Journal of Quantitative Criminology (2020) 36:263–272
https://doi.org/10.1007/s10940-020-09463-9

EDITORIAL

Standard Errors in Quantitative Criminology: Taking Stock and Looking Forward

Gary Sweeten[1]

Published online: 29 May 2020

Introduction

Criticisms of null hypothesis significance testing are as old as the practice itself (e.g. Berkson 1942; Boring 1919; Rozeboom 1960) and have continued at a steady pace to the present, including some from within the discipline of criminology (e.g. Bushway et al. 2006; Maltz 1994; Weisburd et al. 2003). Common reservations raised by critics of this practice include: its ritualized nature, the confounding of statistical and substantive significance, lack of attention to research methods that invalidate the conclusions, misinterpretation of the meaning of *p* values, and encouragement of binary thinking about research evidence. In recent decades new critiques have arisen that focus on the cumulative effect of these problems, most notably in psychology's "replication crisis". The defining moment of psychology's replication crisis was the report, published in *Science*, on a collaborative effort to directly replicate 100 studies that had been published in top-ranked psychology journals in 2008 (Open Science Collaboration 2015). The mean effect size in the replication studies was half that of the original studies. Nearly all (97%) of the original studies produced statistically significant results compared to just 36% of the replications. Less than half of the effect sizes obtained in the replications were within the 95% confidence intervals produced by the original studies. These results fed into a rising chorus of calls to reform the scientific process across multiple disciplines to increase confidence in the reliability of the published record.

Null hypothesis significance testing, and in particular the production and meaning of *p* values, resides at the center of psychology's replication crisis and unreliable research in general. Briefly, this practice begins with a long list of assumptions including a null hypothesis, and the *p* value reflects the likelihood of the observed data assuming the null hypothesis is true and *all other model assumptions are valid* (Greenland et al. 2016; Wasserstein and Lazar 2016). Importantly, a small *p* value does not tell us which of the model's assumptions are unlikely. We may observe a small *p* value because of a violation of the Gauss-Markov assumptions. Fixes to such violations are well-known and some more advanced fixes are presented in this issue (e.g. Moody and Marvell 2020; Thomas et al. 2019). Some less well-known model assumptions are equally important. We may see a small *p* value because the author selected it after screening several alternative models, a

✉ Gary Sweeten
Gary.Sweeten@asu.edu

[1] Arizona State University, Tempe, USA

practice known as *p*-hacking. Or perhaps many research teams attempted the same experiment and we are seeing only the "successful" experiment, a result of publication bias. Perhaps the author of the study stumbled upon an interesting result and has presented a null hypothesis pre-determined to be rejected, a practice known as "harking" – hypothesizing after results are known (Kerr 1998). Only if we can rule out all of these possibilities (and more) can we attribute a small *p* value to a false null hypothesis. Because of the many paths to small *p* values, such findings may fail to replicate.

As the replication crisis rocked the field of psychology and other disciplines, criminology has been slow to take note. There has been increasing interest in replication within criminology (e.g. McNeeley and Warner 2015; Pridemore et al. 2018) and a special issue containing replications was recently published (Savolainen and VanEseltine 2018). However, there has been little attention paid to the causes of the replication crisis, whether criminology suffers from the same issues, and how to address such problems.

This special issue was launched in order to raise awareness of such issues in criminology. The call for papers asked for manuscripts that: (1) identify the causes of false positives, (2) document the prevalence of false positives, and (3) discuss, detail, and demonstrate best practices for moving the discipline forward. I would revise the first two categories today, replacing "false positives" with "unreliable research" as the term "false positive" itself reifies a dichotomous view of scientific evidence that is better viewed as a continuum not primarily about which side of .05 a *p* value is on. The papers in this special issue addressed all of these topic areas, with somewhat more of an emphasis on fixing rather than documenting the problem. I will discuss each topic area in turn, drawing on evidence from outside criminology as well as the papers in this issue.

Causes of Unreliable Research

Researchers in tenure-track positions experience strong pressures to publish peer-reviewed research. Top journals seek novel research findings, which almost always are statistically significant findings. This means that there is a large group of criminologists constantly seeking to produce interesting, statistically significant, findings. They do so in an environment that does not require the sharing of code or data, and that rarely subjects published research to direct replications. This allows and even encourages researchers to engage in practices that increase their chances of obtaining low *p* values, potentially at the cost of producing unreliable research.

The most compelling cause of unreliable research is undisclosed researcher flexibility, detailed by Simmons et al. (2011). Criminologists have practically unlimited researcher degrees of freedom when conducting a study: when to stop data collection, whether to exclude certain cases, which control variables to include or exclude, how to scale independent and dependent variables, whether to transform certain measures by taking the natural log or adding a square term, which interactions to test, which model to use, how to deal with missing data. Consider, for example, a situation in which there are two potential dependent variables one could employ that are correlated at 0.5. If a researcher runs just two regression models and chooses the model with the smaller *p* value for the focal independent variable, the chances that one of the two *p* values is less than .05 in a situation when the true effect is zero for both is 9.5%. Screening two potential dependent variables raises the effective alpha level from .05 to .095. Simmons and colleagues show that with just a few degrees of freedom the nominal alpha level is meaningless: "it is

unacceptably easy to publish 'statistically significant' evidence consistent with *any* hypothesis" (2011:1359, emphasis in original). Practices such as p-hacking, harking and even data fabrication all fall under the umbrella of undisclosed researcher flexibility. The problematic quality of these decisions is that they are undisclosed. Decisions that are disclosed are not problematic because reviewers, editors and readers would be able to assess the evidence in light of these decisions.

Wooditch et al. (2020a, b, this issue) provide compelling new evidence on the effects of undisclosed researcher flexibility. They compare 95 pre-registered clinical trials on substance use to their final publications. Pre-registration is a commonly-cited strategy for limiting researcher degrees of freedom as it is intended to specify all research protocols before the study takes place. Wooditch and colleagues document an alarming degree of deviation from the planned protocols. They also find that studies with more deviations from their pre-registration exhibit higher average effect sizes. Outcomes that are added to publications but not included in the pre-registration have substantively larger (.385) effects sizes in terms of Cohen's D. This study is striking and alarming for a number of reasons. First, it shows that one of the most recommended strategies to curb unreliable research fails to limited researcher flexibility. Second, it documents how the exercise of researcher degrees of freedom is associated with inflated effect sizes. If these processes take place in pre-registered studies, how much more likely are they to take place in un-registered studies subject to much less scrutiny?

The terms "p-hacking" and "harking" imply some ill-intent on the part of the researcher, but researcher degrees of freedom may bend results towards desired outcomes without malfeasance. Careful and earnest scientists need only make data-dependent decisions on the way to their final models in order to inflate effective alpha levels. Gelman and Loken (2014) term the many decisions that are made on the way to the final paper as the "garden of forking paths". This is important to point out because some researchers may feel that their own research is free of problems because they are not actively engaged in p-hacking. Hypothesis testing can be corrupted in much more subtle ways.

Gelman et al. (2020) contribute a thought-provoking piece to this issue on the topic of Weisburd's Paradox, a term coined by Sherman (2007) in reference to the observation that post hoc statistical power does not increase as sample size increases in criminal justice experiments (Weisburd et al. 1993). Weisburd and colleagues pointed out that larger experiments tend to be of lower quality when compared to earlier smaller experiments, a phenomenon widely recognized in the prevention science literature (Gottfredson et al. 2015). Gelman and colleagues suggest an alternative view, that effect size, and thus post hoc statistical power, is inflated in smaller experiments that have passed through the publication bias filter. If statistical significance is practically required for publication, then effect sizes of smaller experiments must be larger in order to qualify for publication since standard errors are larger in small samples. Post-hoc power calculations based on published research are subject to all the same biases of researcher degrees of freedom and so are almost certainly inflated estimates of a priori power (e.g. Barnes et al. 2020). Gelman and colleagues recommend less of a focus on a dichotomous view of statistical evidence and a greater embrace of uncertainty, a point on which Weisburd and Gelman agree (Farrington et al. 2019).

While it is easy for researchers to unconsciously bias their results towards statistical significance, there is also plenty of opportunity for researchers to actively distort the research record. Estimates of lifetime prevalence of data falsification in psychology range from 2% based on self-disclosure (Fanelli 2009) to 10% based on estimates of prevalence among colleagues (John et al. 2012). Other questionable research practices, such as failing to

disclose certain data processing decisions, dropping cases, and rounding down *p* values are much more prevalent. It is not known how prevalent such practices are in the discipline of criminology.

The review process itself is even less often scrutinized than researcher degrees of freedom and questionable research practices. As the gatekeeper for published research, editors and journal reviewers play a profound role in what gets published and where. Apart from the widespread acknowledgement that statistically significant results are more likely to be published than null results, biases may influence publishing decisions. In an intriguing experiment, Mahoney (1977) tested the role of confirmatory bias among reviewers by experimentally manipulating the contents of a paper supposedly under review, simultaneously sent to 75 reviewers. The introduction and methods sections of these papers, the sections that should weigh most heavily in assessments of scientific rigor, were identical across all reviewers. The results and conclusions sections were manipulated so that some reviewers received results that corresponded to their known prior beliefs while others received papers with results contrary to their prior beliefs. Reviewers who received papers that confirmed their prior beliefs rated the papers much more positively and only 25% pointed out a known error in the paper. Reviewers who received papers that disconfirmed their prior beliefs rated the papers much lower and 71% detected the same known error. This quite obviously goes against the norm of disinterestedness in science, whereby participants in the scientific process should not favor one outcome over another (Merton 1942). Should such confirmatory biases exist in criminology, the research record will be distorted so that findings that counter orthodox views will be suppressed.

The most extreme cases of questionable research practices, when detected in the literature, sometimes result in retraction. Records of retraction can shed some light on the causes of unreliable research. First, rates of retracted articles are sharply on the rise (Grieneisen and Zhang 2012). This may indicate a sharp increase in questionable practices, increased detection and sanctioning of such practices, or both. Second, journal impact factor is strongly correlated with retraction rates (Fang and Casadevall 2011). This could indicate that researchers are more apt to engage in questionable research practices in order to publish in the highest impact outlets, that post-publication scrutiny of such publications is much higher, or that editors of higher impact journals are more apt to retract articles when problems arise. Until last year, no journal in the discipline of criminology had ever retracted an article (but see Johnson et al. 2011; Stewart et al. 2018). That the first two retractions were from the flagship journal in the discipline is not surprising. It is more surprising that this had never happened before. A comprehensive analysis of over 4000 retractions from 1928 to 2011 found that less than half (45%) were for questionable data or research misconduct (Grieneisen and Zhang 2012). As many retractions (46%) were for publishing misconduct such a plagiarism, duplicate publication or authorship issues. Like most misconduct, it is quite likely that the vast majority of research misconduct goes undetected.

Evidence of Misleading Research

The greatest asset in detecting whether the body of published literature is in some way a biased record of scientific evidence is to compared it to "gray literatures" such as dissertations, conference abstracts, institutional review board applications, etc. Such literatures are unaffected

by the pressures research is subjected to on the path to publication but are still subject to personal biases of researchers.

In a review of dissertations published in the field of management, O'Boyle et al. (2017) found that 45% of research hypotheses were supported. When they tracked these dissertations to their corresponding peer-reviewed publications they found that 66% of research hypotheses were supported. They detected many different methods of undisclosed researcher degrees of freedom that accounted for the change. Some unsupported hypotheses were dropped, others were reversed, and some new ones were added. Variables were added, others deleted, and there was evidence of data alterations. It is not known whether such changes would be evident in criminology dissertations, but like criminology, the management literature was one where data was not required to be shared and replication was rare.

In another example of leveraging gray literature, Cooper et al. (1997) followed up on 159 psychology studies that had been approved by an institutional review board. Of these studies, 117 had been pursued through data analysis. In 72 instances, the null hypothesis had been rejected and 74% of these had been submitted to a journal. Only 2 of the 45 studies (4%) in which the null hypothesis had not been rejected were submitted to a journal. Thus, the ratio of "successful" to "unsuccessful" studies was 1.6:1 among the studies approved by the IRB, and 26.5:1 among the studies submitted for publication. A person who only read the published literature would have to conclude that the life of a scientist is one of nearly uninterrupted success. This is not a new phenomenon. Sterling (1959) reported that 97% of experiments reported in four top psychology journals had rejected their null hypotheses. The improbably successful nature of the published record was quantified by Francis (2014) as "excess success": rejection of null hypotheses at a higher rate than appropriate given statistical power.

If researchers have their fingers on the scales of statistical significance it will show up in the distribution of p values in the published record. Once a p value dips below .05 there would be no further need to modify the study. Several studies capitalize on this quality by comparing the prevalence of p values just below the .05 threshold to those just above it. If there were no manipulation of the models with the goal of rejection of the null hypothesis, the prevalence of p values on either side of this threshold would be about the same. Studies of p value distributions in sociology and psychology consistently find a much higher prevalence of p values just below .05 (e.g. Gerber and Malhotra 2008; Leggett et al. 2013; Masicampo and Lalande 2012). Gerber and Malhotra (2008) compared these distributions in leading sociology journals in the early 2000s versus the early 1990s and found similar patterns in both timeframes.

In this issue, Wooditch et al. (2020a) implement the p-curve method, first introduced by Simonsohn et al. (2014), to assess whether there is evidence of p-hacking in Campbell Collaborative systematic reviews. These reviews represent some of the strongest researcher in the discipline of criminology so it is striking to note that even in this body of research there is some evidence of p-hacking, with significantly more p values just below the .05 threshold than just above. At the same time, for most Campbell Collaborative reviews, there is evidential value, which implies that these areas of research are more reliable.

Best Practices for Moving the Literature Forward

Numerous proposals have been put forth to increase the reliability of the research record (Munafò et al. 2017; Nosek et al. 2015). The guiding principle for a more reliable literature is transparency and openness. Researcher degrees of freedom and questionable research practices are problematic because they are done in secret, without independent scrutiny.

Data sharing and disclosure of the code that produced a research article are an essential practice of transparent research. To demonstrate this practice, I requested data and code from seven of the authors in this special issue. Five of seven authors complied with this request, providing either code, data, or both. However, two of these were shared as links external to the journal's website, and one author submitted code to the journal but for unknown reasons it was not shared in an online appendix. Sharing of data and code are not yet standard practice in criminology, which likely accounts for the difficulty in systematically sharing data and code for each article in this issue. Of course, data sharing is not a panacea for replicable research. A researcher could, for example, share a version of the data that significantly deviates from the original data collection. Second, verifying the soundness of results in a published article using the uploaded data may be burdensome for reviewers unless the authors put in significant work to put together a replication package. Flagship journals in economics (*American Economic Review*) and political science (*American Journal of Political Science*) require replication packages to be made available before articles are published, and the latter verifies that the replication package actually produces the results presented in the paper.

Pre-registration is another strategy to reduce researcher degrees of freedom. The *American Economic Review*, for example, requires that all randomized control trials submitted to the journal to be pre-registered. However, as shown in this issue (Wooditch et al. 2020b), researchers can deviate from the plan. However, as long as the plan is publicly available, reviewers and editors would be able to assess the degree of deviation from the plan and whether it is justifiable or not.

Another common recommendation is to shift the focus away from p values to effect sizes or confidence intervals, or to abandon frequentist statistics altogether, in favor of Bayesian statistics. P values by themselves tell us next to nothing about the strength of the evidence, the scientific importance of the associated effect, or the likelihood that the result is reproducible (Cumming 2008). Embracing uncertainty and moving away from dichotomous assessment of research findings would be salutary developments. But in the absence of other changes, the high pressure to publish, nearly unlimited researcher degrees of freedom, and the tendency for statistical practice to become routinized, such a shift in emphasis may crystallize into effect size hacking, confidence interval hacking or Bayes factor hacking.

Pragmatists recommend a shift in how we assess this flawed body of research. Meta-analytic techniques such as the trim-and-fill and p-curve methods have been developed to infer the missing studies residing in file drawers and forgotten computer folders (e.g. Ferguson and Brannick 2012; Simonsohn et al. 2014). Winship and Zhou (2020, this issue) suggest rules of thumb for assessing individual studies under the assumption of publication bias. Their stance is tricky, however, because they advise against reviewers and editors using these more stringent thresholds, recommending only that consumers of published research apply more stringent standards of statistical significance. Others have suggested moving the typical threshold for statistical significance from .05 to .005 (Benjamin et al. 2018). However, as Winship and Zhou point out, such a move could increase the severity of p-hacking. It could also drive researchers towards "big data", resulting in a rash of statistically significant but substantively trivial effects.

Berk et al. (2018, accepted for this issue) take an even more pragmatic approach, suggesting that researchers admit that all models are wrong and shift the target of their analyses from causal inference and true effects to linear approximations of the truth. Models should be judged on their usefulness rather than whether they are misspecified, and estimates of causal effects should be reserved for experiments or strong quasi-experimental

methods. Their approach does not call for a significant change in practice, rather they call for a change in our expections of regression models and in how we interpret them.

Another well-worn strategy for avoiding misleading research is to ensure that our standard errors are correct. Moody and Marvell (2020, this issue) address the problem of standard error bias in fixed effects panel data models. They provide a very useful guide for researchers who intend to use such models including a full replication package that includes their Monte Carlo analysis. West et al. (2020, this issue) demonstrate a flexible Monte Carlo method for assessing regression estimate bias by simulating data and a distribution of estimates. This method has the potential to detect fragile estimates and may have potential to uncover p-hacked results. They too provide code for replicating their results.

Finally, Thomas et al. (2019, accepted for this issue) demonstrate a method to determine the fragility of regression estimates by using sensitivity analyses. Their method specifically focuses on quantifying the magnitude of selection bias that would be necessary to push an observed effect below a specific threshold. This threshold can be defined by a p value or an effect size.

Outside of this special issue, many recent articles have focused on replication in criminology. McNeeley and Warner (2015) found that just 2.3% of articles in five leading criminology journals in 2006 to 2010 self-identified as a replication or re-analysis. This compares to 2.8% in the other leading social science journals and 1.4% in leading natural science journals. More recently, Pridemore et al. (2018) found that just 0.45% of nearly 40,000 criminology articles were replications. These studies may have missed repeatability replications that use the same procedure on different data or generalization replications that use different data and a different procedure (Freese and Peterson 2017). There is a need for more direct replications and differentiated replications in order to verify past findings and determine boundary conditions (Farrington et al. 2019). Farrington et al. (2019) suggest the development of a Journal of Criminological Replication. While this could be a welcome addition the ever-growing field of criminology journals, the downside would be its low status in terms of impact factor. Replications will remain rare until they are professionally rewarded. One way to do this would be for the leading criminology journals to encourage and publish replications (Koole and Lakens 2012).

Another way to avoid biases introduced by the review process would be to adopt a two-stage review process where a journal commits to publish a study conditional on a final editorial review based on a proposal that includes the literature review and proposed data and analytic strategy, or by requiring that submitted papers be pre-registered (Greve et al. 2013). These review methods are already widely used for dissertations, which appear to be less biased towards statistically significant results (O'Boyle et al. 2017).

Next Steps

On this topic criminology has much to offer to the academy as a whole. After all, questionable research practices are a form of deviance. We have a long list of theoretical perspectives that may be usefully applied to this issue. Holtfreter et al. (2019), for example, have determined that tenure track faculty across the United States believe questionable research practices are driven primarily by strain and stresses of the tenure track, followed by low probability of being caught, and low self-control. Pratt et al. (2019) have found that researchers favor a punitive strategy for dealing with questionable research practices and there is evidence that the general public feels the same way (Pickett and Roche 2018). Of

course, given the wealth of research on deterrence, criminologists may have something to say about the likely success of a "get tough on questionable research practices" approach.

Although some progress has been made towards understanding the causes and prevalence of misleading research in criminology in this issue, much more work could be done. We do not know, for example, how criminology dissertations compare to journal articles in terms of rejection of null hypotheses. We know little of the extent of questionable research practices in the field, or how the review process itself shapes the body of scientific knowledge. Nor we do know what proportion of research will replicate. There are significant barriers to data sharing and replication in criminology due to the sensitive nature of many of the questions we study and the formidable investments of time and money some research projects require, but we should begin to take steps in the direction of a more open science. To be most effective, this effort should be facilitated by top journals and senior scholars in the field.

Acknowledgements I would like to thank David Weisburd for his support for this special issue. I also thank Shawn Bushway and Emily Owens for editorial advice, Justin Pickett for helping to conceive of the issue, and all of the authors and reviewers for contributing to this effort.

References

Barnes JC, TenEyck MF, Pratt TC, Cullen FT (2020) How powerful is the evidence in criminology? On whether we should fear a coming crisis of confidence. Justice Q 37:383–409

Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers E-J, Berk R, Bollen KA, Brembs B, Brown L, Camerer C, Cesarini D, Chambers CD, Clyde M, Cook TD, De Boeck P, Dienes Z, Dreber A, Easwaran K, Efferson C, Fehr E, Fidler F, Field AP, Forster M, George EI, Gonzalez R, Goodman S, Green E, Green DP, Greenwald AG, Hadfield JD, Hedges LV, Held L, Ho TH, Hoijtink H, Hruschka DJ, Imai K, Imbens G, Ioannidis JPA, Jeon M, Jones JH, Kirchler M, Laibson D, List J, Little R, Lupia A, Machery E, Maxwell SE, McCarthy M, Moore DA, Morgan SL, Munafó M, Nakagawa S, Nyhan B, Parker TH, Pericchi L, Perugini M, Rouder J, Rousseau J, Savalei V, Schönbrodt FD, Sellke T, Sinclair B, Tingley D, Van Zandt T, Vazire S, Watts DJ, Winship C, Wolpert RL, Xie Y, Young C, Zinman J, Johnson VA (2018) Redefine statistical significance. Nat Hum Behav 2:6–10

Berk R, Brown L, Buja A, George E, Zhao L (2018) Working with misspecified regression models. J Quant Criminol 34:633–655

Berkson J (1942) Tests of significance considered as evidence. J Am Stat Assoc 37:325–335

Boring EG (1919) Mathematical vs. scientific significance. Psychol Bull 16:335–338

Bushway SD, Sweeten G, Wilson DB (2006) Size matters: standard errors in the application of null hypothesis significance testing in criminology and criminal justice. J Exp Criminol 2:1–22

Cooper H, DeNeve K, Charlton K (1997) Finding the missing science: the fate of studies submitted for review by a human subjects committee. Psychol Methods 2:447–452

Cumming G (2008) Replications and *p* intervals: *p* values predict the future only vaguely, but confidence intervals do much better. Perspect Psychol Sci 3:286–300

Fanelli D (2009) How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. PLoS ONE 4:e5738

Fang FC, Casadevall A (2011) Retracted science and the retraction index. Am Soc Microbiol 79:3855–3859

Farrington DP, Lösel F, Boruch RF, Gottfredson DC, Mazerolle L, Sherman LW, Weisburd D (2019) Advancing knowledge about replication in criminology. J Exp Criminol 15:373–396

Ferguson CJ, Brannick MT (2012) Publication bias in psychological science: prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. Psychol Methods 17:120–128

Francis G (2014) The frequency of excess success for articles in psychological science. Psychon Bull Rev 4:1180–1187

Freese J, Peterson D (2017) Replication in social science. Annu Rev Sociol 43:147–165

Gelman A, Loken E (2014) The statistical crisis in science: data-dependent analysis—a "garden of forking paths"—explains why many statistically significant comparisons don't hold up. Am Sci 102:460–466

Gelman A, Skardhamar T, Aaltonen M (2020) Type M error might explain Weisburd's paradox. J Quant Criminol 36:1–10

Gerber AS, Malhotra N (2008) Publication bias in empirical sociology research: do arbitrary significance levels distort published results? Sociol Methods Res 37:3–30
Gottfredson DC, Cook TD, Gardner FEM, Gorman-Smith D, Howe GS, Sandler IN, Zafft KM (2015) Standards of evidence for efficacy, effectiveness, and scale-up research in prevention science: next generation. Prev Sci 16:893–926
Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG (2016) Statistical tests, *p* values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol 31:337–350
Greve W, Bröder A, Erdfelder E (2013) Result-blind peer reviews and editorial decisions. Eur Psychol 18:286–294
Grieneisen ML, Zhang M (2012) A comprehensive survey of retracted articles from the scholarly literature. PLoS ONE 7:e44118
Holtfreter K, Reisig MD, Pratt TC, Mays RD (2019) The perceived causes of research misconduct among faculty members in the natural, social and applied sciences. Stud High Educ. https://doi.org/10.1080/03075079.2019.1593352
John LK, Loewenstein G, Prelec D (2012) Measuring the prevalence of questionable researcher practices with incentives for truth telling. Psychol Sci 23:524–532
Johnson BD, Stewart EA, Pickett J, Gertz M (2011) Ethnic threat and social control: examining public support for judicial use of ethnicity in punishment. Criminology 49:401–441 **(Retraction published December 12, 2019, Criminology 58:190)**
Kerr NL (1998) HARKing: hypothesizing after the results are known. Pers Soc Psychol Rev 2:196–217
Koole SL, Lakens D (2012) Rewarding replications: a sure and simple way to improve psychological science. Perspect Psychol Sci 7:608–614
Leggett NC, Thomas NA, Loetscher T, Nicholls MER (2013) The life of *p*: "just significant" results are on the rise. Q J Exp Psychol 66:2303–2309
Mahoney MJ (1977) Publication prejudices: an experimental study of confirmatory bias in the peer review system. Cogn Ther Res 1:161–175
Maltz MD (1994) Deviating from the mean: the declining significance of significance. J Res Crime Delinq 31:434–463
Masicampo EJ, Lalande DR (2012) A peculiar prevlance of *p* values just below .05. Q J Exp Psychol 65:2271–2279
McNeeley S, Warner JJ (2015) Replication in criminology: a necessary practice. Eur J Criminol 12:581–597
Merton RK (1942) Science and technology in a democratic order. J Legal Polit Sociol 1:115–126
Moody CE, Marvell TB (2020) Clustering and standard error bias in fixed effects panel data regressions. J Quant Criminol 36:1–23
Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, du Sert NP, Simonsohn U, Wagenmakers E-J, Ware JJ, Ioannidis PA (2017) A manifesto for reproducible science. Nat Hum Behav 1:1–9
Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Brekler SJ, Buck S, Chambers CD, Chin G, Christensen G, Contestabile M (2015) Promoting an open research culture. Science 348:1422–1425
O'Boyle EH Jr, Banks GC, Gonzalez- Mulé E (2017) The chrysalis effect: how ugly initial results metamorphosize into beautiful articles. J Manag 43:376–399
Open Science Collaboration (2015) Estimating the reproducibility of psychological science. Science 349:aac4716
Pickett JT, Roche SP (2018) Questionable, objectionable or criminal? Public opinion on data fraud and selective reporting in science. Sci Eng Ethics 24:151–171
Pratt TC, Reisig MD, Holtfreter K, Golladay KA (2019) Scholars' preferred solutions for research misconduct: results from a survey of faculty members at America's top 100 research universities. Ethics Behav 29:510–530
Pridemore WA, Makel MC, Plucker JA (2018) Replication in criminology and the social sciences. Annu Rev Criminol 1:19–38
Rozeboom WW (1960) The fallacy of the null-hypothesis significance test. Psychol Bull 57:416–428
Savolainen J, VanEseltine M (2018) Replication and research integrity in criminology: introduction to the special issue. J Contemp Crim Justice 34:236–244
Sherman LW (2007) The power few: experimental criminology and the reduction of harm. The 2006 Joan McCord Prize Lecture. J Exp Criminol 3:299–321
Simmons JP, Nelson LD, Simonsohn U (2011) Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol Sci 22:1359–1366
Simonsohn U, Nelson LD, Simmons JP (2014) P-curve: a key to the file drawer. J Exp Psychol Gen 143:534–547
Sterling TD (1959) Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. J Am Stat Assoc 54:30–34

Stewart EA, Mears DP, Warren PY, Baumer EP, Arnio AN (2018) Lynchings, racial threat, and Whites' punitive views toward Blacks. Criminology 56:455–480 **(Retraction published December 6, 2019, Criminology 58:189)**

Thomas KJ, McGloin JM, Sullivan CJ (2019) Quantifying the likelihood of false positives: using sensitivity analysis to bound statistical inference. J Quant Criminol 35:631–662

Wasserstein RL, Lazar NA (2016) The ASA statement on *p*-values: context, process, and purpose. Am Stat 70:129–133

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

Weisburd D, Lum CM, Yang SM (2003) When can we conclude that treatments or programs "don't work"? Ann Am Acad Pol Soc Sci 587:31–48

West MP, Rorie M, Cohen MA (2020) The "pliability" of criminological analyses: assessing bias in regression estimates using monte carlo simulations. J Quant Criminol 36:1–24

Winship C, Zhuo X (2020) Interpreting t-statistics under publication bias: rough rules of thumb. J Quant Criminol 36:1–18

Wooditch A, Fisher R, Wu X, Johnson NJ (2020a) P-value problems? An examination of evidential value in criminology. J Quant Criminol 36:1–18

Wooditch A, Sloas LB, Wu X, Key A (2020b) Outcome reporting bias in randomized experiments on substance use disorders. J Quant Criminol 36:1–21