Are Most Published Criminological Research Findings Wrong? Taking Stock of Criminological Research using a Bayesian Simulation Approach

Version 1.1: 20 SEPTEMBER 2022

Peer-reviewed and accepted for publication in: Improving Replicable and Open Science [Special Issue]. *International Journal of Offender Therapy and Comparative Criminology* (forthcoming).

Richard E. Niemeyer[1*], K. Ryan Proctor[2], Joseph A. Schwartz[3,4], and Robert G. Niemeyer[5]

[1]United States Air Force Academy
2354 Fairchild 6L-137
USAF Academy, CO 80840
*Corresponding Author. Email: Richard.niemeyer@afacademy.af.edu

[2]Avila University
11901 Wornall Road
Kansas City, MO 64145

[3]College of Criminology and Criminal Justice
Florida State University
145 Convocation Way
304A Eppes Hall
Tallahassee, FL 32306

[4]Center for Social and Humanities Research
King Abdulaziz University
Jeddah, Saudi Arabia

[5] Metropolitan State University of Denver
Department of Mathematics & Statistics
1150 12th St.
Denver, CO 80204

Abstract

This study uses Bayesian simulations to estimate the probability that published criminological research findings are wrong. Toward this end, we employ two equations originally popularized in John P.A. Ioannidis' (in)famous article, "Why Most Published Research Findings are False." Values for relevant parameters were determined using recent estimates for the field's average level of statistical power, level of research bias, level of factionalization, and quality of theory. According to our simulations, there is a very high probability that most published criminological research findings are false-positives, and therefore wrong. Further, we demonstrate that the primary factor contributing to this problem is the poor quality of theory. Stated differently, even when the overall level of research bias is extremely low and overall statistical power is extremely high, we find that poor theory still results in a high rate of false positives. We conclude with suggestions for improving the validity of criminological research claims.

Keywords

Replication crisis; Bayesian simulation; criminological theory; research bias.

Are Most Published Criminological Research Findings Wrong? Taking Stock of Criminological Research using a Bayesian Simulation Approach

Over the past two decades, biostatisticians, medical researchers, behavioral scientists, and philosophers of science have increasingly examined the prevalence of false-positive results in published scientific research (Forstmeier et al., 2017; Loken & Gelman, 2017; Maxwell et al., 2015). Collectively, this research has led many authors to question the validity of significant, seemingly well-established, bodies of literature—a sentiment solemnly summarized by Ioannidis (2005) when he concluded that "[i]t can be proven that most claimed research findings are false" (p. 696). The pervasiveness of replication issues in scientific fields that have looked for them suggests these issues likely extend to other scientific fields that have yet to engage in the same undertaking. In criminology, recent work has provided preliminary evidence suggesting many of the factors responsible for false-positives rates in other fields are also present in criminology (Barnes et al., 2020; Chin, 2021; McNeeley & Warner, 2015; Pridemore et al., 2018; West et al., 2020; Wooditch, Fisher, et al., 2020; Wooditch, Sloas, et al., 2020).

Previous efforts to examine false-positive rates in criminology have been tremendously narrow in scope, examining only statistical power (Barnes et al., 2020) and questionable research practices (QRPs) (Burt, 2020; Chin et al., 2021). Thus, they implicitly assume false positives are the sole result of methodological problems or the publication process more broadly. While such issues certainly are important, we contend theory serves as the *most significant source of false-positive research findings* in the field. In other words: *Theoretical shortcomings that have plagued criminology for decades represent the field's current largest obstacle in producing valid scientific research findings* (Proctor & Niemeyer, 2019; Wikström & Kroneberg, 2022).

In this paper, we employ Bayesian diagnostic equations derived from Ioannidis (2005) to simulate potential rates of false-positive research findings in criminology. Working toward this end, we first provide an overview of the employed Bayesian diagnostic approach. Next, we draw from criminological research to identify possible values for equation parameters. Lastly, we conduct several simulations to identify possible false-positive rates in the field, with a particular eye to how such rates vary as a function of (1) statistical power, (2) prevalence of QRPs, and (3) validity of employed theory.

Identifying False-positive Rates in Science

At the turn of the 21st century, researchers discovered that findings from an alarming number of candidate gene studies could not be replicated (Duncan et al., 2019; Munafò, 2009). In response, Wacholder et al. (2004) proposed using Bayesian statistical methods to evaluate the probability that a statistically significant result is false. The calculation Wacholder et al. (2004) employed is called the *false-positive report probability (FPRP)* and is widely used in clinical medicine as a diagnostic tool (Fletcher, 2020). According to Park et al. (2019, p. 170), the false-positive report probability is widely "regarded to be one of the important statistical methods to judge" the likelihood that a reported statistically significant association between a candidate gene and a disease is not a false-positive.

Following Wacholder et al. (2004), Ioannidis (2005) introduced a second Bayesian statistical method called the *positive predictive value (PPV)*. Although Wacholder et al. (2004) and Ioannidis' (2005) equations are mathematically related such that $PPV = (1 - FPRP)$, Ioannidis (2005) extended Wacholder et al.'s (2004) argument in two critical ways. First, the proposed Bayesian models could be extended beyond molecular epidemiology to other scientific fields. Second, it elaborated the model to consider how research biases and other research conditions potentially exacerbate the likelihood of false positives. Based on these extensions, Ioannidis (2005) (in)famously concluded *most* published scientific research findings are false—*not just those findings published in molecular and genetic epidemiology.*

The basic version of Ioannidis's (2005) PPV is calculated using the following parameters: (1) the study's statistical power; (2) the level of statistical significance; and (3) the pre-study odds that the research hypothesis being evaluated is true. Formally, PPV is calculated as:

$$\text{Equation 1: PPV} = \frac{([1-\beta]\ \mathrm{R})}{(\mathrm{R} - \beta\mathrm{R} + \alpha)}$$

Here, α represents the study's Type I error rate, β is the Type II error rate, and $(1 - \beta)$ calculates level of statistical power. The variable R represents the "ratio of the number of 'true relationships' to 'no relationships' among those being tested" within a scientific field (Ioannidis, 2005, p. 0696). In other words, R is the ratio of true hypotheses (i.e., they correspond to reality) to the number of false hypotheses (i.e., they fail to correspond to reality). Thus, the ratio can be interpreted as the probability a given hypothesis in a field is true.

A shortcoming of Equation 1 is that it calculates PPV under the unrealistic assumption that a field's published research findings are not impacted by various forms of research bias (i.e., QRPs). Accordingly, Ioannidis (2005) modifies Equation 1 to include an additional term representing the level of bias affecting the likelihood a researcher will publish a false-positive finding, with the modified equation can be expressed as:

$$\text{Equation 2: PPV} = \frac{([1-\beta]\mathrm{R} + \mathrm{u}\beta\mathrm{R})}{(\mathrm{R} + \alpha - \beta\mathrm{R} + \mathrm{u} - \mathrm{u}\alpha + \mathrm{u}\beta\mathrm{R})}$$

In the equation, u represents overall bias. While all forms of bias ultimately affect the post-study probability a statistically significant finding is true, it is important to distinguish between different categories of bias. Previous studies, for example, have recognized deliberate acts of scientific misconduct (Gross, 2016) as well as more ethically questionable practices—

such as the so-called "four horsemen of irreproducibility" (i.e., publication bias, low statistical power, *P*-hacking, and hypothesizing after the results are known (Bishop, 2019)—as contributing to bias. Bias can also result from established normative practices when a field's explananda are ambiguous, its methods and conceptual definitions are unstandardized and highly flexible, and when numerous conceptual definitions, theoretical frameworks, and evidentiary standards co-exist (Ioannidis, 2005). These collective sources of bias stem from a combination of methodological (the former) and theoretical (the latter) sources, demonstrating the importance of considering multiple forms of bias when assessing the validity of research findings.

Lastly, Ioannidis (2005) further elaborates Equation 1 to account for the effect of multiple scientists/scientific teams pursuing the same research question. According to Ioannidis (2005, p. 0697), "[as] research efforts are globalized, it is practically the rule that several research teams, often dozens of them, may probe the same or similar questions." Assuming a constant level of statistical power across all studies for computational ease, Ioannidis (2005, p. 0697) says, "[the] probability that at least one study, among several focused on the same question, claims a statistically significant research finding" can be calculated as

$$\text{Equation 3: PPV} = \frac{(1-\beta^{n})}{(R+1[1-\alpha]^{n}-R\beta^{n})}$$

where n represents the number of researchers/research teams examining a given research question. The negative association between *n* and PPV may be due to multiple possibilities. The simplest explanation stems from the low signal-to-noise ratio associated with small effect sizes and low-powered research designs. Given these conditions, the probability that any one study will publish a true finding is low (Ioannidis 2005; see also Equation 1). Consequently, increasing the number of scientists or scientific teams researching a topic increases the probability that a false-positive finding will be identified and published.

Subsequent work has proposed two additional explanations. The first explains the negative association between n and PPV as an instantiation of the "winner's curse," drawing from the economic phenomenon commonly observed in auctions whereby the winning bid for an item is often much higher than the item's actual market value due to imperfect information (Young et al., 2008). Although a bidder may 'win' an item at auction, they are nonetheless 'cursed' for probably having paid too much for it. According to Young et al. (2008), scientific publishing is like an auction whereby individual journals 'bid' on research findings in the hope of landing a significant or notable study that will increase their journal's reputation or impact factor. Unfortunately, the 'value' of the article (i.e., the reported novelty of the statistically significant effect) may be overinflated or false and not worth the opportunity cost and possible embarrassment associated with its publication. Consequently, the greater the motivation for publishers to publish prominent findings and the greater the number of labs generating articles for them to bid on, the greater the probability a journal will publish a false-positive.

A second proposed explanation for the negative association between n and PPV stems from a 'winner-takes-all' reward structure in science (Casadevall & Fang, 2012; Hoppe-Wewetzer et al., 2021). Specifically, Hoppe-Wewetzer et al. (2021) argue that because of limited funding, limited employment opportunities, and a publish-or-perish culture, "[scientists] seek to establish priority by being first to publish an advance in knowledge and are concerned at being preempted in this by another scientist" (p. 2). Consequently, researchers are highly motivated to submit the smallest publishable finding as quickly as possible to establish a 'first mover advantage' that comes with being the first researcher or research team to claim a discovery (Newman, 2009; Sabatier & Chollet, 2017). Moreover, because academic journals rarely publish confirmations or reproductions of previously reported findings, rival scientists are also motivated to change their research focus in the hope of making a novel discovery elsewhere. To the degree that the value of n represents the level of competition within a scientific field, the probability of a novel discovery being a false-positive increases as a function of the number of researchers or research teams increases.

Calculating False-positive Rates in Criminology

To date, criminologists have not employed the equations specified by Ioannidis (2005) to estimate PPV values for criminology as a whole or for specific research programs. Nonetheless, criminologists have amassed knowledge in numerous areas that can be used to help identify plausible values for the parameters in the previously discussed equations. This knowledge directly relates to: (1) the statistical power $(1 - \beta)$ of criminological studies; (2) sources of bias related to methodological and theoretical practices (u); (3) the ratio of true to false hypotheses (R); and (4) the possible presence of the “winners curse” (n).

Statistical Power of Scientific Criminological Research

In an important recent study, Barnes et al. (2020) estimated the power of criminological studies to serve as an indicator of whether criminological research findings possess the same high false-positives rates as other fields. To arrive at field-wide estimates of statistical power, Barnes and colleagues examined 81 meta-analyses published between 1990 and 2015. The authors calculated statistical power for 270 effects, yielding a mean power of .605 and median power of .706. In terms of dispersion, the 25th percentile statistical power was .236 and the 75th percentile was .992. Since the mean and median power scores were less than the conventional value of .8, the authors observe studies in criminology are generally underpowered in terms of their effect sizes during the period observed. Regarding dispersion, a quarter of all studies were severely underpowered, while another quarter had very high levels of statistical power. These findings provide a robust estimate of the statistical power $(1 - \beta)$ of criminological research suitable for use in calculating PPV for the field of criminology.

Biases in Scientific Criminological Research

Criminologists are now beginning to study bias in criminological research by examining the prevalence of questionable research practices in the field. In what is currently the only study within criminology examining this issue, Chin et al. (2021) examine a variety of QRPs among a sample of researchers who have published in criminology journals. The authors examined HARKing (presenting exploratory findings as if they were produced using *a priori* hypotheses), underreporting results, hiding encountered problems, hiding imputed data, omitting non-significant studies or results, selectively dropping covariates, rounding *p*-values to meet specific alpha value thresholds, selectively excluding data, collecting additional data after inspecting if the current data produced statistically significant results, or changing the statistical technique employed to achieve a desired statistical finding.

The results indicate that criminologists engage in QRPs at similar rates to scientists in other fields. More specifically, 87% (*95% CI* = 84-89%) of criminology and penology authors of quantitative articles reported engaging in at least one QRP, with respondents reporting an average 2.7 different QRPs (*95% CI* = 2.6-2.9%). Additionally, those who engaged in a specific QRP tended to do so repeatedly, engaging in such practices 29-47% of the time depending upon the QRP. Importantly, the QRPs examined by Chin et al. (2021) were exclusively *methodological* in nature. That is their resulting influence is confined to the u parameter in Equation 2.

QRPs do not, however, capture other sources of biases identified by Ioannidis (2005)—such as ambiguous theoretical concepts—that serve as a source for increased researcher degrees of freedom and the ease with which various QRPs may be employed. As noted by Gibbs (1985), criminological theories are so ambiguously stated that none of them can be subjected to adequate empirical testing. In line with Gibbs, Bruinsma (2016), observes:

> [Criminology] possesses a mixture of hundreds of perspectives, definitions, ideas, sketches, multiple factors, theories and single hypotheses that are partly true and partly untrue, and none are completely true or untrue…criminologists in fact apply hardly any rule to distinguish between true and untrue theories (p. 1).

In line with these observations, criminologists have long discussed challenges in defining even the most fundamental concepts, including the very concept of crime (Henry & Lanier, 2001). Criminologists, for example, have long debated whether the criminality of an act: (1) should be identified using behavioral categories established by scientists or legal categories used by the legal system (Sellin, 1938); (2) can only be established within the legal system through prosecution and conviction (Tappan, 1947); (3) can be identified if a scientist believes an act would be defined as a crime by legal institutions, regardless of conviction status (Sutherland 1949); and (4) is marked by the punishability of an act by a representative of conventional society, regardless of its legal status (Hirschi, [1969] 2002). Moreover, criminologists further disagree over whether crime represents a homogenous class of behavior or a heterogenous one in which the only shared characteristic of criminal behaviors is its status as a legally proscribed act (Quinney, 1964). Taken together, these varying conceptions of crime—*the concept most central to the science of criminology*—provides criminologists with innumerable degrees of freedom when conducting research. And as Gibbs (1985) observes, problems of ambiguity are not limited to the field's primary *explanandum*.

Ratio of True to False Hypotheses in Criminology

Unlike statistical power, the ratio of true hypotheses to false ones (R) has not been directly estimated within criminology, leaving us to look to other indicators for identifying a reasonable range of values for this parameter. One strategy previously employed in other fields is to rely on theoretical development in selecting appropriate values, wherein bodies of literature guided by strong, well-developed theory tend to have larger R values compared to literatures with less developed theory (Bird, 2020). Ioannidis (2005) and others (Diekmann, 2016) have even directly tied the expected value of R in a given field to the underlying rigor of the theories that comprise a given body of literature. For example, Stroebe (2016) says:

> … even though we may not know the prior probability of the validity of a specific hypothesis, we can assume that hypotheses that are logically deduced from well-corroborated scientific theories have a greater likelihood of being valid than hypotheses based on ad hoc hunches .... Since from a Bayesian perspective, the prior probability of a hypothesis being valid (before doing the study) is one of the determinants of a subsequent research finding being "true" … *a discipline that mainly focuses on theory-guided research has a greater likelihood of producing valid findings than a field that engages mainly in theoretical research (*p.136, emphasis added).

Given the previously identified connection between theoretical rigor and R, there is sufficient reason to believe the ratio of true relationships to no relationships among hypotheses in criminology is exceedingly low. This should come as little surprise as criminologists have continuously raised various concerns related to the validity of key theories over the years (Bernard, 1990; Gibbs, 1985; Proctor & Niemeyer, 2019; Tittle, 1985). While the list of criticisms aimed at criminological theories is exceptionally long, we believe two criticisms are most relevant for the purposes of determining probable values of R: (1) criminological theories are rarely falsified; and (2) criminological theories are fragmented into various schools of thought, not all of which are scientific.

Falsification

Numerous criminologists have observed that the traditional Popperian model of scientific progress has failed in criminology (Bernard, 1990; Bernard & Snipes, 1996; Dooley & Goodison, 2019; Elliot, 1985; Proctor & Niemeyer, 2019; Robinson & Beaver, 2020; Walsh, 2002). Elliot (1985), for example, advocates for theory development through theory integration over theory competition and falsification precisely because he sees theory falsification as having failed in criminology. Likewise, Bernard and Snipes (1996) abandon theory development through falsification and instead contend criminologists should focus efforts on estimating the effects of theoretically informed risk-factors. Lastly, in their study of theory falsification examining a random sample of 501 criminological research articles, Dooley and Goodison (2019) found criminological theories are falsified by atrophy rather than through empirical falsification—a process whereby criminologists lose interest in specific theories due to socio-political factors that can create selection pressures for the use of certain theories over others.

A significant cause of falsification by atrophy is likely criminology's treatment of theories as logically closed systems whose truth rest upon the relations between a given theory's assumptions, *explanans*, and *explanandum*—the totality of which is compared to reality as a package to determine its 'truth.' This view of theory arises from what Hirschi (1989) calls the oppositional tradition of theory development whereby theories were explicitly formulated to contradict the assumptions and explanations of prior theories. Within this tradition, theory is seen as an "all or nothing" (Bernard, 1990, p. 330) affair and empirical tests inevitably demonstrate 'some' support for a theory. Consequently, theories are rarely seen as being falsified, instead they are simply abandoned to atrophy, as Dooley and Goodison (2019) observe.

Fragmentation

Another issue is that criminology is a heavily fragmented interdisciplinary field in which not everyone agrees the scientific enterprise should be its purpose (Agnew, 2011; Tittle, 1985). As noted by Proctor and Niemeyer (2019), scientific fields possess certain philosophical assumptions and practices that separate them from other epistemic fields. Inherent to these assumptions is a specific model of knowledge production which is aimed at the reliable prediction, explanation, and control of natural phenomena. This differs from alternative, non-scientific conceptions of social science that might emphasize statistical prediction (Friedman, 1970), meaning and understanding (Weber, 1978), social critique and/or social justice (Tittle, 1985), or empirical-philosophical explanations that rely heavily on essentialist assumptions—such as those of human nature (Hirschi, [1969] 2002). As these non-scientific views reject the assumptions of scientific knowledge and its conception of truth, the 'truth' of their claims is suspect for no other reason than the assumptions they use to produce knowledge are incommensurable with those of science. Thus, every non-scientific theory is false insofar as it fails to conform to the assumptions of scientific fields, and every scientific claim is equally false in so far as it also lacks coherence with the assumptions of science (for a review of the assumptions of science, see Mahner, 2007). By extension, this fragmentation within criminological research also suggests a low R value.

Given these collective observations and the numerous ways that the theoretical problems plaguing criminology may contribute to an increased number of false relationships, it seems reasonable to assign a relatively low value for R in the estimated simulations.

The Negative Effect of Multiple Competing Researchers and Research Teams

Criminology research findings are undoubtedly subject to the ‘winner’s curse’ and the ‘winner-takes-all’ effects associated with the presence of multiple competing researchers and research teams. We contend, however, that the field’s historical fragmentation affects how these phenomena manifest in the field in several ways. First, the winner’s curse in criminology is tied to the field’s interdisciplinary character, failure to falsify theories, and failure to update theories in light of empirical findings. Because of these characteristics, the winner’s curse in criminology involves more than a novel discovery or theoretical claim: it also potentially represents the founding of a *school of thought*. According to Collins (1994), low-consensus, slow-discovery sciences (LCSD)—such as criminology—are marked by the absence of a single, unified body of knowledge that can be taken for granted by scientists. Instead, LCSDs contain various schools of thought, with each possessing a founding scholar or scholars, lines of succession, and unique stocks of knowledge. Moreover, each school seeks to promote and defend its body of knowledge while at the same time discrediting opposing schools of thought, an observation reflected in Hirschi’s (1989) conception of the oppositional tradition of theory development.

Within criminology, schools of thought vary tremendously in terms of their ontological and epistemological assumptions, conception of ‘truth,’ and core aims. Whereas Ionnidas (2005) originally assumed the winner’s curse involved members of the scientific community, no such assumption can be made in criminology—and it is possible the school claiming to discover a truth is operating on assumptions that conflict with those of science. Since the assumptions of non-scientific fields are scientifically false to the degree they violate the assumptions of science, the discoveries they make are equally false, scientifically speaking. Thus, schools of thought may rest on nonscientific assumptions that alter the ability of others to reliably predict, control, and explain phenomena. Ultimately, as Collin's (1994) observes, LCDSs tend to be stagnant for long periods of time, promote disagreement, emphasize the interpretation or commenting on past scholarship, and rarely engage in research at the forefront of discovery.

Second, while schools of thought bring prestige to their founders, inventing a new theory is a rare phenomenon (Tittle, 1995). For other criminologists, prestige is more tied to publication counts and receipt of grant funding (Bernard, 1990). Thus, criminologists have a high incentive to be the first to test a theoretical claim (often coming from a school of thought), contributing to the winner's curse and incentivizing the publication of the smallest possible finding.

Present Study

Methods

The present study estimates the probability that a published criminological research finding is true. To do this, we perform two sets of simulations using Equations 2 and 3. First, we calculate the PPV for published criminological research using Equation 2, given a range of probable levels of research bias (u), statistical power (1-β), and estimated ratio of criminology's true to false hypotheses (R). Second, we calculate the value of PPV using Equation 3, allowing us to also examine the potential influence of increasing numbers of competing researchers, research teams, and "competing schools of thought" (n) on PPV.

Selection of α (statistical significance) parameter value

For both set of simulations, α was set at .05 to reflect the normative convention within criminology (and other behavioral sciences).

Selection of 1-β (statistical power) parameter values.

The parameter 1-β was set in line with the average level of statistical power in criminological research reported by Barnes et al. (2020), with β values 0.01, 0.39, and 0.76

selected to reflect high (.99), moderate (.61), and low (.24) levels of statistical power (respectively).

Selection of R (ratio of true to false hypotheses) parameter values

While the values for parameters α and 1-β can be determined by disciplinary convention and past research, there is generally no convention for specifying the prior probability that a hypothesis is true (R). Previously, in situations where a true value of R is unknown, researchers have suggested the use of a wide range of prior probabilities aimed at reflecting high, moderate, and low probabilities (Wacholder et al., 2004). Directly in line with this suggestion, we looked to the existing criminological literature to find values that potentially represent high, moderate, and low probabilities of R. For the high prior probability category, we relied on findings from a recent study examining the extent to which criminologists falsify theories within their field's stocks of knowledge (Dooley & Goodison, 2019). The results of this study indicated that life-course perspectives and eight other theories exhibited confirmation rates of approximately 70% or greater. Given these findings, we selected .7 to represent the high prior probability category. For the moderate prior probability category, previous observations regarding Martinson's (1974) now infamous assessment of the effectiveness of treatment programs in which Cullen noted that regardless of the treatment selected, "half the time it worked, half the time it didn't work" (cf Dooley & Goodison, 2019, p. 38). Given this observation, it seems reasonable to assume that this same "50/50" convention may also apply to a broader range of criminological hypotheses, generating a true hypothesis as often as a false one. Consequently, we set the moderate prior probability category for R at .5. Finally, for the low prior probability category, we point to the criticisms outlined above (i.e., lack of falsification, imprecision, and fragmentation), which suggest that the true value of R within criminology is likely quite low. Given these observations, we define the low prior probability category as a range of .01 to .1.

Selection of u (researcher bias) parameter values

Like R, the true value of research bias within criminology is unknown. However, Chin et al.'s (2021) recent analysis suggests that the true value is quite high. Given these findings, along with Ioannidis' (2005) original example, we selected a range of values for the u parameter in our final calculations to reflect very low (.05), low (.20), moderate (.50), and high (.80) levels of researcher bias.

Selection of n (competing schools of thought) parameter values

Like R and u, there is no standardized definition for the number of competing research groups or competing schools of thought (n). However, recently Dooley (2018) identified twenty-six perspectives within criminology that were then grouped into four basic categories: (1) structural approaches (e.g., Marxist, anomie, strain) (2) cultural approaches (e.g., differential association, reintegrative shaming, legitimation of violence), (3) control approaches (e.g., control, rational choice, life-course), and (4) bio-social approaches (e.g., biological/genetic perspective and Moffit's developmental taxonomy). Based on this classification, the value of the parameter n was set at 1 to reflect a single, unified approach, 4 to reflect the number of identified approaches, and 26 to reflect the maximum number of identified perspectives.

Plan of Analysis

As described above, the analysis involved two sets of simulations. The first set of simulations used Equation 2 to calculate the PPV across varying levels of statistical power (1-β), the estimated ratio of criminology's true to false hypotheses (R), and research bias (u). The calculations were carried out in three steps. First, we calculate PPV when power is high, and allow levels of R and u vary, demonstrating changes in PPV as a function of changes in R and u. Second, we performed the same calculations a second time, but decreased power to the average level observed by Barnes et al. (2020) to better simulate conditions observed in a typical criminological study. Third, to simulate situations in which studies are underpowered, we further decreased 1-β. Finally, we also plotted levels of PPV as a function of changes in u and across levels of R to better visualize changes in PPV across various conditions.

The second set of simulations were like the first but used Equation 3 to incorporate n. The first set of calculations provided estimates of PPV across varying levels of R and n when 1-β is high. These calculations allow for an examination of changes in PPV across levels of R and when n varies between 1 and 26. We additionally calculated PPV across levels of R and n, when 1-β was set at the average for criminological studies, and the third set provided the same calculations when 1-β was low. Finally, PPV was plotted as a function of changes in n across levels of R to reflect the compounded impact of both parameters.

Results

Recall that the value R measures the ratio of criminology's true to false hypotheses, and the value u measures the level of research bias within criminological literature. Table 1 presents the results of all simulations using Equation 2. The first set of calculations estimate PPV across levels of R and u when power (1-β) is set extremely high (i.e., β = .01). As can be seen in the top panel of Table 1, when R and 1-β are both high and u is low, the probability that a published criminological finding is true is quite high—87.67%. However, despite maintaining a high value for both R and 1-β, when u is also high (u =.8), the PPV decreases to 46.31%. Further, in situations where 1-β is extremely high and u is extremely low (u = .05), PPV ranges between 9.22% and 50.39% when R is low (between .01 and .1, respectively). These findings suggest that even in artificial situations where researcher bias can be nearly perfectly ruled out and statistical power is extremely high, the presence of poor theory significantly limits the probability of observing a true finding within criminology. Finally, the PPV drops to between 1.22% and 10.97% when both R is low and u is high, suggesting that the compounded effect of low R and high u washes out any positive impact of extremely high levels of statistical power (e.g., $1\text{-}\beta = 0.99$).

Insert Table 1 About Here

The middle panel of Table 1 provides the results of a similar set of calculations with the exception that $1\text{-}\beta$ was set to the average in power level observed in criminological research ($1\text{-}\beta = .61$; $\beta = .39$), better simulating realistic conditions within criminological research. A similar pattern is observed, wherein the PPV is 81.88% when R is high (.7) and u is very low (.05). Once again, however, PPV becomes extremely low (1.13%) when R is low (.01) and u is high (.8). Based on the discussion above, these conditions are likely to be the most representative of criminological research (i.e., moderately high levels of statistical power, moderately high levels of researcher bias, and poor theoretical development), suggesting that under these conditions, somewhere between 1.13% and 10.22% of findings from criminology are actually based in reality. Finally, the bottom panel of Table 1 presents the findings from calculations for underpowered studies ($1\text{-}\beta = .24$; $\beta = .76$). Once again, the general pattern of findings remains the same with the exception that even when R is high and u is very low, PPV drops to only 66.62%, demonstrating the negative impact of limited power.

To better visualize the findings presented in Tables 1, we have plotted the PPV as a function of each preselected value of u across levels of R with $1\text{-}\beta$ set to the conventional value of .8 in Figure 1, Panel A. The purpose of this figure is to better summarize the overall findings and to also demonstrate the compounded impact of changes in both R and u on PPV. As can be seen in the figure, the relationship between R and PPV is nonlinear, such that the low values of R have a much stronger impact on the value of PPV than moderate to higher values of R. Additionally, when research bias is exceptionally high (i.e., when u = .8) PPV is reduced by almost 40%, even when power is high and R is exceptionally high.

Insert Figure 1 About Here

The top panel in Table 2 presents the results of a set of simulations using Equation 3 to calculate changes in PPV given high, moderate, and low levels of R and different numbers of competing schools of thought (n) when 1-β is exceptionally high. When R is high, the PPV ranges between 48.73% (n = 26) and 93.27% (n = 1), indicating that even when power and R are extremely high, increased values of n drives PPV down. In addition, when R is at the upper bounds of the low category (i.e., R = .1), PPV ranges between 11.96% (when n = 26) and 66.45% (when n = 1). A similar set of findings was also observed when 1-β was set to .61, with the results presented in the middle panel of Table 2. More specifically, PPV was lowest when R was low and n was high. This pattern of findings was also observed for underpowered studies (1-β = .24), with the results presented in the bottom panel of Table 2. The primary difference is that even under nearly perfect conditions (i.e., R = .7 and n = 1), the expected PPV is 77.06%, demonstrating the negative effects of low statistical power.

Insert Table 2 About here

Finally, to better visualize the findings from simulations using Equation 3, calculated PPVs have been plotted as a function of the preselected values of n across levels of R with 1- β set to the conventional value of .80 in Figure 1, Panel B Once again, the relationship between R and PPV is nonlinear, wherein low values of R have a stronger effect on PPV relative to moderate and higher values. The graph also demonstrates the negative impact of increased levels of n on PPV, wherein a greater number of competing perspectives are present.

Discussion and Conclusion

For over twenty years, various scientific community members have drawn attention to the high prevalence of false-positive research findings in numerous scientific fields. However, only recently have criminologists begun examining whether criminology possesses a high number of false-positive research findings. The performed simulations demonstrate that criminologists have good reason to worry about the validity of their research findings. Criminologists have rightly pointed out that low statistical power and the use of QRPs can generate false-positive findings, but as our simulations reveal, issues related to researcher bias extend beyond typical concerns surrounding the use of QRPs. They also include theoretical sources of bias, such as the presence of unstandardized concepts and the presence of competing theoretical frameworks and evidentiary standards that afford criminological researchers increased degrees of freedom. Efforts to target QRPs to reduce bias and lower potential false-positive rates in the field should undoubtedly be pursued; however, expectations should be tempered surrounding the benefits of such approaches. Since they only address one factor contributing to false-positives and fail to address the contributions to researcher bias stemming from theoretical or metatheoretical issues, *our findings demonstrate that even completely eradicating QRPs would not resolve this issue.*

The strongest evidence of the importance of theory to generating false-positive research findings is evident in our findings related to the R parameter. Since criminology is a highly divided field, possessing both non-scientific and scientific approaches, numerous criminological theories are likely false for no other reason than they violate the philosophical assumptions of scientific fields. While some would likely oppose such judgment of non-scientific theories, such critiques are warranted given that hypothesis testing is a scientific endeavor and criminologists often fail to distinguish between scientific and non-scientific theories. Social control theory (Hirschi, [1969] 2002), for example, eschews scientific determinism in favor of the assumptions of classical criminology, such as those of (Hobbes, 2015) that hold individuals are innately hedonistic and rational. As Proctor and Niemeyer (2019) note, these assumptions violate core assumptions of scientific fields, such as the *ex-nihilo-nihil-fit* principle that holds phenomenon must come from somewhere and cannot vanish into nothingness. Within essentialist frameworks, assumptions of human nature—such as those in social control theory—have no origins in reality. Consequently, social control theory provides a scientifically false explanation of crime.

Even criminological theories that broadly conform to science's philosophical assumptions adversely contribute to R. Social learning theory (Akers, 2009), for example, contends individuals learn attitudes (i.e., moral evaluations) through operant conditioning. However, as recently observed by Proctor and Niemeyer (2019, 2020), discoveries made in cognitive neuroscience have revealed that symbolic knowledge—such as one's moral beliefs—are stored as schemas in semantic memory. Since semantic memories are learned through symbolical communication and are consciously accessible to individuals, it is physiologically impossible to acquire them through operant conditioning. Operant conditioning, alternatively, is the primary mechanism by which people learn skills that are stored as motor maps and are not consciously accessible. Thus, social learning theory also contributes to false positives in the field because it incorrectly specifies learning mechanisms.

Ambiguous theories also significantly contribute to possible false-positive rates because of their contribution to bias (u). Without clear theoretical criteria for guiding empirical studies, researchers are freer to engage in HARKing, selectively dropping covariates, selectively switching analysis to attain significance, etc. While open science practices can help eliminate many QRPs, having correctly specified theories that clearly define concepts, the precise propositional relationships among them, and the specific conditions under which a theoretical relationship holds can do a great deal to limit the number of *ad hoc* research practices criminologist engage in. Thus, longstanding criticisms surrounding the state of scientific criminological theory should be taken very seriously.

While our simulations demonstrate that poor theory is likely the largest source of potential false-positive research findings in criminology, this does not mean the field should ignore efforts to address QRPs, publication bias, statistical power, or any other potential contributors to false-positive rates. As evident in our simulations, these practices also contribute (sometimes in important ways) to false-positive rates and may be more amenable to intervention in many situations. Particularly low-hanging fruit lies in improving the statistical power of studies and adopting open science practices. Nonetheless, our simulations highlight improvements in these areas alone are insufficient to ensure low false-positive rates in the field, as poor theory can neutralize the gains of these practices.

A final source of false-positive research findings relates to the number of competing researchers in a field (n). While some in criminology contend criminology's mixture of numerous scientific and non-scientific schools of thought contributes to the richness of field (Cullen et al., 2018), viewed scientifically, few things can be worse for the field's false-positive rate than criminology's current stock of knowledge that contains numerous contradictory and paradoxical explanations of crime. Thus, in addition to improving the clarity and precision of criminological theory, scientific criminologists must also begin identifying ways of dismantling existing schools of thought through the defensible falsification and integration of theories—two methods of theory development that have been debated extensively in criminology (Bernard & Snipes, 1996; Hirschi, 1979, 1989; Messner et al., 1989; Tittle, 1995).

In conclusion, this study represents a first effort in estimating potential false-positive rates in criminology. As such, the study possesses several weaknesses. First, since criminologists are just now starting to assess the possibility of a credibility crisis empirically, little research exists that can be used to calculate simulation parameters. As more research is carried out, new simulations should be performed to continue monitoring potential false-positive rates in the field. Second, while theory is the most crucial contributor to false-positive rates (as it affects R, u, and n), it is also the most difficult to quantify. Future efforts should identify ways to measure theoretical parameters more precisely to understand better how theory directly and indirectly (by its influence on researcher degrees of freedom) affects false-positive rates. This study is but a first effort to determine the credibility of criminological knowledge. Lastly, the simulations conducted in this study are based upon a Bayesian diagnostic tool and not a scientific theory of science. As such, it is essential to recognize that our simulations are simply one means of trying to understand how we might improve scientific criminology.

References

Agnew, R. (2011). *Toward a unified criminology: Integrating assumptions about crime, people and society*. New York University Press.

Akers, R. L. (2009). *Social learning and social structure: A general theory of crime and deviance*. Transaction Publishers.

Barnes, J. C., TenEyck, M. F., Pratt, T. C., & Cullen, F. T. (2020). How powerful is the evidence in criminology? On whether we should fear a coming crisis of confidence. *Justice Quarterly*, *37*(3), 383-409. https://doi.org/10.1080/07418825.2018.1495252

Bernard, T. J. (1990). Twenty years of testing theories: What have we learned and why? *Journal of Research in Crime and Delinquency*, *27*(4), 325-347. https://doi.org/10.1177/0022427890027004002

Bernard, T. J., & Snipes, J. B. (1996). Theoretical integration in criminology. *Crime and Justice*, *20*, 301-348.

Bird, A. (2020). Understanding the replication crisis as a base rate fallacy. *The British Journal for the Philosophy of Science*, *72*(4), 965-993. https://doi.org/10.1093/bjps/axy051

Bishop, D. (2019). Rein in the four horsemen of irreproducibility [Report]. *Nature*, *568*, 435.

Bruinsma, G. (2016). Proliferation of crime causation theories in an era of fragmentation: Reflections on the current state of criminological theory. *European Journal of Criminology*, *13*(6), 659-676. https://doi.org/10.1177/1477370816667884

Burt, C. (2020). Doing better science: Improving review & publication protocols to enhance the quality of criminological evidence. *The Criminologist*, *45*(4), 1-6.

Casadevall, A., & Fang, F. C. (2012). Reforming science: Methodological and cultural reforms. In (Vol. 80, pp. 891-896): Am Soc Microbiol.

Chin, J. M., Pickett, J. T., Vazire, S., & Holcombe, A. O. (2021). Questionable research practices and open science in quantitative criminology. *Journal of Quantitative Criminology*. https://doi.org/10.1007/s10940-021-09525-6

Collins, R. (1994). Why the social sciences won't become high-consensus, rapid-discovery science. *Sociological Forum*, *9*(2), 155-177.

Cullen, F. T., Agnew, R., & Wilcox, P. (2018). Understanding criminological theory: A guide for readers. In F. T. Cullen, R. Agnew, & P. Wilcox (Eds.), *Criminological Theory: Past to Present* (pp. 1-18). Oxford Universty Press.

Diekmann, A. (2016). Are most published research findings false? In *Methodological Artefacts, Data Manipulation and Fraud in Economics and Social Science* (pp. 628-635). De Gruyter Oldenbourg.

Dooley, B. D. (2018). Whither criminology? The search for a paradigm over the last half century. *The American Sociologist*, *49*(2), 258-279. https://doi.org/10.1007/s12108-018-9370-8

Dooley, B. D., & Goodison, S. E. (2019). Falsification by atrophy: The Kuhnian process of rejecting theory in US criminology. *The British Journal of Criminology*, *60*(1), 24-44. https://doi.org/10.1093/bjc/azz026

Duncan, L. E., Ostacher, M., & Ballon, J. (2019). How genome-wide association studies (GWAS) made traditional candidate gene studies obsolete. *Neuropsychopharmacology*, *44*(9), 1518-1523. https://doi.org/10.1038/s41386-019-0389-5

Elliot, D. S. (1985). The assumption that theories can be combined with increased explanatory power: Theoretical integrations. In R. F. Meier (Ed.), *Theoretical Methods in Criminology* (pp. 123-149). Sage Publications.

Fletcher, G. S. (2020). *Clinical epidemiology : the essentials* (6th ed.). Wolters Kluwer Health/Lippincott Williams & Wilkins.

Forstmeier, W., Wagenmakers, E.-J., & Parker, T. H. (2017). Detecting and avoiding likely false-positive findings – a practical guide. *Biological Reviews, 92*(4), 1941-1968. https://doi.org/https://doi.org/10.1111/brv.12315

Friedman, M. (1970). The methodology of positive economics. In M. Friedman (Ed.), *Essays in Positive Economics* (pp. 3-43). Universtiy of Chicago Press. (Reprinted from 1953)

Gibbs, J. P. (1985). The methodology of theory construction in criminology. In R. F. Meier (Ed.), *Theoretical Methods in Criminology* (pp. 23-50). Sage Publication.

Gross, C. (2016). Scientific misconduct. *Annual review of psychology*, *67*(1), 693-711. https://doi.org/10.1146/annurev-psych-122414-033437

Henry, S., & Lanier, M. (2001). *What is crime?: Controversies over the nature of crime and what to do about it*. Rowman & Littlefield.

Hirschi, T. (1979). Separate and unequal is better. *Journal of Research in Crime and Delinquency*, *16*(34), 34-38.

Hirschi, T. (1989). Exploring alternatives to integrated theory. In S. F. Messner, M. D. Krohn, & A. Liska (Eds.), *Theoretical Integration in the Study of Deviance and Crime* (pp. 37-49). State University of New York Press.

Hirschi, T. ([1969] 2002). *Causes of delinquency*. Transaction Publishers.

Hobbes, T. (2015). *Leviathan*. Top 100 Classics Publishers. (1651)

Hoppe-Wewetzer, H., Katsenos, G., & Ozdenoren, E. (2021). *The Effects of Rivalry and Information Exchange on Scientific Progress*. Available at SSRN 3321231

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*(6325), 584-585. https://doi.org/doi:10.1126/science.aal3618

Mahner, M. (2007). Demarcating science from non-science. In T. A. Kuipers (Ed.), *Handbook of the Philosophy of Science: General Philosophy of Science–Focal Issues* (pp. 515-575). Elsevier.

Martinson, R. (1974). What works? Questions and answers about prison reform. *The public interest*, *35*, 22.

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*(6), 487-498. https://doi.org/10.1037/a0039400

McNeeley, S., & Warner, J. J. (2015). Replication in criminology: A necessary practice. *European Journal of Criminology*, *12*(5), 581-597. https://doi.org/10.1177/1477370815578197

Messner, S. F., Krohn, M. D., & Liska, A. E. (Eds.). (1989). *Theoretical integration in the study of deviance and crime*. State University of New York Press.

Munafò, M. R. (2009). Reliability and replicability of genetic association studies. *Addiction*, *104*(9), 1439-1440. https://doi.org/10.1111/j.1360-0443.2009.02662.x

Newman, M. E. J. (2009). The first-mover advantage in scientific publication. *EPL (Europhysics Letters)*, *86*(6), 68001. https://doi.org/10.1209/0295-5075/86/68001

Park, J. H., Geum, D. I., Eisenhut, M., van der Vliet, H. J., & Shin, J. I. (2019). Bayesian statistical methods in genetic association studies: Empirical examination of statistically non-significant genome wide association study (GWAS) meta-analyses in cancers: A systematic review. *Gene*, *685*, 170-178. https://doi.org/https://doi.org/10.1016/j.gene.2018.10.057

Pridemore, W. A., Makel, M. C., & Plucker, J. A. (2018). Replication in criminology and the social sciences. *Annual Review of Criminology*, *1*(1), 19-38. https://doi.org/10.1146/annurev-criminol-032317-091849

Proctor, K. R., & Niemeyer, R. E. (2019). *Mechanistic criminology*. Routlege Taylor & Francis Group.

Proctor, K. R., & Niemeyer, R. E. (2020). Retrofitting social learning theory with contemporary understandings of learning and memory derived from cognitive psychology and neuroscience. *Journal of Criminal Justice*, *66*, 101655. https://doi.org/https://doi.org/10.1016/j.jcrimjus.2019.101655

Quinney, E. R. (1964). The study of white collar crime: Toward a reorientation in theroy and research. *Journal of Criminal Law, Criminology and Political Science*, *55*(2), 208-214.

Robinson, M. B., & Beaver, K. M. (2020). *Why crime?* (Third edition. ed.). Carolina Academic Press, LLC.

Sabatier, M., & Chollet, B. (2017). Is there a first mover advantage in science? Pioneering behavior and scientific production in nanotechnology. *Research Policy*, *46*(2), 522-533.

Sellin, J. T. (1938). *Culture conflict and crime*. Social science research council.

Stroebe, W. (2016). Are most published social psychological findings false? *Journal of Experimental Social Psychology*, *66*, 134-144.

Sutherland, E. H. (1949). *White collar crime*. Holt, Rinhart, and Winston.

Tappan, P. (1947). Who is the criminal? *American Sociological Review*, *12*(1), 96-102.

Tittle, C. R. (1985). The assumption that general theories are not possible. In R. F. Meier (Ed.), *Theoretical Methods in Criminology* (pp. 93-122). Sage Publication.

Tittle, C. R. (1995). *Control balance: Toward a general theory of deviance*. Westview Press.

Wacholder, S., Chanock, S., Garcia-Closas, M., El Ghormli, L., & Rothman, N. (2004). Assessing the probability that a positive report is false: An approach for molecular epidemiology studies. *JNCI: Journal of the National Cancer Institute*, *96*(6), 434-442. https://doi.org/10.1093/jnci/djh075

Walsh, A. (2002). *Biosocial criminology: introduction and integration*. Anderson Publishing Company.

Weber, M. (1978). *Economy and society: an outline of interpretive sociology* (G. Roth & C. Wittich, Eds.). University of California Press.

West, M. P., Rorie, M., & Cohen, M. A. (2020). The “pliability” of criminological analyses: Assessing bias in regression estimates using monte carlo simulations. *Journal of Quantitative Criminology*, *36*(2), 371-394. https://doi.org/10.1007/s10940-018-9398-5

Wikström, P.-O. H., & Kroneberg, C. (2022). Analytic criminology: Mechanisms and methods in the explanation of crime and its causes. *Annual Review of Criminology*, *5*(1), 179-203. https://doi.org/10.1146/annurev-criminol-030920-091320

Wooditch, A., Fisher, R., Wu, X., & Johnson, N. J. (2020). p-value problems? An examination of evidential value in criminology. *Journal of Quantitative Criminology*, *36*(2), 305-328. https://doi.org/10.1007/s10940-020-09459-5

Wooditch, A., Sloas, L. B., Wu, X., & Key, A. (2020). Outcome reporting bias in randomized experiments on substance use disorders. *Journal of Quantitative Criminology*, *36*(2), 273-293. https://doi.org/10.1007/s10940-018-9388-7

Young, N. S., Ioannidis, J. P. A., & Al-Ubaydli, O. (2008). Why current publication practices may distort science. *PLoS Medicine*, *5*(10), e201. https://doi.org/10.1371/journal.pmed.0050201

Table 1: Equation 2 Simulated Values for PPV Across Different Values of Bias.

		Bias (u)			
Statistical Power (1-β)	*R*	*.05*	*.2*	*.5*	*.8*
Statistical Power = .99	.01	9.22%	3.97%	1.86%	1.22%
	.1	50.39%	29.25%	15.93%	10.97%
	.5	83.55%	67.39%	48.65%	38.12%
	.7	87.67%	74.32%	57.02%	46.31%
Statistical Power = .61	.01	6.07%	2.79%	1.51%	1.13%
	.1	39.23%	22.28%	13.30%	10.22%
	.5	76.35%	58.90%	43.40%	36.27%
	.7	81.88%	66.74%	51.77%	44.35%
Statistical Power = .24	.01	2.77%	1.61%	1.17%	1.04%
	.1	22.19%	14.04%	10.56%	9.48%
	.5	58.77%	44.95%	37.13%	34.36%
	.7	66.62%	53.34%	45.25%	42.29%

Notes: The parameter R measures the prior probability that a hypothesis is true. Categories of R: low (R = .01 or .1), medium (R = .5), high (R = .7).

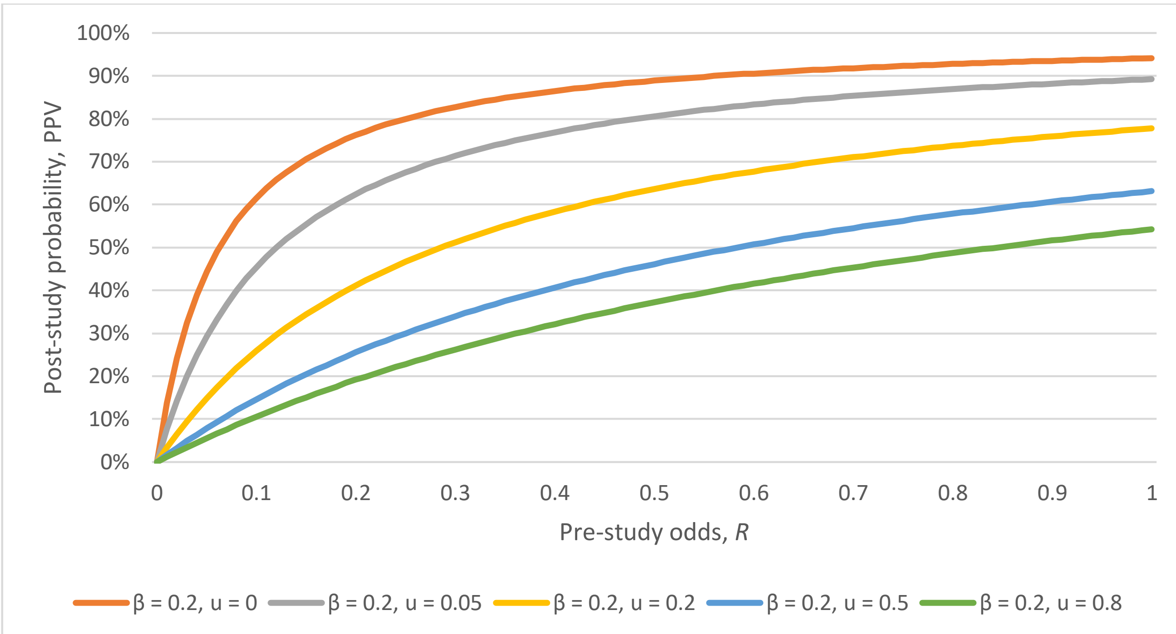
Table 2: Equation 3 Simulated Values of PPV Across Differing Numbers of Competing Researchers and Schools.

		Number of Competing Researchers/Schools (n)		
Statistical Power (1-β)	*R*	*1*	*4*	*26*
Statistical Power = .99	.01	16.53%	5.12%	1.34%
	.1	66.44%	35.03%	11.96%
	.5	90.83%	72.94%	40.44%
	.7	93.27%	79.05%	48.73%
Statistical Power = .61	.01	10.87%	5.00%	1.34%
	.1	54.96%	34.50%	11.96%
	.5	85.92%	72.48%	40.44%
	.7	89.52%	78.66%	48.73%
Statistical Power = .24	.01	4.58%	3.47%	1.34%
	.1	32.43%	26.43%	11.95%
	.5	70.59%	64.24%	40.42%
	.7	77.06%	71.55%	48.71%

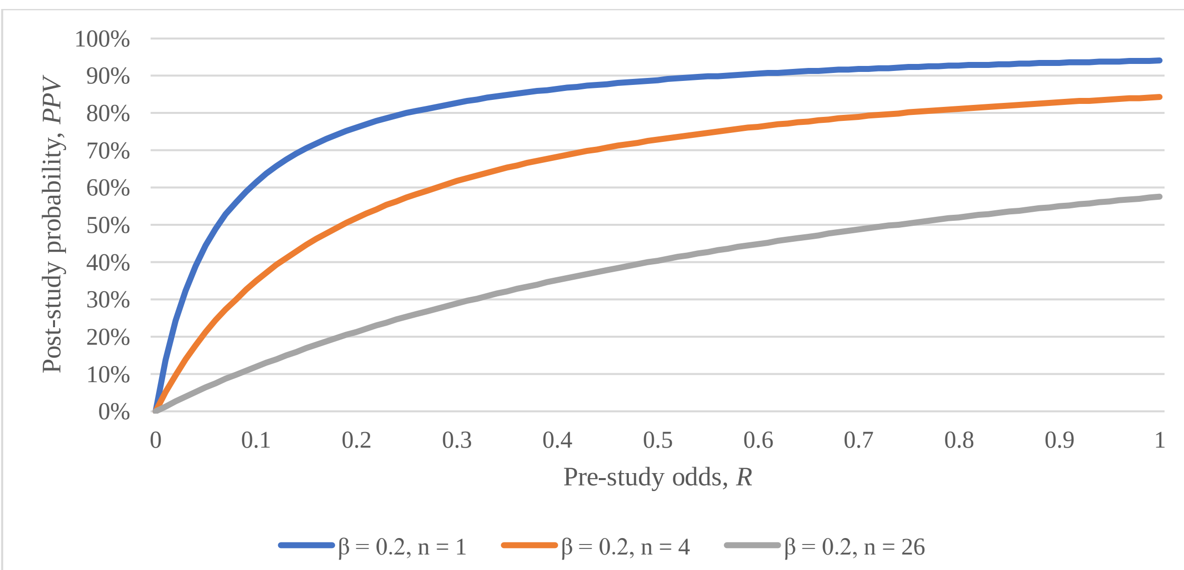
Notes: The parameter R measures the prior probability that a hypothesis is true. Categories of R: low (R = .01 or .1), medium (R = .5), high (R = .7).

Figure 1: Graphs for Equations 2 and 3 Holding Statistical Power Constant at .8

Panel A: Bias and PPV (Simulating Equation 2)



Panel B: Number of Competing Researchers and Schools (Simulating Equation 3)



Notes: The function where u = 0 in Panel A represents no bias, which is essentially Equation 1. This function is included here for comparison to show how bias affects PPV when level of statistical power is held constant.