterintuitive findings and its ability to forecast false-discovery rates (Button et al., 2013).

It might come as a surprise, then, that criminologists have not studied statistical power more thoroughly. Indeed, as we will show below, over the past 25 years, there have only been four attempts to estimate statistical power in criminological research. As such, we will assess and summarize the level of statistical power that prevails among studies in criminology. This issue matters generally in terms of the quality of basic scientific research that the discipline produces, but it also matters in a very practical way. Criminology is an applied field, with its published knowledge used to inform policy and practice. In an era of increasing calls for the field to be evidence-based in policy matters, criminology has the potential to be influential in unprecedented ways (Clear, 2010; Petrosino et al., 2001; Sherman et al., 2002). To do so effectively and ethically, that research evidence must be reliable and trustworthy, which is to say it should be gleaned from studies that are likely to identify true-positive results. Do we have reason to be confident that the knowledge we share with policymakers and practitioners is true?

Or, similar to other disciplines, is criminology going to confront a crisis of confidence? An assessment of statistical power will indicate where things stand.

Accordingly, in the present paper we have three objectives. First, we consider what statistical power is and how we should be thinking about it in the context of criminological research. Second, inspired by work in economics, psychology, and neuroscience (Bakker et al., 2012; Button et al., 2013; Ioannidis et al., 2015), the present study will report on one of the first discipline-wide power analyses (see Brown, 1989; Weisburd, Lum, & Yang, 2003). Third, we will explore the implications of our findings for evidence-based policy and for criminological research more broadly.

“Ingredients” for Statistical Power

Statistical power is best understood in the context of the null hypothesis significance test (NHST). NHSTs estimate the compatibility between the observed data and the null hypothesis (H_0), which is typically set to zero. The compatibility between the observed data and H_0 is summarized with a P value, where smaller P values (i.e., those closer to 0.00) indicate the data and H_0 are less compatible (Wasserstein & Lazar, 2016). To understand how this works, let us consider the sampling distribution under H_0 . A sampling distribution represents the universe of estimates that would be observed with repeated sampling from the population of focus. The mean of a sampling distribution will equal the population parameter of interest, but note that there will be variation. Indeed, there will be a distribution, meaning some estimates will be larger than the population parameter and some will be smaller.

Any given NHST will result in one of four outcomes, each of which is shown in Table 1. The top row of Table 1 reveals the two possible outcomes of an NHST when H_0 is actually true. Notice that in this scenario there is only one distribution and it is centered on the H_0 value. When

H_0 is in fact true, the researcher will commit a Type I error by rejecting H_0 . This will occur at the rate of α and it is represented as the shaded region in the top-left panel of Table 1. We demarcate the α -level with the vertical drop lines, so the Type I error region is the region that lies at or beyond the α threshold. Alternatively, the researcher can make a correct decision by failing to reject H_0 , which will occur at the rate of $1 - \alpha$. This is represented as the shaded region in the top-right panel of Table 1.

Insert Table 1 about Here

When H_0 is in fact false, two distributions will prevail and there are two possible outcomes of an NHST (bottom row of Table 1). As before, there will be a distribution under H_0 (denoted with the bold line). The second distribution—denoted with the grey line—is sometimes referred to as the non-central distribution and it reveals the sampling distribution that would prevail if the population parameter were not equal to H_0 . In these situations, the researcher will reject H_0 , which is the correct decision, at the rate defined by the level of statistical power ($1 - \beta$). This rate can be calculated as the proportion of the non-central distribution that lies beyond the α -level threshold, which is set in the distribution under H_0 . We represent this as the shaded region in the bottom-left panel of Table 1. The other scenario is that the researcher can make a Type II error by failing to reject H_0 . This will occur at the rate of β , which is represented as the shaded region in the bottom right panel of Table 1.

So what are the “ingredients” for statistical power? In other words, what makes the shaded region in the bottom-left panel of Table 1 larger or smaller? There are, generally speaking, three factors that affect the statistical power of any study (Britt & Weisburd, 2010;

Cohen, 1988)⁴: 1) the critical region, which is demarcated by the α -level specified by the researcher; 2) the effect size of the relationship in question; and 3) the sample size (n) of the study. The first ingredient, the α -level set by the researcher, is a threshold that differentiates a statistically significant effect from one that is not statistically significant (Fisher, 1925[1973]). This value is set by the researcher *a priori* and convention is $\alpha \leq 0.05$. Note that if a researcher chooses a more stringent α -level (e.g., $\alpha = 0.01$), the statistical power of that study will be reduced compared to a less stringent α -level. This can be visualized by looking at the bottom-left panel of Table 1 and imagining the vertical bars are shifted more toward the tails of the distribution under H_0 . Doing so would reduce the shaded region in the non-central distribution.

The second ingredient is the effect size. Larger effect sizes, all else being equal⁵, push the non-central distribution further from the distribution under H_0 . Smaller effect sizes allow these two distributions to overlap. The degree to which the two distributions overlap directly affects the statistical power of the study because statistical power is calculated as the proportion of the non-central distribution that lies beyond the α -level threshold. This can be visualized in the bottom-right panel of Table 1. A larger effect size would move the non-central distribution further to the right, resulting in a greater proportion of that distribution lying beyond the α -level threshold (i.e., the shaded region would increase).

The last ingredient is the n for the study. All else being equal, studies with a larger n will produce sampling distributions with less variation, meaning any estimate gleaned from the

⁴ Sherman (2007) raises a fourth factor—heterogeneity in effect size—that should also be considered. Because we seek to offer a broad discussion of statistical power here, we consider heterogeneity of effect sizes to be part of the more general effect size component but we do recognize its unique contribution to average estimates of statistical power. Interested readers are encouraged to see Sherman’s comments.

⁵ We thank an anonymous reviewer for pointing out that, in applied research contexts, all else is typically not equal. Funding constraints mean researchers must balance various issues, where maximizing sample size is just one concern of many. It is important to note, then, that we are not implying criminologists have intentionally overlooked statistical power. It is more likely the case that statistical power is but one concern among a landscape of issues that confront scientists in any project. Our goal here is to provide context for why statistical power is and should remain a key concern.

sampling distribution will have greater precision. This is reflected by standard errors, which are estimates of the standard deviation of the sampling distribution. Standard errors decrease as n increases. As n grows, the non-central distribution is narrower with a much more obvious peak around the population parameter. The reverse is also true; as n drops, the sampling distribution has greater variation.

Considering these three ingredients reveals a few ways a researcher might affect the statistical power of any given analysis. First, s/he could adjust the α -level. Increases in the α -level will increase statistical power (e.g., move from $P < 0.05$ to $P < 0.10$). But this is not a recommended strategy—and we do not mean to imply that researchers actually use this strategy—because it increases the probability of a false-positive finding. Second, all else being equal, an increase in the effect size would increase statistical power. But it goes without saying that the researcher cannot change the effect size of a relationship, so this is not a viable strategy. This leaves sample size as the primary target. If a researcher wants to increase statistical power in any given study, the most appropriate strategy for doing so is to increase the sample size for the analysis.⁶ For this reason, in our analysis we will highlight the sample sizes that are typical in criminological research.

Prior Estimates of Statistical Power in Criminology

⁶ But, as an anonymous reviewer noted, it is not always easy for researchers to increase sample sizes given a) budgetary constraints on primary data collection efforts and b) the reality that secondary data analysis affords no flexibility on sample size. We agree with these concerns but would also point out that calculating statistical power prior to conducting a study (whether it be primary or secondary data) allows one to gain insight into whether prevailing sample sizes are appropriate for testing the question at hand given an expected effect size. Doing so also affords the researcher the opportunity to estimate the probability of other types of inferential errors prior to carrying out an analysis (Gelman & Carlin, 2014). This, however, should not give the misleading impression that statistical significance is the goal of any study and that one simply needs a “big enough” sample size to achieve $P < 0.05$. We will return to these points in the Discussion section, particularly when we recommend that criminologists consider adopting alternative approaches.

Our discussion has led to the central question that motivates this study: how (statistically) *powerful* is the evidence in criminology? Discussions of statistical power are rare in the criminological literature, but attempts to estimate and summarize it are not completely absent. To date, four studies have sought to provide an overall (or an average) estimate of the prevailing level of statistical power for certain areas of criminological research (Brown, 1989; Nelson et al., 2015; Weisburd et al., 1993; Weisburd et al., 2003).

The first of these was published more than a quarter-century ago (Brown, 1989). Brown (1989) analyzed 53 individual articles published in eight leading criminology journals.⁷ Brown's analysis revealed that roughly half of the published studies analyzed samples that were smaller than 250 and fully 83% of the studies analyzed samples smaller than 1,000. Based on the different effect size thresholds defined by Cohen's *d* (1988), Brown reported that very few of the published studies had enough power—typically, statistical power of 0.80 is the threshold—to detect small effect sizes. And just over half of the studies had enough statistical power to reliably detect moderate effect sizes. The focus on small and moderate effect sizes is quite important because, as Weisburd and Piquero's (2008) analysis revealed, the typical criminology study is expected to produce small-to-moderate effect sizes.

The second study was conducted by Weisburd and colleagues (1993) who sought to determine whether design features like the sample size impacted the findings gleaned from criminology experiments. Their analysis of 74 intervention studies suggested that only 15% had enough statistical power (i.e., power > 0.80) to reliably detect a small effect size (again, using Cohen's *d*). They did conclude, though, that most studies had plenty of statistical power to detect moderate and large effect sizes.

⁷ The journals were *Journal of Criminal Law & Criminology*, *Criminology*, *Journal of Research in Crime and Delinquency*, *Crime & Delinquency*, *Journal of Criminal Justice*, *Journal of Police Science & Administration*, *Criminal Justice Review*, and *Criminal Justice and Behavior*.

The third study, Weisburd et al.'s (2003) analysis of 58 effect sizes from the "Maryland Report" by Sherman et al. (1997), found that the vast majority (70%) of the studies reported an average effect size that was substantively small (Cohen's d between 0.00 and 0.20). Based on the assumption of a small effect size, Weisburd and colleagues (2003) calculated that 83% of the studies analyzed were underpowered to detect the effects that they produced.

And finally, Nelson and colleagues (2015) drew effect size estimates from 66 experiments in the Campbell Collaboration's Crime and Justice Coordinating Group. These authors reported that experiments in the Campbell Collaboration tended to be underpowered, with a total average estimate of statistical power coming in at 0.32, well below the suggested 0.80 threshold.

The four analyses that are currently available suggest statistical power in criminology is quite low (Brown, 1989; Nelson et al., 2015; Weisburd et al., 1993; Weisburd et al., 2003), but there is reason to be cautious about generalizing to the typical criminological study. Two concerns give us pause. First, several of these prior studies are now somewhat dated (e.g., Brown, 1989; Weisburd et al., 1993; Weisburd et al., 2003). Given the recent popularity of large, nationally representative datasets in criminology (e.g., the Add Health data), it may be the case that statistical power has increased over the past two-to-three decades.

Second, three of the prior power analyses were focused on specific research topics or designs used by criminologists. Weisburd and colleagues (1993) focused exclusively on randomized experiments, Weisburd et al. (2003) only analyzed studies in crime prevention, and Nelson et al. (2015) restricted their analysis to randomized controlled trials drawn from the Campbell Collaboration. Each of these studies might be expected to produce lower levels of statistical power (relative to other areas of inquiry) due to the smaller sample sizes that are

typical of studies in those areas. In other words, it may not be appropriate to generalize the results from any of the available power analyses to the rest of the field.

Thus, we sought to expand the existing literature in two ways. First, we provide an up-to-date assessment of statistical power. Second, we estimate statistical power across a broad range of research topics and research designs in criminological studies. This sort of “metascience”—studying science itself (Munafò et al., 2017)—is not frequently conducted in criminology (for notable examples, see Bushway et al., 2006; Gelman et al., 2018; McNeely & Warner, 2015; Pridemore et al., 2018). Thus, while studying statistical power is important in its own right, we believe there is value in conducting metascience reviews like this because they can act as a check on the direction our field is going.

Before we move on, however, it is important to explicitly acknowledge a key point. That is, statistical power might be considered a second-order concern, conditional on the assumption that an effect size has been properly identified. Put a different way, a reasonable argument can be made that concerns over statistical power should be secondary to concerns over confounding and other sources of bias. We certainly agree that proper identification of effect sizes is a major issue in criminological research because most hypotheses and research questions in criminology are not amenable to experimental designs (Sampson, 2010), raising questions about whether any given effect size estimate has been properly identified. But to note the concern of effect size identification does not diminish the importance of statistical power for an applied field like criminology. If anything, we believe it heightens its relevance because properly identified effect size estimates can only be trusted proportional to the statistical power of the test used to produce them. As Gelman and Carlin (2014) demonstrated, variation in the sampling distribution of an effect size is a function of statistical power. If a study has low statistical power, then the effect

size estimate—even if it is properly identified—will vary and could, in some cases, come out in the wrong direction.

Methods

Adopting the strategies that have been applied in psychology (Bakker et al., 2012; Nuijten et al., 2018) and neuroscience (Button et al., 2013), we gathered average effect size and average sample size (n) estimates from criminology meta-analyses. There are several benefits to relying on evidence from meta-analyses.⁸ First, the effect sizes estimated in meta-analyses are often more stable than those estimated from any single study (Lipsey & Wilson, 2001). Second, meta-analyses can provide estimates of the average n that prevails among criminology studies. Third, meta-analyses are topic-specific, meaning we can summarize vast bodies of evidence by drawing on the information reported in a manageable number of sources. Thus, in the sections that follow, we will outline the adopted procedures for identifying meta-analyses in criminology and how we calculated the average effect size and the average n .⁹ Knowing these bits of information allowed us to estimate the average level of statistical power in criminological research.¹⁰

⁸ But, as an anonymous reviewer pointed out, published meta-analyses do not perfectly represent the complexity of research in criminology. For example, focusing on meta-analyses means we necessarily are restricted to quantitative studies. It is not clear whether findings from our study have anything to say about qualitative research. This point, to our knowledge, has not been raised in other fields that are currently experiencing the crisis of confidence. We therefore encourage criminologists to engage in this discussion given the rich pieces of information that have been gleaned from qualitative research in our field.

⁹ Ideally, we would have collected additional information such as the sample size per cell, manipulations that were carried out, and statistical tests that were estimated. But most of the meta-analyses did not include this information. We encourage scholars to follow-up on our efforts with more nuanced foci like those mentioned here. To do so will most likely necessitate a review of primary studies instead of meta-analyses.

¹⁰ An anonymous reviewer noted that meta-analysis "...provides an effective solution to the problems of sampling error and low power and precision in individual studies" (Schmidt & Oh, 2016:34). This is an important point. Because we rely on the average effect size and the average sample size reported in each meta-analysis, we are able to provide an estimate of the average level of statistical power for each meta-analysis included in our sample. But, to the degree that reliance on meta-analysis has affected our estimates of statistical power, given the comments by Schmidt and Oh, one might be inclined to conclude that our estimates are inflated (i.e., that statistical power is actually lower than we report). Also, by extension, our study does not consider research that has not been meta-

Literature Search, Inclusion Criteria, and Coding

In order to include as many meta-analyses as possible, we searched Google Scholar and the Web of Science databases for any studies published between 1990 and 2015 that included some combination of key-word terms like: “meta-analysis,” “crime,” “delinquency,” or “recidivism”.¹¹ To be considered for inclusion in the analysis, all studies were required to be written in English, peer-reviewed, and published in a criminology journal. Additionally, three other inclusion criteria were established: 1) the study must provide an effect size estimate (e.g., Cohen’s d , r); 2) the meta-analysis must report on the association between two variables (i.e., descriptive meta-analyses were not included); and 3) the outcome of the study must be a form of criminal behavior if using individual-level data or a form of criminal activity/crime rates if macro-level data. Although not an inclusion criterion *per se*, all of the meta-analyses were searched for an indicator of sample size (n) for the original studies.¹²

After applying the inclusion criteria, 81 meta-analyses were deemed eligible and were included in the analysis. This number of meta-analyses is consistent with Farrington et al. (2016) who recently found 43 systematic reviews published between 2000 and 2016 that were focused on the risk factors for violence, offending, and delinquency. Recall our time window spanned 1990 to 2015 so it is reasonable that we found nearly twice as many as Farrington and colleagues. Nonetheless, we conduct supplemental analyses to gauge how robust our estimates are to overlooked studies (see the Findings section below).

analyzed. Based on Schmidt and Oh’s (2016) arguments, we have reason to believe statistical power will be higher in our study than if one were to assess a sample of findings in criminology that have not been meta-analyzed.

¹¹ Specifically, the advanced search codes for the Web of Science search were: “TI=(meta*) AND TI=(crim* OR delinq* OR antisocial OR anti-social OR recidivism) AND WC=(Criminology & Penology) Timespan: 1990-2015.” The codes for the Google Scholar search were: “allintitle: meta-analysis crime, OR delinquency, OR antisocial, OR anti-social, OR recidivism” and the date range was restricted to studies published in 1990-2015.

¹² An example of a study that did not meet the inclusion criteria and was, therefore, omitted from the analysis is the meta-analysis on the relationship between self-control and victimization by Pratt and colleagues (2014). This study met the first two criteria, but it did not meet the third criterion.

These 81 meta-analyses covered more than 6,000 individual primary studies. All studies included in the analysis are indexed in the reference list with an asterisk (*). The information coded from each meta-analysis is available in downloadable format on the first author’s GitHub page (<https://github.com/jcbarnescrim/power>). Additionally, all of the *Stata* and *R* code used to conduct the analysis is posted to that same GitHub page. Thus, readers are free to reproduce the present findings or expand on the analysis as they see fit (see Nosek et al., 2015).

Summarizing a body of literature is often challenging due to different reporting conventions, the variety of statistical techniques available, and the various operationalizations of key constructs. These were important issues to be dealt with in the current analysis. As a result, we adopted the following guidelines for study coding. Whenever available, we coded the weighted—rather than the unweighted—effect sizes. In a few cases, effect sizes from both fixed and random effects models were reported. Given the option, we always coded the random effects estimates because random effects models allow for variation between studies that is not strictly due to random error (Lipsey & Wilson, 2001). Our primary focus was on “overall” or “average” effect sizes, so we coded those in lieu of effect sizes from moderator analyses. This was done to keep the analysis tractable and so that the between study variation in effect sizes would not be a function of moderator effects being explored in study A but not in study B.¹³

Findings

¹³ For example, the meta-analysis by Welsh and Farrington (2009) analyzed the effect of CCTV on local crime rates. Moderator analyses explored whether the effect of CCTV was specific to a particular crime type (violent versus vehicle crime). For the purposes of this study, we recorded information on the overall crime analysis. Another example is the meta-analytic results presented by Gobeil et al. (2016). This study presented the results of interventions on recidivism for female offenders. The authors provide an “overall” effect, which was included in our analysis. Gobeil and colleagues also performed several additional analyses to address questions about whether the type of intervention affected the outcome. These latter estimates were not coded for the present analysis. Again, our point here was to gather broad estimates of effect sizes and average sample sizes across the discipline—not to explore variability of effects due to moderator analyses.

Effect Sizes in Criminology

Effect sizes were collected from each of the 81 meta-analyses and were coded according to the metric used in the original study. Seven different effect size metrics were observed: Cohen's d , r , $z(r)$, phi , Hedge's g , odds ratio, and risk ratio.¹⁴ We relied on well-documented conversion formulae to transform the above effect sizes into the Pearson product-moment correlation r (Borenstein et al., 2009; Lipsey & Wilson, 2001; Rosenthal & DiMatteo, 2001). Thus, all effect sizes discussed from this point forward are on the standard r scale. Because we are not testing the directionality of any given relationship, but instead are simply interested in the magnitude of effect sizes in criminology, we carried out an absolute value transformation of r (i.e., $|r|$). Positive and/or negative values are thus treated equivalently in our analysis.

As shown in the first row of Table 2, there were 322 effect sizes reported across the 81 meta-analyses. The distribution of effect sizes is presented in Figure 1. The mean of the distribution is plotted as an open circle, along with the 25th percentile value (triangle on the left), the median (diamond), and the 75th percentile value (triangle on the right). The mean was $|\bar{r}| = 0.148$, revealing the average effect size in criminology is small-to-medium (Cohen, 1988). This finding is consistent with prior work (Sherman et al., 1997; Weisburd et al., 1993).

Insert Table 2 about Here

Insert Figure 1 about Here

We performed four supplemental analyses to examine the robustness of our estimate of the average effect size. First, we estimated the degree to which overlooked meta-analyses might have biased our findings. In essence, we tested the sensitivity of our estimate to missing data. Let

¹⁴ Risk ratios were treated as odds ratios during conversion, which is unlikely to have biased the results due to the fact that our analysis generally focuses on rare outcomes. Risk ratios are approximately equal to odds ratios when the outcome is rare, but they will diverge such that the odds ratio will grow larger than the risk ratio as the prevalence of the outcome increases. Only 21 of the total 322 effects sizes were reported as risk ratios and our effect size estimates do not substantively change when we omit the risk ratios from the analysis.

us imagine our literature search overlooked 180 effect size estimates that all had an effect size of 0.20 (i.e., a large effect size relatively speaking). If we were to add these estimates to our sample, the resulting mean value would be $|\bar{r}|_{\text{revised}} = 0.166$. In other words, even if we overlooked a very large number of relatively large effect sizes, the conclusions we reach below are unlikely to be affected because the resulting bias will only change $|\bar{r}|$ by roughly 0.018 units.

The second supplemental analysis estimated whether the inclusion of more than one effect size estimate per study had any impact on our estimate of $|\bar{r}|$. When we randomly drew an effect size estimate for each meta-analysis, the resulting value for $|\bar{r}|_{\text{unclustered}}$ was 0.151, a value that is only 0.003 units larger than the $|\bar{r}|$ reported above. This served to indicate that our findings are robust to any concerns that might arise due to the clustering of effect sizes within studies.

For the third supplemental test, we considered whether effect sizes among intervention studies (41 total) were different than effect sizes from non-intervention studies (281 total). Estimates were identified as coming from an intervention study if it (the effect size estimate) reflected the impact of a program on criminal behavior or crime rates. Meta-analyses with at least one intervention estimate are demarcated with a superscript “i” in the reference list. Our analysis revealed no meaningful substantive difference between intervention and non-intervention studies. The mean effect size for intervention studies was $|r_{\text{int}}| = 0.120$ and it was $|r_{\text{non-int}}| = 0.152$ for non-intervention studies. A combined histogram is provided in Panel A of Appendix A.

Our fourth supplemental test assessed whether effect sizes among individual-level studies (253 total) were different than effect sizes from macro-level studies (69 total). There did not

appear to be a meaningful difference in the effect size estimates between individual-level studies ($|r_{ind-level}| = 0.147$) and macro-level studies ($|r_{macro-level}| = 0.150$). A combined histogram is provided in Panel A of Appendix B.

Sample Sizes (n) in Criminology

The average n reported in criminology meta-analyses is provided in the second block of rows in Table 2. We report on n observed among the studies that were used to compute the corresponding effect sizes. There were a number of studies that did not provide any details we could use to draw this information. Specifically, n was only available for 271 effect sizes (84.2%). We performed a t -test to assess whether the effect sizes among studies with available n differed from those with missing n . This test showed only a small substantive difference and the confidence interval included 0.00 (mean difference in $|r| = 0.016$, 95% confidence interval = [-0.015, 0.046], $t = 0.989$, $P = 0.323$ [two-tailed test]), so we assume there is no systematic difference among effect sizes gleaned from studies with and without n .

As shown in Table 2, the mean was $\bar{n} = 2,929.821$, but the median was $n = 327.429$, indicating considerable right skew. Approximately 70% of all effect sizes were drawn from studies with n less than 1,000 and more than 90% of all effect sizes were drawn from studies with n less than 5,000. Thus, we also present summary statistics for all studies where $n \leq 5,000$ (see the third block of rows in Table 2). It is worth noting the similarity of these results to those reported by Brown (1989), who found the vast majority of studies used samples smaller than $n = 1,000$.

The histogram plotted in Figure 2 shows the distribution of n for all corresponding 252 effect sizes where $n \leq 5,000$. Even among this (slightly) restricted sample, the right skew is obvious. The mean of the distribution, which is marked with an open circle, appears at $\bar{n} =$

710.122. The median, which is marked by the diamond, appears much further to the left, at $n = 291.125$.

Insert Figure 2 about Here

As with the effect size analysis, we performed two supplemental tests: 1) to determine whether intervention studies substantively differed from non-intervention studies and 2) to determine whether individual-level studies differed from macro-level studies. First, intervention studies were found to have a more restricted range of sample sizes compared to non-intervention studies. More specifically, the mean was $\bar{n} = 627.084$ among all intervention studies (the 95% range was 27.333 to 3,492.763) and $\bar{n} = 3,184.632$ among all non-intervention studies (the 95% range was 44 to 24,353.630) (results for studies where $n \leq 5,000$ are provided in the third block of rows in Table 2). In short, non-intervention studies have a wider range of n and a larger mean n . This could impact statistical power in a predictable way: we might expect intervention studies to have lower statistical power compared to non-intervention studies. See Panel B of Appendix A for a combined histogram of n for intervention and non-intervention studies where $n \leq 5,000$.

As for individual-level and macro-level studies, individual-level studies presented with a mean $\bar{n} = 3,292.281$ (the 95% range was 62 to 24,911.130) and macro-level studies presented with a mean $\bar{n} = 485.804$ (the 95% range was 44 to 6,476.400). Similar to our prediction about intervention studies, these observations suggest macro-level studies will have lower statistical power compared to individual-level studies—at least on average. A combined histogram is presented in Panel B of Appendix B.

Statistical Power ($1 - \beta$) in Criminology

Statistical power estimates were calculated for each effect size/sample size combination that was observed in the meta-analyses. All power calculations were conducted using the `power`

package in *Stata* 14.1. As was noted above, all meta-analyses provided an effect size, but not all of them provided the information necessary to calculate n . Thus, we were only able to estimate statistical power using 270 effect sizes. This number is one value smaller than the total number of sample size estimates (271) because one study reported a mean effect size of 0.00. Statistical power calculations are unnecessary for this case because the effect size was equal to the null hypothesis value.

Statistical power estimates are presented in the last block of rows in Table 2 and as a histogram in Figure 3. The figure has a distinct U shape, with a cluster of very low-powered studies on the left and a cluster of high-powered studies on the right. In that respect, the distribution looks very similar to statistical power estimates observed in other behavioral science disciplines (for a similar pattern from neuroscience, see Button et al., 2013; Nuijten et al., 2018). The open circle demarcates the average level of statistical power, which was 0.605. The average level of statistical power can be interpreted as the expected probability that any randomly drawn study will reject the null hypothesis when the null is in fact false. Because the average level of statistical power is below the typical 0.80 threshold, we can conclude that the average study is *underpowered* to detect the effect sizes that are observed. This conclusion is consistent with the four statistical power analyses that preceded ours (Brown, 1989; Nelson et al., 2015; Weisburd et al., 1993; Weisburd et al., 2003).

Insert Figure 3 about Here

Note that the median level of statistical power is somewhat higher than the mean. Specifically, the median level of statistical power (denoted by the diamond in Figure 3) was 0.706. Based on this value and using the standard 0.80 power threshold as a benchmark, one can conclude that *more than half of all criminology studies are underpowered to detect the effect*

sizes they observe. This finding is sobering evidence that much of the published literature is reporting results that could be incorrect.

But there was variation in the observed level of statistical power. The 75th percentile value was located at statistical power = 0.992, meaning that at least 25% of all criminology research is very well powered. This finding is mainly driven by the right-skew observed for n ; about 25% of all studies have very large n and, therefore, are well powered to detect even small effect sizes. Still, a large number of studies have troublingly low levels of statistical power. Note that the 25th percentile was located at statistical power = 0.236. This means that nearly one-quarter of all research has statistical power below 0.24.

We performed two supplemental analyses to determine whether statistical power varied between 1) intervention and non-intervention studies and 2) between individual-level and macro-level studies. As was predicted due to the observation that intervention studies tended to have smaller n , we found that the mean level of statistical power among intervention studies was smaller (0.534) than the mean level of statistical power among non-intervention studies (0.613). Yet the distribution of statistical power is substantively similar across the two types of studies. Panel C of Appendix A reveals the distribution of statistical power estimates for intervention studies and Panel D reveals the distribution for non-intervention studies. Both types of studies presented with the same U shaped distribution that was observed among all studies. As can be seen, the problem of low statistical power thus marks both types of studies.

Similarly, our prediction for the difference in statistical power between individual-level and macro-level studies was borne out: the mean level of statistical power for individual-level studies (0.654) was substantively larger than the mean for macro-level studies (0.274).

Distributions can be found in Panel C (individual-level studies) and Panel D (macro-level studies) of Appendix B.

Discussion

How much of our evidence base is trustworthy? Although answering this question is difficult, one can rely on the level of statistical power that prevails in the literature as an indicator because low statistical power portends high false-negative rates and high false-discovery rates (Ioannidis, 2005). Based on the results of our assessment of more than 300 effect sizes and more than 250 sample sizes gleaned from more than 80 meta-analyses—meta-analyses that cut across a wide range of criminological topics and span a quarter-century of research—two key findings emerged. We will now consider those two key findings and we will discuss the implications that stem from them. After that, we will consider potential solutions to the problems that can be caused by low statistical power.

Key Findings

The first key point to take away from this analysis is that studies in criminology have relatively high levels of statistical power compared to other areas of behavioral and social science. It is, however, important to emphasize the word *relatively* here. Criminology has high statistical power (on average) compared to fields like psychology (Bakker et al., 2012), neuroscience (Button et al., 2013), and behavioral genetics (Duncan & Keller, 2011; Sham & Purcell, 2014), but those disciplines have confronted very low levels of statistical power. Thus, we should not consider ourselves free from the problems that come from low statistical power. At the same time, it is important to point out that high statistical power—either in an absolute sense or a relative one—does not mean all is well. Higher statistical power may be an indicator

that researchers perceive more flexibility in research designs, meaning there are greater perceived researcher degrees of freedom. As we discussed in the introduction to this study, flexibility in research design is a major concern and likely a key contributor to the crisis of confidence in psychology and other areas of behavioral research (Simmons et al., 2011).¹⁵ Also, high statistical power might lead researchers to put too much emphasis on substantively small effect sizes, directing attention and resources to efforts that may have little impact. Thus, for these reasons, we encourage criminologists to remain cautious when interpreting results—even if the study boasts a high level of statistical power.

Nevertheless, our second—and related—point is that there is wide variability in statistical power across studies in our field. Indeed, the 95% range of statistical power estimates is between 0.052 and nearly 1.00. Comfortingly, a good portion of studies exhibit high statistical power (about 25% were in the 0.99 to 1.00 range), something that should not be surprising given that scholars in our field often rely on large, publicly available datasets to conduct their research (Woodward et al., 2016; Worrall, 2000).

But there is also a large portion of studies that are (on average) dangerously underpowered. About 25% of all studies have power between 0.01 and 0.24. This means that roughly one-quarter of all studies in criminology have levels of statistical power that make it nearly impossible to identify the effects they are estimating. And, given that low statistical power forces one to find larger effects to reject the null, this result raises the possibility that a sizable portion of our evidence base suffers from the “winner’s curse” (Button et al., 2013). The “curse”, recall, is that the researcher who finds a statistically significant result in an underpowered study is likely to have overstated the effect size (Gelman & Carlin, 2014; Ioannidis, 2005).

¹⁵ We thank an anonymous reviewer for raising this point.

Implications

For decades, criminology was marginalized and dismissed as irrelevant when it came to informing policy decisions about how to control crime (Austin, 2003; Cressey, 1978). In more recent times, however, criminology has earned a seat at the criminal justice policy-making table. This new-found legitimacy has been fueled primarily by the broader acceptance of an evidence-based approach to criminal justice policy (Braga & Apel, 2016; Petrosino et al., 2001; Sherman et al., 2002); an approach that defers to criminologists as the experts who produce the necessary evidence and who should be in a good position to translate that evidence into practice. In short, criminology arguably matters more now than it ever has.

Thus, a sizable body of underpowered research is cause for concern. Because nearly any study with a statistically significant P value is likely to be published somewhere, problems arising from low levels of statistical power could have an impact on evidence-based practices. This raises an interesting question: which studies are policymakers reading? If policymakers are reading the top journals in our field, then it may be that concerns about low statistical power are not as acute (working on the assumption that higher ranked journals tend to published stronger studies [in this context, meaning higher powered]). But the rate at which policymakers read *Justice Quarterly*, for instance, is an unknown and certainly a question worthy of empirical scrutiny. Thus, for the present study, we relied on the assumption that all articles are treated equally by policymakers and should, therefore, be given equal weight in our analysis. This was both practical and necessary because meta-analyses summarize evidence from numerous primary studies, so it would be exceedingly difficult to try and identify an appropriate weighting mechanism that would take these points into account. Nonetheless, this is an important

assumption that should be examined in future work—perhaps by performing power analyses on primary studies that have proven impactful in the policy arena.¹⁶

To the extent that scholars make recommendations based on underpowered research, it is possible that they will advise policymakers with faulty evidence (see, e.g., Maruna, 2015). This may mean policymakers overlook interventions or policy changes that work and instead waste time and resources on those that do not. If policymakers are advised based on underpowered studies, they may invest in strategies that are unlikely to yield returns that would be expected based on the research evidence. Criminologists should prioritize studies and research strategies that maximize statistical power because it will afford policymakers the ability to differentiate between evidence that is actionable and evidence that is not.

Arguably the most risky kind of evidence is a large effect size gleaned from a statistically underpowered study. When we see findings like this, our reaction should be one of “organized skepticism” (Merton, 1973)—where caution is exercised and a call for replication is made before we get too excited about the single study (see also Kulig, Pratt, & Cullen, 2017; Pratt, Turanovic, & Cullen, 2016). But demonstrating such caution is not something humans are either wired for or socialized to do very well (Kahneman, 2011). And as a discipline, criminology has a sketchy track record with demonstrating caution in the wake of a splashy new finding. Recall the case of Project HOPE.

Moving Forward

This all raises a very important question: what are we to do? One obvious recommendation comes to mind: scholars should, all else being equal, prioritize and prefer studies with larger sample sizes. Scholars and the public alike tend to trust evidence that comes

¹⁶ We thank an anonymous reviewer for raising these points.

from larger studies. But we would be remiss if we did not acknowledge recent recommendations (Gelman et al., 2018) that advised against the use of statistical power altogether. The reason for this recommendation was based on the fact that a focus on statistical power implicitly touts the importance of statistical significance as a research goal. We agree that statistical significance should not be viewed as a research goal, but alternative approaches (meaning alternatives to the null hypothesis significance test) require a drastic shift in the way criminological research gets done. Specifically, it would require a shift toward Bayesian statistical analysis. We do not have the space necessary to compare and contrast Bayesian statistics with the standard frequentist paradigm that currently dominates criminological research practices (for some discussion, see Gill [2014] or Lee [2012]). Suffice to say, we second Gelman and colleagues' (2018) recommendation and we encourage scholars in the discipline to take these points seriously. In the meantime, we have three additional recommendations.

First, criminologists should place more emphasis on *substantive* significance and less emphasis on *statistical* significance (see, generally, Maltz, 1994; McShane & Gal, 2015). Criminologists have an abundance of theories that guide us when we seek to answer questions about relationships of interest. Those theories are almost always silent, though, on how large the relationships (i.e., effect sizes) are expected to be. Thus, a systematic consideration of the expected effect sizes for the various areas of criminology would be beneficial to the discipline because it would give scholars a more realistic benchmark to use when considering the relative importance of any given result. It might also allow scholars to perform tests for statistical power prior to any inferential statistical analysis (Cohen, 1988; see, for example, Weisburd et al., 2003). A systematic consideration of effect sizes in criminology could help scholars assess the degree to which certain estimates might have been inflated/exaggerated (Gelman & Carlin, 2014).

Second, criminologists should place a greater emphasis on confidence intervals than they do on *P* values and statistical significance. Rosnow and Rosenthal (1989:1277) noted that, “surely, God loves the .06 nearly as much as the .05”. This quote highlights the reality that the $P < 0.05$ statistical significance criterion is arbitrary. Emphasizing confidence intervals might, therefore, help to redirect our focus back to the magnitude of the effects we are estimating—something that may ultimately prove to be more useful than *P* values. But it is important to note confidence intervals come with their own set of misunderstandings and fallacies (Morey et al., 2015). Thus, while we urge criminologists to think about confidence intervals as a substitute for *P* values, we do so with full understanding that confidence intervals are not the end game. Rather we see them as a stepping-stone that moves us in the right direction—toward a better appreciation of substantive significance and, perhaps eventually, to Bayesian inferential methods.

Third, we call upon authors, journal editors, and manuscript reviewers to place a greater emphasis on conducting and presenting sensitivity analyses in published studies. When we estimate our models and we get the result we think we want, how committed are we to avoiding the trap of confirmation bias (Jussim et al., 2016)? How often do we subject our finding to all of the alternative specifications that might have been thrown at it (see, e.g., Gorman, 2015; Nickerson, 1998)? Are the results robust enough to withstand empirical challenges, or do they disappear upon introducing the slightest methodological tweak? Put simply, what we are suggesting is that all of those engaged in the research enterprise ask themselves one simple but tough question before publishing any result: do I trust this finding?

Conclusion

The primary theme of this study concerns whether criminologists can be confident in the evidence-base we have created. A binary response to such an inquiry would not do justice to the nuances we have uncovered, but a broad-based assessment would suggest criminology is not immune to the same ills that caused the crisis of confidence in other disciplines like psychology. The indicators are already beginning to emerge: Weisburd's paradox (1993; see also Nelson et al., 2015) is likely explainable as a result of the winner's curse and publication bias toward statistically significant findings (Gelman et al., 2018); more than half of the studies we analyzed were underpowered; more than one-fourth of all criminology studies have dangerously low levels of statistical power; and statistical power is quite low, on average, in macro-level research.

It is important to recognize that scientific evidence about the causes of crime is inherently difficult to pin down. Criminologists face the dual task of producing evidence and constructing a narrative that may explain that evidence. It is therefore easy to understand how a crisis of confidence might emerge. If said evidence is drawn into question, then the narratives built on top of it become suspect. This brings to mind a point made by the Nobel Laureate Daniel Kahneman (2011:212): "Confidence is a feeling, which reflects the coherence of the information and the cognitive ease of processing it. It is wise to take admissions of uncertainty seriously, but declarations of high confidence mainly tell you that an individual has constructed a coherent story in his mind, not necessarily that the story is true."

Criminologists have the difficult job of trying to piece together noisy evidence to arrive at a coherent story. We have reason to be optimistic that many studies in criminology show something real about the causes of crime because statistical power is relatively high on average and many studies have a very high level of statistical power. But separating the signals from the

noise is never an easy task. It is, however, a little easier when researchers use well powered research designs.

References

(*appears in analysis; ⁱ contributed at least one intervention effect)

- American Academy of Arts & Sciences. (2018). *Perceptions of science in America*. Cambridge, MA: American Academy of Arts & Sciences
- *ⁱAndrews, D. A., Zinger, I., Hoge, R. D., Bonta, J., Gendreau, P., & Cullen, F. T. (1990). Does correctional treatment work? A clinically relevant and psychologically informed meta-analysis. *Criminology*, 28, 369-404.
- Austin, J. (2003). Why criminology is irrelevant. *Criminology and Public Policy*, 2, 557-564.
- *Baier, C. J., & Wright, B. R. E. (2001). "If you love me, keep my commandments": A meta-analysis of the effect of religion on crime. *Journal of Research in Crime and Delinquency*, 38, 3-21.
- Bailystok, E., Kroll, J. F., Green, D. W., MacWhinney, B., & Craik F. I. M. (2015). Publication bias and the validity of evidence: What's the connection? *Psychological Science*, in press.
- Baker, M. (2016). 1,500 scientists lift the lid on reproducibility: Survey sheds light on the 'crisis' rocking research. *Nature*, 533, 452-54.
- Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, 7, 543-54.
- *ⁱBarnes, T. N., Smith, S. W., & Miller, M. D., (2014). School-based cognitive-behavioral interventions in the treatment of aggression on the United States: A meta-analysis. *Aggression and Violent Behavior*, 19, 311-321.
- Bartels, L. (2017). *Swift, certain and fair: Does Project HOPE provide a therapeutic paradigm for managing offenders?* Basingstoke: Palgrave MacMillan.
- *ⁱBennett, T., Holloway, K., & Farrington, D. (2006). Does neighborhood watch reduce crime? A systematic review and meta-analysis. *Journal of Experimental Criminology*, 2, 437-458.
- *Bennett, T., Holloway, K., & Farrington, D. (2008). The statistical association between drug misuse and crime: A meta-analysis. *Aggression and Violent Behavior*, 13, 107-118.
- *Blais, J., Solodukhin, E., & Forth, A. E. (2014). A meta-analysis exploring the relationship between psychopathy and instrumental versus reactive violence. *Criminal Justice and Behavior*, 41, 797-821.
- *Bonta, J., Blais, J., & Wilson, H. A. (2014). A theoretically informed meta-analysis of the risk for general and violent recidivism for mentally disordered offenders. *Aggression and Violent Behavior*, 19, 278-287.

- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. New York: Wiley.
- Braga, A. A., & Apel, R. (2016). And we wonder why criminology is sometimes considered irrelevant in real-world policy conversations. *Criminology and Public Policy*, in press.
- *ⁱBraga, A. A., Papachristos, A. V., & Hureau, D. M. (2014). The effects of hot spots policing on crime: An updated systematic review and meta-analysis. *Justice Quarterly*, *31*, 633-663.
- *ⁱBraga, A. A., & Weisburd, D. (2012). The effects of focused deterrence strategies on crime: A systematic review and meta-analysis of the empirical evidence. *Journal of Research in Crime and Delinquency*, *49*, 323-358.
- *ⁱBraga, A. A., Welsh, B. C., & Schnell, C. (2015). Can policing disorder reduce crime? A systematic review and meta-analysis. *Journal of Research in Crime and Delinquency*, *52*, 567-588.
- Britt, C. L., & Weisburd, D. (2010). Statistical power. In Piquero, A. R. and Weisburd, D. (eds.), *Handbook of Quantitative Criminology*, Springer, New York, pp. 313-332.
- Brown, S. E. (1989). Statistical power and criminal justice research. *Journal of Criminal Justice*, *17*, 115-22.
- Bushway, S. D., Sweeten, G., & Wilson, D. B. (2006). Size matters: Standard errors in the application of null hypothesis significance testing in criminology and criminal justice. *Journal of Experimental Criminology*, *2*, 1-22.
- Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*, 365-376.
- *Carriaga, M., & Worrall, J. L. (2015). Police levels and crime: A systematic review and meta-analysis. *The Police Journal: Theory, Practice and Principles*, *88*, 315-333.
- Carter, E. C., & McCullough, M. E. (2014). Publication bias and the limited strength model of self-control: Has the evidence for ego depletion been overestimated? *Frontiers in Psychology*, *5*, 1-11.
- Clear, T. (2010). Policy and evidence: The challenge to the American Society of Criminology: 2009 Presidential Address to the American Society of Criminology. *Criminology*, *48*, 1-25.
- Clemens, M. A. (2017). The meaning of failed replications: a review and proposal. *Journal of Economic Surveys*, *31*, 326-342.

- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd edition). New York: Lawrence Erlbaum Associates.
- Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of *p*-values. *Royal Society Open Science*, *1*, 1-16.
- *Collins, R. E. (2010). The effect of gender on violent and nonviolent recidivism: A meta-analysis. *Journal of Criminal Justice*, *38*, 675-684.
- *Cottle, C. C., Lee, R. J., & Heilburn, K. (2001). The prediction of criminal recidivism in juveniles: A meta-analysis. *Criminal Justice and Behavior*, *28*, 367-394.
- *ⁱCox, S. M., Davidson, W. S., & Bynum, T. S. (1995). A meta-analytic assessment of delinquency-related outcomes of alternative education programs. *Crime & Delinquency*, *41*, 219-234.
- Cressey, D. R. (1978). Criminological theory, social science, and the repression of crime. *Criminology*, *16*, 171-191.
- Cullen, F. T., Pratt, T. C., & Turanovic, J. J. (2016). It's hopeless: Beyond zero-tolerance supervision. *Criminology and Public Policy*, *15*, 1215-1227.
- *Derzon, J. H. (2010). The correspondence of family features with problem, aggressive, criminal, and violent behavior: A meta-analysis. *Journal of Experimental Criminology*, *6*, 263-292.
- *ⁱDowden, C., & Andrews, D. A. (2003). Does family intervention work for delinquents? Results of a meta-analysis. *Canadian Journal of Criminology and Criminal Justice*, *45*, 327-342.
- *Dowden, C., & Brown, S. L. (2002). The role of substance abuse factors in predicting recidivism: A meta-analysis. *Psychology, Crime, & Law*, *8*, 243-264.
- Duncan, L. A., & Keller, M. C. (2011). A critical review of the first 10 years of candidate gene-by-environment interaction research in psychiatry. *American Journal of Psychiatry*, *168*, 1041-49.
- Duvendack, M., Palmer-Jones, R., & Reed, W. R. (2017). What Is Meant by "Replication" and Why Does It Encounter Resistance in Economics? *American Economic Review*, *107*, 46-51.
- Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*, 1-11.

- *Edens, J. F., Campbell, J. S., & Weir, J. M. (2007). Youth psychopathy and criminal recidivism: A meta-analysis of the psychopathy checklist measures. *Law and Human Behavior, 31*, 53-75.
- Farrington, D. P., Gaffney, H. & Ttofi, M. M. (2016). Systematic reviews of explanatory risk factors for violence, offending, and delinquency. *Aggression and Violent Behavior*, forthcoming.
- *ⁱFarrington, D. P., Gill, M., Waples, S. J., & Argomaniz, J. (2007). The effect of closed-circuit television on crime: Meta-analysis of an English national quasi-experimental multi-site evaluation. *Journal of Experimental Criminology, 3*, 21-38.
- Ferguson, C. J. (2013). Spanking, corporal punishment and negative long-term outcomes: A meta-analytic review of longitudinal studies. *Clinical Psychology Review, 33*, 196-208.
- Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science, 7*, 555-561.
- Fisher, R. A. (1925/1973). *Statistical Methods for Research Workers*. New York: Hafner.
- Franco, A., Malhotra, N. & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science, 345*, 1502-05.
- *Franklin, T. W., Franklin, C. A., & Pratt, T. C. (2006). Examining the empirical relationship between prison crowding and inmate misconduct: A meta-analysis of conflicting research results. *Journal of Criminal Justice, 34*, 401-12.
- Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type s (sign) and type m (magnitude) errors. *Perspectives on Psychological Science, 9*, 641-651.
- Gelman, A., & Loken, E. (2014a). The AAA tranche of subprime science. *Chance, 27*, 51-56.
- Gelman, A., & Loken, E. (2014b). Data-dependent analysis—a “garden of forking paths”—explains why many statistically significant comparisons don't hold up. *American Scientist, 102*, 460-465.
- Gelman, A., Skardhamar, T., & Aaltonen, M. (2018). Type M error can explain Weisburd's paradox. *Journal of Quantitative Criminology*, in press.
- *ⁱGendreau, P., Little, T. & Goggin, C. (1996). A meta-analysis of the predictors of adult offender recidivism: What works! *Criminology, 34*, 575-607.
- Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on “Estimating the reproducibility of psychological science.” *Science, 351*, 1037a.

- Gill, J. (2014). *Bayesian Methods: A Social and Behavioral Sciences Approach*, Chapman & Hall/CRC, New York.
- *ⁱGobeil, R., Blanchette, K., & Stewart, L. (2016). A meta-analytic review of correctional interventions for women offenders: Gender-neutral versus gender-informed approaches. *Criminal Justice and Behavior*, *43*, 301-322.
- *Goncalves, L. C., Goncalves, R. A., Martins, C., & Dirkzwager, A. J. E. (2014). Predicting infractions and health care utilization in prison: A meta-analysis. *Criminal Justice and Behavior*, *41*, 921-942.
- Gorman, D. M. (2015). "Everything works": The need to address confirmation bias in evaluations of drug misuse prevention interventions for adolescents. *Addiction*, *110*, 1539-1540.
- *Gutierrez, L., Wilson, H. A., Rugge, T. & Bonta, J. (2013). The prediction of recidivism with Aboriginal offenders: A theoretically informed meta-analysis. *Canadian Journal of Criminology and Criminal Justice*, *55*, 55-99.
- Hagan, J. (1973). Labeling and deviance: A case study in the "sociology of the interesting." *Social Problems*, *20*, 447-458.
- Hamilton, Z., Campbell, C. M., van Wormer, J., Kirgl, A., & Posey, B. (2016). Impact of swift and certain sanctions: Evaluation of Washington State's policy for offenders on community supervision. *Criminology and Public Policy*, *15*, 1009-1072.
- Hawken, A. & Kleiman, M. A. R. (2009). *Managing drug involved probationers with swift and certain sanctions: Evaluating Hawaii's HOPE*. Washington, DC: National Institute of Justice.
- *ⁱHelmond, P., Overbeek, G., Brugman, D., & Gibbs, J. C. (2015). A meta-analysis on cognitive distortions and externalizing problem behavior: Associations, moderators, and treatment effectiveness. *Criminal Justice and Behavior*, *42*, 245-262.
- *Helmus, L., Hanson, R. K., Babchishin, K. M., & Mann, R. E. (2013). Attitudes supportive of sexual offending predict recidivism: A meta-analysis. *Trauma, Violence, & Abuse*, *14*, 34-53.
- *Helmus, L. M., Hanson, R. K., Thornton, D., Babchishin, K. M., & Harris, A. J. R. (2012). Absolute recidivism rates predicted by Static-99R and Static-2002R sex offender risk assessment tools vary across samples: A meta-analysis. *Criminal Justice and Behavior*, *39*, 1148-1171.
- *Helmus, L. M., & Thornton, D. (2015). Stability and predictive and incremental accuracy of the individual items of Static-99R and Static-2002R in predicting sexual recidivism. *Criminal Justice and Behavior*, *42*, 917-937.

- Higdon, J. V., & Frei, B. (2006). Coffee and health: A review of recent human research. *Critical Reviews in Food Science and Nutrition*, 46, 101-123.
- *ⁱHoogsteder, L. M., Starns, G. J. J. M., Figge, M. A., Changeo, K., van Horn, J. E., Hendriks, J., & Wissink, I. B. (2015). A meta-analysis of the effectiveness of individually oriented Cognitive Behavioral Treatment (CBT) for severe aggressive behavior in adolescents. *The Journal of Forensic Psychiatry & Psychology*, 26, 22-37.
- *Hsieh, C. C., & Pugh, M. D. (1993). Poverty, income inequality, and violent crime: A meta-analysis of recent aggregate data studies. *Criminal Justice Review*, 18, 182-202.
- *Hubbard, D. J., & Pratt, T. C. (2002). A meta-analysis of the predictors of delinquency among girls. *Journal of Offender Rehabilitation*, 34, 1-13.
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine* 2: e124.
- Ioannidis, J. P. A., Stanley, T. D., & Doucouliagos, H. (2015). The power of bias in economics research. Working paper: https://www.deakin.edu.au/_data/assets/pdf_file/0007/477763/2016_1.pdf
- Johnson, V. E., Payne, R. D., Wang, T., Asher, A., & Mandal, S. (2017). On the reproducibility of psychological science. *Journal of the American Statistical Association*, 112, 1-10.
- *Jones, S. E., Miller, J. D., & Lynam, D. R. (2011). Personality, antisocial behavior, and aggression: A meta-analytic review. *Journal of Criminal Justice*, 39, 329-337.
- *Joyal, C. C., Beaulieu-Plante, J., & de Chantérac, A. (2014). The neuropsychology of sex offenders: A meta-analysis. *Sexual Abuse: A Journal of Research and Treatment*, 26, 149-177.
- Jussim, L., Crawford, J. T., Anglin, S. M., Stevens, S. T., & Duarte, J. L. (2016). Interpretations and methods: Towards a more effectively self-correcting social psychology. *Journal of Experimental Social Psychology*, 66, 116-33.
- Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.
- *Kelly, P. E., Polanin, J. R., Jang, S. J., & Johnson, B. R. (2015). Religion, delinquency and drug use: A meta-analysis. *Criminal Justice Review*, 40, 505-523.
- King, G. (1995). Replication, replication. *Political Science and Politics*, 28, 444-452.
- *ⁱKoehler, J. A., Humphreys, D. K., Akoensi, T. D., de Ribera, O. S., & Lösel, F. (2014). A systematic review and meta-analysis on the effects of European drug treatment programs on reoffending. *Psychology, Crime & Law*, 20, 584-602.

- Kulig, T. C., Pratt, T. C., & Cullen, F. T. (2017). Revisiting the Stanford prison experiment: A case study in organized skepticism. *Journal of Criminal Justice Education, 28*, 74-111.
- Lajous, M., Bijon, A., Fagherazzi, G., Balkau, B., Boutron-Ruault, M. C., & Chapelon, F. (2015). Egg and cholesterol intake and incident type 2 diabetes among French women. *British Journal of Nutrition, 114*, 1667-1673.
- Lansford, J. E., Sharma, C., Malone, P. S., Woodlief, D., Dodge, K. A., et al. (2014). Corporal punishment, maternal warmth, and child adjustment: A longitudinal study in eight countries. *Journal of Clinical Child and Adolescent Psychology, 43*, 670-685.
- *¹Latimer, J. (2001). A meta-analytic examination of youth delinquency, family treatment, and recidivism. *Canadian Journal of Criminology, 43*, 237-53.
- Lattimore, P. K., MacKenzie, D. L., Zajac, G., Dawes, D., Arsenault, E., & Tueller, S. (2016). Outcome findings from the HOPE demonstration field experiment: Is swift, certain, and fair an effective supervision strategy? *Criminology and Public Policy, 15*, 1103-1141.
- Lee, P. M. (2012). *Bayesian Statistics: An Introduction*, 4th ed. West Sussex, UK: Wiley.
- *Lestico, A. R., Salekin, R. T., DeCoster, J., & Rogers, R. (2008). A large-scale meta-analysis relating the Hare measures of psychopathy to antisocial conduct. *Law and Human Behavior, 32*, 28-45.
- *Leschied, A., Chiodo, D., Nowicki, E., & Rodger, S. (2008). Childhood predictors of adult criminality: A meta-analysis drawn from the prospective longitudinal literature. *Canadian Journal of Criminology and Criminal Justice, 50*, 435-467.
- *¹Lipsey, M. W. (2009). The primary factors that characterize effective interventions with juvenile offenders: A meta-analytic overview. *Victims and Offenders, 4*, 124-147.
- Lipsey, M. W., & Wilson, D. B. (2001). *Practical Meta-Analysis*. Thousand Oaks, CA: Sage.
- *Lytle, D. J. (2014). The effects of suspect characteristics on arrest: A meta-analysis. *Journal of Criminal Justice, 42*, 589-597.
- Maltz, M. D. (1994). Deviating from the mean: The declining significance of significance. *Journal of Research in Crime and Delinquency, 31*, 434-63.
- Maruna, S. (2015). Qualitative research, theory development, and evidence-based corrections: can success stories be “evidence.” Pp. 311-337 in Jody Miller and Wilson R. Palacios (Eds.), *Qualitative research in criminology—Advances in criminological theory*. Vol. 20. New Brunswick, NJ: Transaction.

- Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does “failure to replicate” really mean? *American Psychologist*, *70*, 487-98.
- *McCann, K., & Lussier, P. (2008). Antisociality, sexual deviance, and sexual reoffending in juvenile sex offenders. *Youth Violence and Juvenile Justice*, *6*, 363-385.
- McNeeley, S., & Warner, J. J. (2015). Replication in criminology: A necessary practice. *European Journal of Criminology*, *12*, 581-97.
- McShane, B. B., & Gal, D. (2015). Blinding us to the obvious? The effect of statistical training on the evaluation of evidence. *Management Science*, *62*, 1707-18.
- Merton, R. K. (1957). Priorities in scientific discovery: A chapter in the sociology of science. *American Sociological Review*, *22*, 635-659.
- Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. Chicago: University of Chicago Press.
- *Miller, J. D., & Lynam, D. (2001). Structural models of personality and their relation to antisocial behavior: A meta-analytic review. *Criminology*, *39*, 765-798.
- *ⁱMitchell, O., Wilson, D. B., Eggers, A., & MacKenzie, D. L. (2012). Assessing the effectiveness of drug courts on recidivism: A meta-analytic review of traditional and non-traditional drug courts. *Journal of Criminal Justice*, *40*, 60-71.
- *Moore, T. M., Scarpa, A., & Raine, A. (2002). A meta-analysis of serotonin metabolite 5-HIAA and antisocial behavior. *Aggressive Behavior*, *28*, 299-316.
- Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E. J. (2015). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*, 103-23.
- Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., du Sert, N. P., Simonsohn, U., Wagenmakers, E. J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto on reproducible science. *Nature Human Behavior*, *1*, 1-9.
- Nakamura, Y., Iso, H., Kita, Y., Ueshima, H., Okada, K. et al. (2006). Egg consumption, serum total cholesterol concentrations and coronary heart disease incidence: Japan Public Health Center-based prospective study. *British Journal of Nutrition*, *96*, 921-928.
- National Academies of Sciences, Engineering, and Medicine. (2016). *Statistical challenges in assessing and fostering the reproducibility of scientific results: Summary of a workshop*.
- Nelson, M. S., Wooditch, A., & Dario, L. M. (2015). Sample size, effect size, and statistical power: A replication study of Weisburd’s paradox. *Journal of Experimental Criminology*, *11*, 141-63.

- Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *General Review of Psychology*, 2, 175-220.
- *Nivette, A. E. (2011). Cross-national predictors of crime: A meta-analysis. *Homicide Studies*, 15, 103-131
- Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., et al. (2015). Promoting an open research culture. *Science*, 348, 1422-1425.
- Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, 7, 615-31.
- Nuijten, M. B., van Assen, M. A. L. M., Augusteijn, H. E. M., Cromptvoets, E. A. V., & Wicherts, J. M. (2018). Effect sizes, power, and biases in intelligence research: A meta-meta-analysis. Preprint retrieved from <https://psyarxiv.com/ytsvw>.
- O'Connell, D. J., Brent, J. J., & Visher, C. A. (2016). Decide your time: A randomized trial of a drug testing and graduated sanctions program for probationers. *Criminology and Public Policy*, 15, 1073-1102.
- *Ogilvie, C. A., Newman, E., Todd, L., & Peck, D. (2014). Attachment and violent offending: A meta-analysis. *Aggression and Violent Behavior*, 19, 322-339.
- *Ogilvie, J. M., Stewart, A. L., Chan, R. C. K., & Shum, D. H. K. (2011). Neuropsychological measures of executive function and antisocial behavior: A meta-analysis. *Criminology*, 49, 1063-1107.
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349, aac4716.
- Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, 7, 528-30.
- Petrosino, A., Boruch, R. F., Soydan, H., Duggan, L., & Sanchez-Meca, J. (2001). Meeting the challenges of evidence-based policy: The Campbell Collaboration. *Annals of the American Academy of Political and Social Science*, 578, 14-34.
- Pickett, J. T., & Roche, S. P. (2018). Questionable, objectionable or criminal? Public opinion on data fraud and selective reporting in science. *Science and Engineering Ethics*, 24, 151-71.
- *ⁱPiquero, A. P., Farrington, D. P., Welsh, B. C., Tremblay, R., & Jennings, W. G. (2009). Effects of early family/parent training programs on antisocial behavior and delinquency. *Journal of Experimental Criminology*, 5, 83-120.

- *Portnoy, J., & Farrington, D. P. (2015). Resting heart rate and antisocial behavior: An updated systematic review and meta-analysis. *Aggression and Violent Behavior, 22*, 33-45.
- *Pratt, T. C., & Cullen, F. T. (2000). The empirical status of Gottfredson and Hirschi's general theory of crime: A meta-analysis. *Criminology, 38*, 931-964.
- *ⁱPratt, T. C., & Cullen, F. T. (2005). Assessing the macro-level predictors and theories of crime: A meta-analysis. *Crime and Justice, 32*, 373-450.
- *Pratt, T. C., Cullen, F. T., Blevins, K. R., Daigle, L. E., & Madensen, T. D. (2006). The empirical status of deterrence theory: A meta-analysis. In Cullen, F. T., Wright, J. P., and Blevins, K. R. (eds.), *Taking Stock: The Status of Criminological Theory—Advances in Criminological Theory, Volume 15*, Transaction, New Brunswick, NJ, pp. 367-395.
- *Pratt, T. C., Cullen, F. T., Blevins, K. B., Daigle, L. E., & Unnever, J. D. (2002). The relationship of Attention Deficit Hyperactivity Disorder to crime and delinquency: A meta-analysis. *International Journal of Police Science and Management, 4*, 344-360.
- *Pratt, T. C., Cullen, F. T., Sellers, C. S., Winfree Jr. L. T., Madensen, T. D., Daigle, L. E., Fearn, N. E., & Gau, J. M. (2010). The empirical status of social learning theory: A meta-analysis. *Justice Quarterly, 27*, 765-802.
- *Pratt, T. C., McGloin, J. M. & Fearn, N. E. (2006). Maternal cigarette smoking during pregnancy and criminal/deviant behavior. *International Journal of Offender Therapy and Comparative Criminology, 50*, 672-690.
- Pratt, T. C., Turanovic, J. J., & Cullen, F. T. (2016). Revisiting the criminological consequences of exposure to fetal testosterone: A meta-analysis of the 2D:4D digit ratio. *Criminology, 54*, 587-620.
- Pratt, T. C., Turanovic, J. J., Fox, K. A., & Wright, K.A. (2014). Self-control and victimization: A meta-analysis. *Criminology, 52*, 87-116.
- *ⁱPrendergast, M. L., Pearson, F. S., Podus, D., Hamilton, Z. K., & Greenwell, L. (2013). The Andrews' principles of risk, needs, and responsivity as applied in drug treatment programs: Meta-analysis of crime and drug use outcomes. *Journal of Experimental Criminology, 9*, 275-300.
- Pridemore, W. A., Makel, M. C., & Plucker, J. A. (2018). Replication in criminology and the social sciences. *Annual Review of Criminology, 1*, 19-38.
- *Pyrooz, D. C., Turanovic, J. T., Decker, S. H., & Wu, J. (2016). Taking stock of the relationship between gang membership and offending. *Criminal Justice and Behavior, 43*, 365-397.

- *ⁱRedondo, S., Sánchez-Meca, J., & Garrido, V. (1999). The influence of treatment programmes on the recidivism of juvenile and adult offenders: An European meta-analytic review. *Psychology, Crime & Law*, 5, 251-78.
- *ⁱReitzel, L. R., & Carbonell, J. L. (2006). The effectiveness of sexual offender treatment for juveniles as measured by recidivism: A meta-analysis. *Sex Abuse*, 18, 401-421.
- Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology*, 52, 59-82.
- Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, 44, 1276-1284.
- Sampson, R. J. (2010). Gold standard myths: Observations in the experimental turn in quantitative criminology. *Journal of Quantitative Criminology*, 26, 489-500.
- *Savage, J., & Yancey, C. (2008). The effects of media violence exposure on criminal aggression: A meta-analysis. *Criminal Justice and Behavior*, 35, 772-791.
- Schmidt, F. L., & Oh, I. S. (2016). The crisis of confidence in research findings in psychology: Is lack of replication the real problem? Or is it something else? *Archives of Scientific Psychology*, 4, 32-37.
- *ⁱSchmucker, M., & Lösel, F. (2015). The effects of sexual offender treatment on recidivism: An international meta-analysis of sound quality evaluations. *Journal of Experimental Criminology*, 11, 597-630.
- *ⁱSchwalbe, C. S., Gearing, R. E., MacKenzie, M. J., Brewer, K. B. & Ibrahim, R. (2012). A meta-analysis of experimental studies of diversion programs for juvenile offenders. *Clinical Psychology Review*, 32, 26-33.
- Sham, P. C., & Purcell, S. M. (2014). Statistical power and significance testing in large-scale genetic studies. *Nature Reviews Genetics*, 15, 335-46.
- Sherman, L. 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.
- Sherman, L. W., Farrington, D. P., Welsh, B. C., & MacKenzie, D. L. (2002). *Evidence-Based Crime Prevention*. New York: Routledge.
- Sherman, L. W., Gottfredson, D., MacKenzie, D., Eck, J., Reuter, P., & Bushway, S. (1997). *Preventing Crime: What Works, What Doesn't, What's Promising: A Report to the United States Congress*. Washington, DC: National Institute of Justice.

- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359-1366.
- *Smith, P., Cullen, F. T. & Latessa, E. J. (2009). Can 14,737 women be wrong? A meta-analysis of the LSI-R and recidivism for female offenders. *Criminology & Public Policy, 8*, 183-208.
- Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science, 9*, 59-71.
- *ⁱTaheri, S. A., & Welsh, B. C. (2015). After-school programs for delinquency prevention: A systematic review and meta-analysis. *Youth Violence and Juvenile Justice, 14*, 272-290.
- *Teng, Z., Liu, Y., & Guo, C. (2015). A meta-analysis of the relationship between self-esteem and aggression among Chinese students. *Aggression and Violent Behavior, 21*, 45-54.
- *ⁱTolan, P. H., Henry, D. H., Schoeny, M. S., Lovegrove, P., & Nichols, E. (2014). Mentoring programs to affect delinquency and associated outcomes of youth at risk: A comprehensive meta-analytic review. *Journal of Experimental Criminology, 10*, 179-206.
- *van Langen, M. A. M., Wissink, I. B., van Vugt, E. S., van der Stouwe, T. & Stams, G. J. J. M. (2014). The relation between empathy and offending: A meta-analysis. *Aggression and Violent Behavior, 19*, 179-189.
- *van Vugt, E., Gibbs, J., Stams, G. J., Bijleveld, C., Hendriks, J. & van der Laan, P. (2011). Moral development and recidivism: A meta-analysis. *International Journal of Offender Therapy and Comparative Criminology, 55*, 1234-1250.
- *ⁱVisher, C. A., Winterfield, L., & Coggeshall, M. B. (2005). Ex-offender employment programs and recidivism: A meta-analysis. *Journal of Experimental Criminology, 1*, 295-315.
- Vostal, F. (2016). *Accelerating academia. The changing structure of academic time*. Basingstoke: Palgrave Macmillan.
- *Walters, G. D. (1992). A meta-analysis of the gene-crime relationship. *Criminology, 30*, 595-613.
- *Walters, G. D. (2003). Predicting institutional adjustment and recidivism with the psychopathy checklist factor scores: A meta-analysis. *Law and Human Behavior, 27*, 541-558.
- *Walters, G. D. (2006). Risk-appraisal versus self-report in the prediction of criminal justice outcomes: A meta-analysis. *Criminal Justice and Behavior, 33*, 279-304.

- *Walters, G. D. (2012). Criminal thinking and recidivism: Meta-analytic evidence on the predictive and incremental validity of the Psychological Inventory of Criminal Thinking Styles (PICTS). *Aggression and Violent Behavior, 17*, 272-278.
- *Walters, G. D. (2013). Testing the specificity postulate of the violence graduation hypothesis: Meta-analyses of the animal cruelty-offending relationship. *Aggression and Violent Behavior, 18*, 797-802.
- Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on *p*-values: Contest, process, and purpose. *The American Statistician, 70*, 129-33.
- Weisburd, D., Lum, C. M., & Yang, S. M. (2003). When can we conclude that treatments or programs "don't work"? *The Annals of the American Academy of Political and Social Science, 587*, 31-48.
- Weisburd, D., Petrosino, A. & Mason, G. (1993). Design sensitivity in criminal justice experiments. *Crime and Justice, 17*, 337-79.
- Weisburd, D., & Piquero, A. R. (2008). How well do criminologists explain crime? Statistical modeling in published studies. *Crime and Justice, 37*, 453-502.
- *ⁱWelsh, B. C., & Farrington, D. P. (2009). Public area CCTV and crime prevention: An updated systematic review and meta-analysis. *Justice Quarterly, 26*, 716-745.
- *ⁱWilson, D. B., Gallagher, C. A., & MacKenzie, D. L. (2000). A meta-analysis of corrections-based education, vocation, and work programs for adult offenders. *Journal of Research in Crime and Delinquency, 37*, 347-368.
- *ⁱWilson, D. B., Mitchell, O., & MacKenzie, D. L. (2006). A systematic review of drug court effects on recidivism. *Journal of Experimental Criminology, 2*, 459-487.
- *ⁱWilson, H. A. (2014). Can antisocial personality disorder be treated? A meta-analysis examining the effectiveness of treatment in reducing recidivism for individuals with ASPD. *Journal of Forensic Mental Health Services, 13*, 36-46.
- *ⁱWilson, H. A., & Hoge, R. D. (2013). The effect of youth diversion programs on recidivism. *Criminal Justice and Behavior, 40*, 497-518.
- *Wilson, H. A., & Gutierrez, L. (2014). Does one size fit all? The meta-analysis examining the predictive ability of the Level of Service Inventory (LSI) with Aboriginal offenders. *Criminal Justice and Behavior, 41*, 196-219.
- *ⁱWood, S., & Mayo-Wilson, E. (2012). School-based mentoring for adolescents: A systematic review and meta-analysis. *Research on Social Work Practice, 22*, 257-269.

- Woodward, M., & Tunstall-Pedoe, H. (1999). Coffee and tea consumption in the Scottish Heart Health Study follow up: Conflicting relations with coronary risk factors, coronary disease, and all cause mortality. *Journal of Epidemiology and Community Health, 53*, 481-487.
- Woodward, V. H., Webb, M. E., Griffin III, O. H., & Copes, H. (2016). The current state of criminological research in the United States: An examination of research methodologies in criminology and criminal justice journals. *Journal of Criminal Justice Education, 27*, 340-361.
- Worrall, J. L. (2000). In defense of the “quantoids”: More on the reasons for the quantitative emphasis in criminal justice education and research. *Journal of Criminal Justice Education, 11*, 353-361.

Table 1. Four Results of a Null Hypothesis Significance Test

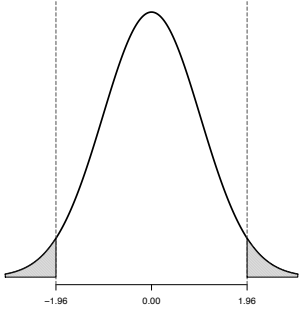
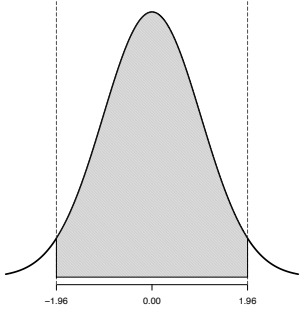
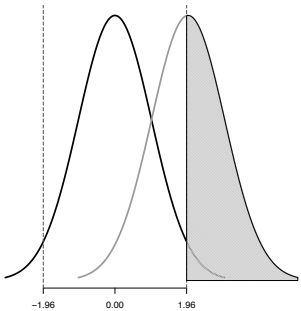
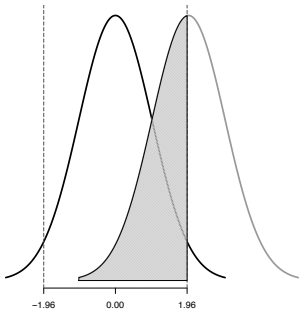
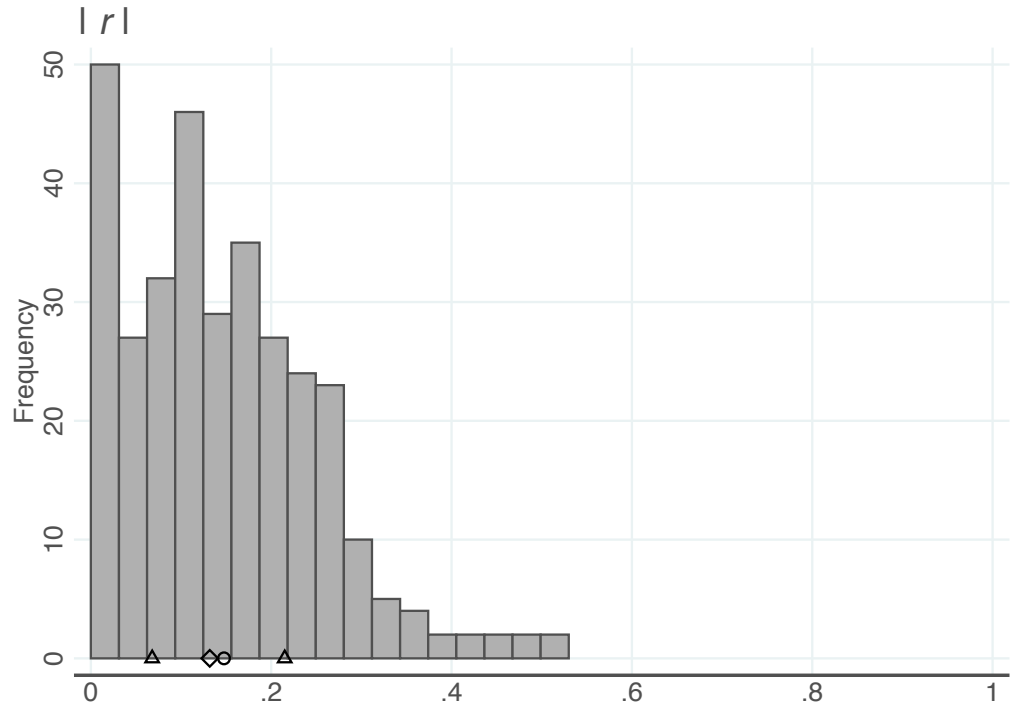
		Researcher's Decision	
		Reject H_0	Fail to Reject H_0
Reality $H_0 = \text{True}$	α ; Type I Error		
$H_0 = \text{False}$	$1 - \beta$; Statistical Power; Correct Decision		

Table 2. Effect Sizes ($|r|$), Sample Sizes (n), & Statistical Power ($1 - \beta$) Drawn from Criminology Meta-analyses

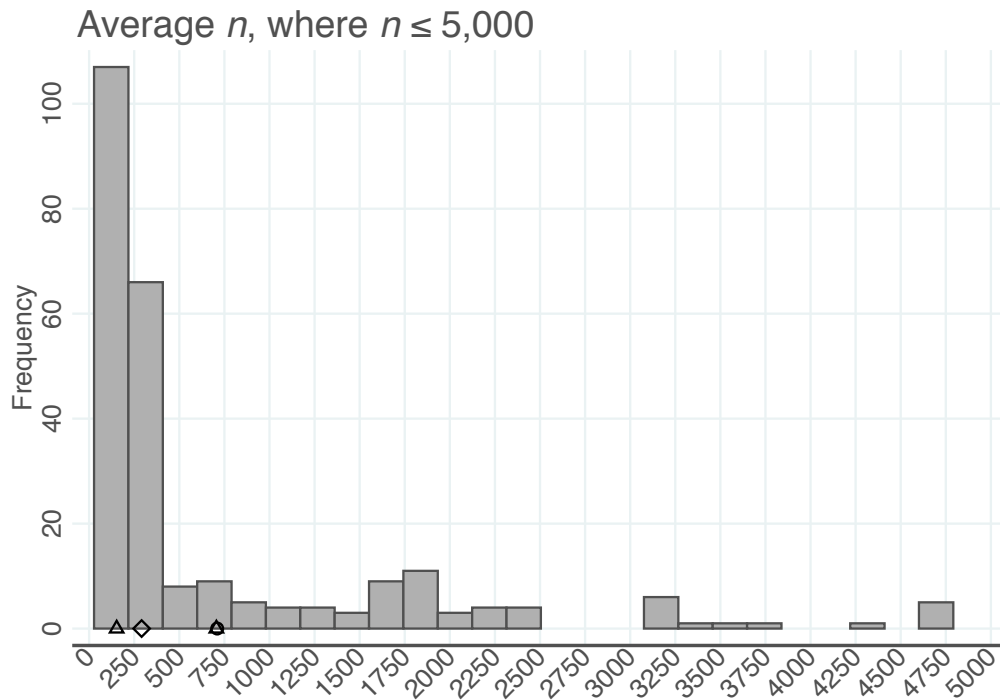
	Observations	Mean	Median	Mode	Standard Deviation	Min	Max	95% Range
Effect Size, r								
All Studies	322	0.148	0.132	0.030	0.103	0.000	0.530	0.010; 0.409
Intervention Studies	41	0.120	0.110	0.138	0.095	0.015	0.495	0.015; 0.485
Non-Intervention Studies	281	0.152	0.143	0.030	0.104	0.000	0.530	0.010; 0.409
Individual-Level Studies	253	0.147	0.138	0.030	0.101	0.000	0.530	0.010; 0.380
Macro-Level Studies	69	0.150	0.110	0.072	0.109	0.012	0.445	0.012; 0.441
Sample Size, n								
All Studies	271	2,929.821	327.429	44	14,196.600	27.333	190,586.500	44; 21,890.740
Intervention Studies	27	627.084	331.778	3,492.763	774.695	27.333	3,492.763	27.333; 3,492.763
Non-Intervention Studies	244	3,184.632	320.239	44	14,940.510	44	190,586.500	44; 24,353.630
Individual-Level Studies	236	3,292.281	351.638	62	15,172.19	27.333	190,586.500	62; 24,911.130
Macro-Level Studies	35	485.804	44	44	1,541.675	44	6,476.400	44; 6,476.400
Sample Size, where $n \leq 5,000$								
All Studies	252	710.122	291.125	44	1,005.293	27.333	4,791.833	44; 4,138.697
Intervention Studies	27	627.084	331.778	3,492.763	774.695	27.333	3,492.763	27.333; 3,492.763
Non-Intervention Studies	225	720.086	288.250	44	1,030.455	44	4,791.833	44; 4,433.586
Individual-Level Studies	219	798.632	332.613	62	1,040.195	27.333	4,791.833	62; 4,491.143
Macro-Level Studies	33	122.738	44	44	382.651	44	2,225.667	44; 2,225.667
Statistical Power ($1 - \beta$)								
All Studies	270	0.605	0.706	1.000	0.364	0.051	1.000	0.052; 1.000
Intervention Studies	27	0.534	0.515	1.000	0.352	0.055	1.000	0.055; 1.000
Non-Intervention Studies	243	0.613	0.755	1.000	0.365	0.050	1.000	0.052; 1.000
Individual-Level Studies	235	0.654	0.786	1.000	0.351	0.051	1.000	0.055; 1.000
Macro-Level Studies	35	0.274	0.184	0.184	0.270	0.051	1.000	0.051; 1.000

Figure 1. Distribution of 322 Effect Sizes Observed in 81 Criminology Meta-analyses



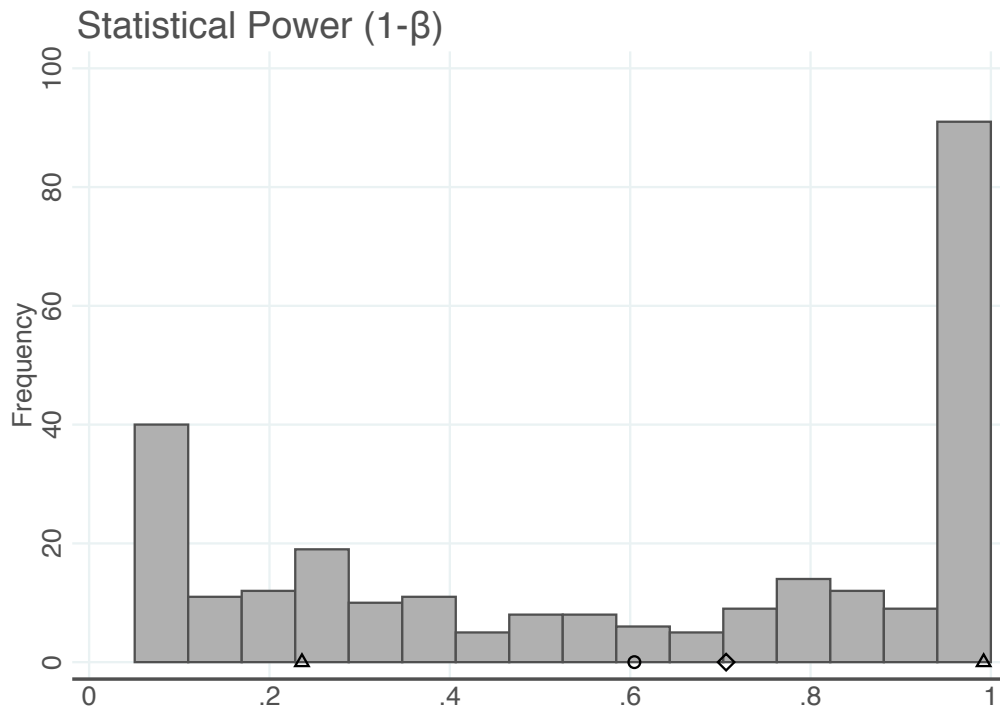
Notes: The location of the 25th percentile value is denoted with the *triangle* marker on the left, the location of the median value is denoted with the *diamond* marker, the location of the mode value is denoted with the *circle* marker, and the location of the 75th percentile value is denoted with the *triangle* marker on the right.

Figure 2. Distribution of 252 Average Sample Sizes Observed in Criminology Meta-analyses, where $n \leq 5,000$



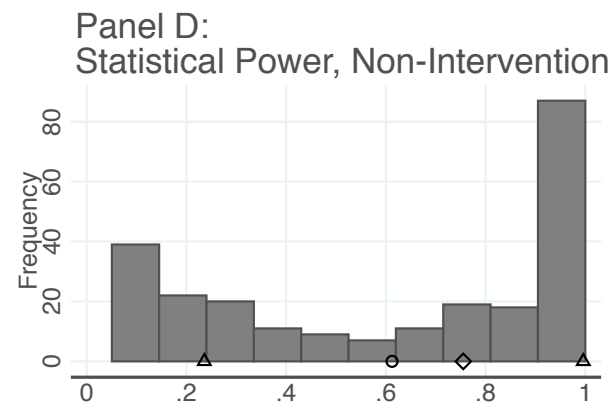
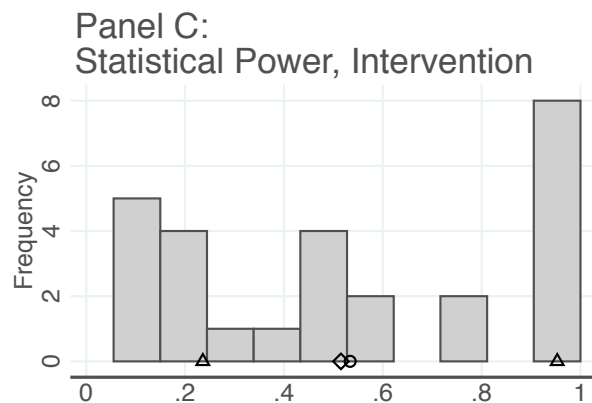
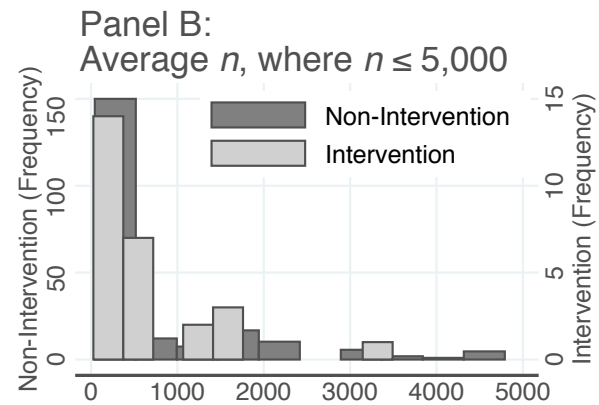
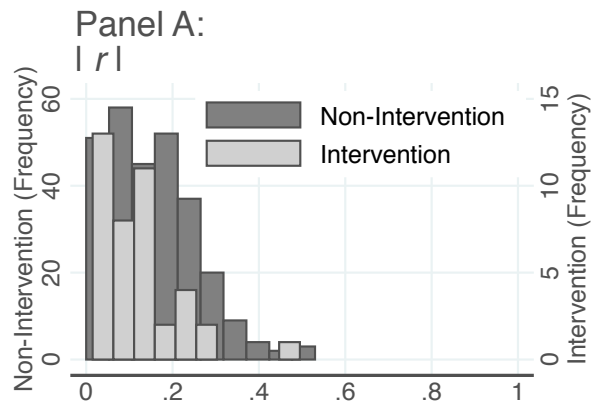
Notes: The location of the 25th percentile value is denoted with the *triangle* marker on the left, the location of the median value is denoted with the *diamond* marker, the location of the mean value is denoted with the *circle* marker, and the location of the 75th percentile value is denoted with the *triangle* marker on the right.

Figure 3. Statistical Power Estimates from 270 Effect Size & Sample Size Combinations Observed in Criminology Meta-Analyses



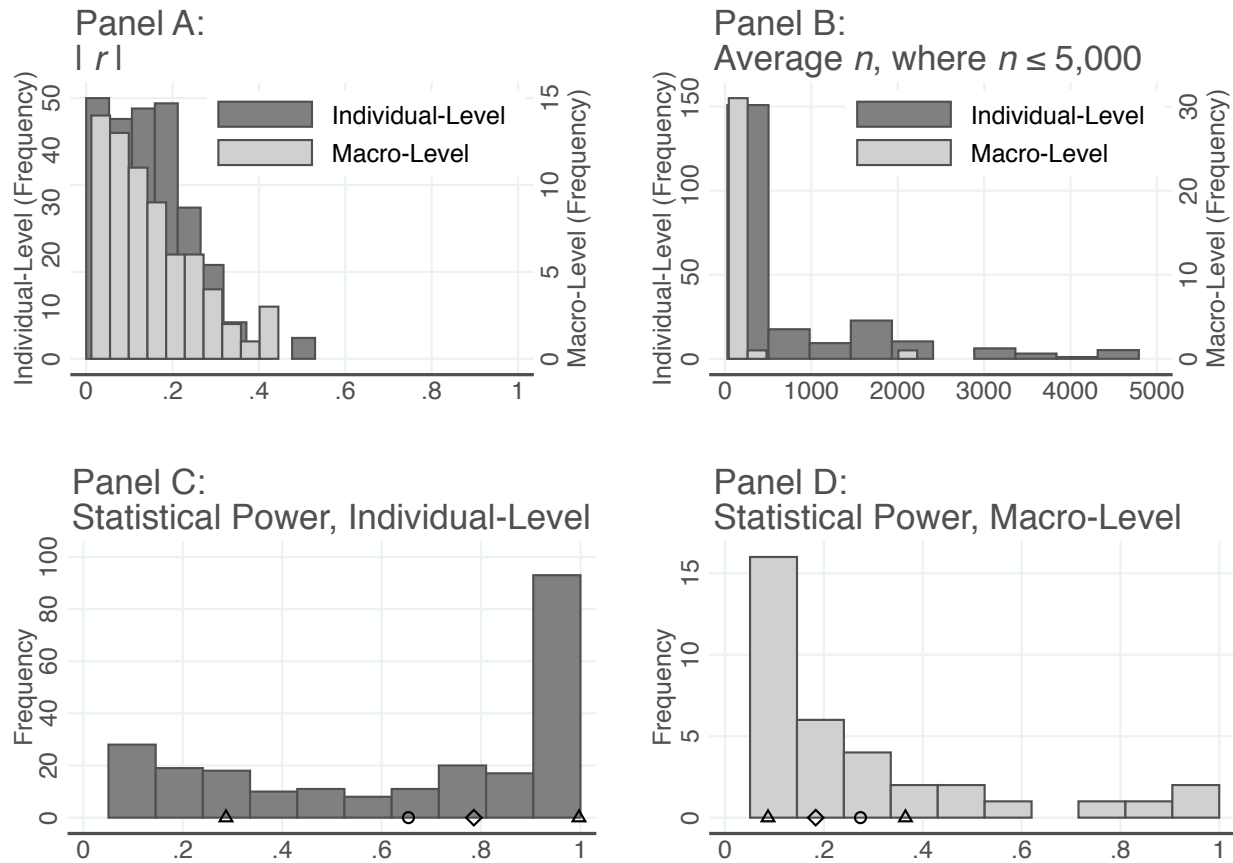
Notes: The location of the 25th percentile value is denoted with the *triangle* marker on the left, the location of the median value is denoted with the *diamond* marker, the location of the mean value is denoted with the *circle* marker, and the location of the 75th percentile value is denoted with the *triangle* marker on the right.

Appendix A. Intervention & Non-Intervention Studies



Notes: The location of the 25th percentile value is denoted with the *triangle* marker on the left, the location of the median value is denoted with the *diamond* marker, the location of the mean value is denoted with the *circle* marker, and the location of the 75th percentile value is denoted with the *triangle* marker on the right.

Appendix B. Individual-level & Macro-level Studies



Notes: The location of the 25th percentile value is denoted with the *triangle* marker on the left, the location of the median value is denoted with the *diamond* marker, the location of the mean value is denoted with the *circle* marker, and the location of the 75th percentile value is denoted with the *triangle* marker on the right.