Institute for Policy Research
Northwestern University
Working Paper Series

WP-15-10

The Emergence of Forensic Objectivity

Jeremy Freese
Ethel and John Lindgren Professor of Sociology
Faculty Fellow, Institute for Policy Research
Northwestern University

David Peterson
Department of Sociology
Northwestern University

Version: July 2015

DRAFT
Please do not quote or distribute without permission.

2040 Sheridan Rd. ◆ Evanston, IL 60208-4100 ◆ Tel: 847-491-3395 Fax: 847-491-9916
www.northwestern.edu/ipr, ◆ ipr@northwestern.edu

Abstract

A central goal of modern science, objectivity, is a concept with a documented history. Its meaning in any specific setting reflects historically situated understandings of both science and self. Recently, various scientific fields have confronted growing mistrust about the replicability of findings. Statistical techniques familiar to forensic investigations have been deployed to articulate a "crisis of false positives." In response, epistemic activists have invoked a decidedly economic understanding of scientists' selves. This has prompted a set of proposed reforms including regulating disclosure of "backstage" research details and enhancing incentives for replication. Freese and Peterson argue that, together, these events represent the emergence of a new formulation of objectivity. *Forensic objectivity* assesses the integrity of research literatures in the results observed in collections of studies rather than in the methodological details of individual studies and, thus, positions meta-analysis as the ultimate arbiter of scientific objectivity. Forensic objectivity not only presents a challenge to scientific communities but also raises new questions for the sociology of science.

THE EMERGENCE OF FORENSIC OBJECTIVITY

INTRODUCTION

"Objectivity" is a core aspiration of conventional science. Yet historians have documented how the goals of producing objective knowledge often come into conflict with the expertise needed to produce science (Daston 1992; Daston & Galison 1992; 2010; Porter 1996). Scientists are lauded for their uncommon skill and judgment, but these are also regarded as barriers to the universality and transparency implicit in objectivity. During periods of scandal or controversy, scientific judgment may come to be seen as a potential source of bias and even corruption. According to Daston and Galison (2010), debates about validity in science revolve around specific, historically-situated articulations both of epistemic vices and of the epistemic virtues that scientists are pressed to adopt to overcome these vices. Throughout this history, the virtues of objectivity have been defined against particular weaknesses perceived in subjectivity, and, in this way, broad developments in objectivity also reflect and reveal ascendant understandings of the self.

Presently, various scientific fields are said to be threatened by a "crisis of credibility" that centers on concerns about the replicability of published research. For instance, National Institutes of Health director Francis Collins described replicability concerns as a "cloud" over biomedical research (Hughes 2014). When a private research firm attempted to replicate 53 landmark studies in cancer research, they failed to replicate 47 (89%) (Begley & Ellis 2012). When another group of scientists at Bayer conducted a survey of heads of laboratories regarding the validity cancer research, they found that only 20-25% of published articles matched what the labs themselves had found (Prinz et al. 2011). Another effort to replicate 100 sampled findings from leading psychology journals had only 39% success in terms of statistical significance and only 59% success in finding even "moderately similar" results (Baker 2015). These highly-reported examples point toward broader anxieties over replication failures that have become acute. In November 2014, *Science* and *Nature*—along with 30 other journals—published an unprecedented joint editorial making specific commitments to replicable science (McNutt 2014).

We argue that fears about replicability across the sciences reflect the development of a new and powerful means of articulating epistemic vice. In response, epistemic activists, intent on mitigating the threat, have developed a correspondingly novel formulation of objectivity, which, for reasons we describe below, we call *forensic objectivity*.[1] The central feature of forensic objectivity is the projection of debates about objectivity and subjectivity onto the patterns of results produced by *collections* of studies rather than the methodological details of *individual* studies. The decentering and demotion of individual studies to mere data points within larger analyses not only undermines traditional interpretations of scientific evidence but reveals, in ways that are invisible when studies are evaluated in isolation, how currently acceptable forms of expert discretion can lead to systematic problems in literatures.

By reframing objectivity as a cumulating achievement, activists have also redefined epistemic vice. Rather than the incursion of individual subjectivity into objective research, they target the collective failure that results from the misalignment of institutional incentives. In what follows, we outline how this understanding has inspired a package of institutional reforms which present fundamental challenges to both disclosure practices and data interpretation. In the essay's conclusion we argue that recent changes to scientific practice represent the restatement of classical debates regarding objectivity onto a new, collective plane.

We present our argument using recent events in (psychological) social psychology as an extended example. The authors represent an unusual collaboration between a (sociological) social psychologist with extensive background in statistical methodology and a science studies ethnographer who has spent three years conducting fieldwork and archival research in psychology. Appendix A describes this background further and its relation to the particular materials invoked as examples in this essay. Social psychology's proximity to sociology allows us to outline the emergence of forensic objectivity in a field that is more accessible to sociologists than similar debates that are occurring in, for instance, neuroscience (Button et al. 2013) or medical genetics (Greene et al. 2009). Also, psychology's chronically insecure status as science has, historically, led psychologists to aggressively pursue new technologies of

[1] Like many scientific and intellectual social movements (Frickel & Gross 2005), the scholar "activists" we investigate are heterogeneous group. Their level of participation varies and there are disagreements about specific policies. However, they share a profound sense of unease with the current state of psychology and a broad agreement on the types of institutional changes required to improve their field. Because forensic objectivity represents an attempt to transform epistemic cultures, we use the term "epistemic activists" throughout the article. Although they share a focus on the politics of knowledge, advocates of forensic objectivity represent a conservative counterpoint to the epistemic activism previous outlined by science studies scholars (e.g., Epstein 1996).

objectivity (Danziger 1990; Porter 1995). However, we will make connections to concurrent developments in other fields throughout the article to highlight the growing role of forensic objectivity across various sciences.

EXPERTISE AND OBJECTIVITY

"All epistemology begins in fear," write Daston and Galison (2007:372). During periods of high anxiety and suspicion--periods in which *what we know* is no longer secure--issues of *how we know* come to the fore. Historically, scientific epistemology has been motivated by the fear of subjectivity because the achievement of objective knowledge has been understood to be possible only through "the suppression of some aspect of the self, the countering of subjectivity" (36).

Rather than static concepts, however, theories of objectivity and subjectivity have coevolved through historical debates over the metaphysical, methodological, and moral dimensions of these concepts (Daston 1992). In their historical study of scientific atlases, Daston and Galison (2007) outline the "epistemic virtues" and "vices" that dominated attention in different historical periods. Importantly, although each virtue and vice is a product of a particular historical moment, they should *not* be understood as "phases." As they explain, "Epistemic virtues do not replace one another like a succession of kings. Rather, they accumulate into a repertoire of possible forms of knowing" (112-113).

Broadly, each form of knowing may be understood as another movement in the interplay between the valorization of expertise, which is the embodied unification of objectivity and subjectivity, and the drive to erect strong barriers between objectivity and subjectivity that occurs when experts lose credibility. Changes in social and technical conditions continually challenge prevailing practices and motivate new epistemic virtues, which pull fields from trusting experts to demanding objectivity and back again. Here, we briefly outline the major movements outlined by previous research (Figure 1).

FIGURE 1 HERE

Idealization. Before the historical advent of the modern concept of "objectivity," science was guided by the Platonic belief that nature provided only imperfect examples of pure, objective forms. Naturalists were responsible for synthesizing their observations into "ideal" or "characteristic" portrayals. Early scientific atlas makers were guided by an epistemic virtue in which they produced idealized drawings that were meant to "portray the underlying type [...] rather than any individual specimen" (Daston & Galison 2007:20).

Mechanization. The idealizations characteristic of early atlases led to increasing anxieties about the potential for researchers to unintentionally aestheticize or theorize their images. In the mid-19th Century, concerns over subjective contamination motivated a turn toward "mechanical objectivity." Mobilizing an ethic of scientific asceticism, researchers restrained from idealizing depictions of nature through the use of machines and the strict adherence to protocols. In scientific atlases, this trend was ushered in by advancements in photography which produced representations purportedly free from human input.

Formalism and expertise. Although mechanical objectivity marginalized human intervention in data capture and data processing, doing so resulted in new problems. Without the subjective intervention which gave form to nature, researchers were left with nothing but chronic irregularity. As it became clear that mechanization would produce only complex and undigested data, two new epistemic virtues emerged in reaction. One further radicalized the split between objectivity and subjectivity through a strong emphasis on formalism. Mechanical objectivity attempts to extirpate idealizations from scientific representations, but advocates of strong formalism—whom Daston and Galison (2007: 259) call "ascetics among ascetics"—sought to remove the representations altogether, holding that the only truly objective facts were structures like logic and mathematics that could be universally communicated.

The other reaction to mechanical objectivity instead pushed back against the denial of subjectivity. Rather than a mere contaminant, these scientists argued that a scientific imagination tempered by experience was vital for understanding complex data (Daston 1998). That is, as the role of the scientist gained social stature, the need to justify the objectivity of science through ascetic self-denial ebbed. Making an epistemic virtue of "trained judgment," researchers no longer discounted all personal interpretation as distorting bias; rather, "*trained judgment* came increasingly to be seen as a necessary supplement to any image the machines might produce" (Daston & Galison 2007: 314).

Coordination. Increases in the scale and interconnectedness of science resulted in challenges that earlier scientists could not have envisioned. Rather than undermine the individual products of science, however, the growing scope of scientific research poses new threats to the integration of the scientific community. Differences in subjects, research protocols, and data analysis procedures produce incongruous literatures that threaten to fragment research. In order to counter this problem, researchers- especially in medical fields- have increasingly turned to guidelines, rules, standards, and regulations to enforce integration (Berg et al. 2000; Timmermans & Epstein 2010). Cambrosio and colleagues (2006; 2009) have labeled this move toward centralized coordination "regulatory objectivity." Like trained judgment, regulatory objectivity depends upon trust in expertise but, because regulatory objectivity is primarily concerned with the "collective production of evidence" (2009:654), it represents a break from the concern with the individualist form of objectivity which dominated earlier periods and a reformulation of these issues on a higher, collective plane.

FORENSIC OBJECTIVITY

We argue that the five major developments in the unfolding of objectivity described above have more recently been joined by a sixth. We refer to it as the rise of "forensic objectivity," and its signature is the grounding of objectivity in the aggregated assessment of the coherence of results reported by multiple studies. Forensic objectivity is a response to the fear that the various sorts of interests involved in the production and publication of results may sometimes be so profound and so pervasive as to enable a self-reinforcing pair of problems. One is a vast proliferation of exaggerated knowledge claims, and the other is a weakened capacity for exaggerated claims to be subsequently "corrected."

As with mechanical objectivity, the reforms we associate with forensic objectivity are rooted in a concern about how researchers' subjectivities may prompt erroneous idealizations of data. However, unlike mechanical objectivity, forensic objectivity is primarily concerned with how the cumulative effects of these idealizations bias entire literatures. The primary object of scrutiny in forensic objectivity is not the individual study but a *population of studies*. This move toward conceptualizing objectivity as a quality of collected studies is anticipated by the emphases on standardization and regulation of research practice that mark regulatory objectivity. However, while regulatory objectivity centralizes expertise and integrates research communities by implementing rules regarding the production of data, forensic objectivity if concerned with transforming how research is reported, aggregated, and interpreted.

This form of objectivity is "forensic" in two senses. First, like the original meaning of "forensic" as a public forum (e.g., a "forensic debate"), forensic objectivity privileges "publicness" and "openness" and does not attempt to restrain the act of judgment to a "core-set" of qualified researchers (Collins 1981). Second, like the more common use of "forensic," forensic objectivity involves using data analytic methods to *investigate* studies, scholars, and even entire literatures. While these methods have been used to investigate individual scholars for research fraud, the same methods and logic have been extended more broadly to investigate biases in the research process that are far more mundane, but potentially also more widespread and ultimately damaging.

We make no claim that phenomena we describe have gone previously unrecognized. In fact, many recent studies have investigated aspects of forensic objectivity including recent literature on "meta" science (Edwards et al. 2011; Evans & Foster 2011; Zimmerman 2008), the expansion of forensic science (Kruse 2012; Lynch et al. 2008), evidence-based medicine (Lambert 2006; Timmermans & Berg 2003; Mykhalovskiy & Weir 2004), and the explicit codification of rules for conducting and reporting research (Castel 2009; Frow 2012; Leahey 2008; Montgomery & Oliver 2009). However, what makes forensic objectivity a unique and significant development is that it combines these different ideas into a potent package that includes a philosophy of science, a set of statistical tools, and a set of demands regarding changes in scientific practice. Because forensic objectivity is concerned with the products of entire fields, everything from graduate training to journal editing is implicated.

We outline the emergence of forensic objectivity in two parts. In the first section, we describe how epistemic activists have used analyses of collections of studies to raise the possibility of a "crisis of false positives" (e.g., Wilson 2014), making visible a threat to objectivity that is hidden when studies are considered on their own. Then, we explain how this threat to objectivity has been dominantly depicted by epistemic activists in terms of a particular view of scientists' selves: namely, scientists as economic actors led to bad practices by a poorly aligned system of incentives.

In the second part, we discuss how this understanding shapes the two complementary and mutually reinforcing reforms that epistemic activists have pressed. One is increasing and standardizing the disclosure of details of the research process that were previously left to the "backstage" of science. The other is the cultivation of collections of studies that allow techniques of collective evaluation to be more powerfully applied.

THE CHALLENGE OF COLLECTIVE ASSESSMENT

Outside science, the capacity for inferential statistics to *articulate unlikeliness* gives it enormous forensic force. In fingerprint analysis, for example, inscriptions of fingertips are represented as a series of standardized, categorizable "points of identification" (Cole 1998). Many people share categorizations for particular points but, as the number of points considered increases, statistics allows investigators to make claims regarding the unlikeliness that anyone other than a suspect could have produced a particular fingerprint. Likewise, in forensic accounting, the statistical distribution of first digits listed in ledgers or balance sheets can be contrasted to the expected distribution in which smaller first digits occur much more often, with the first digit of `1' occurring 6 times as often as `9'. Substantial deviations from "Benford's law" serve as initial evidence of some irregularity in the process by which the presented numbers were generated (Durtschi et al. 2004; Nigrini & Mittermaier 1997). Subsequent investigation may reveal this to be caused by misbehavior including outright fraud.

Similar forensic demonstrations have been used to detect fraud in science. Social psychology's biggest fraud case—of prominent Dutch social psychologist Diederik Stapel—instead followed the more familiar route of being instigated by graduate students who worked with him and so had access to "backstage" information about research processes (Levelt Committee 2012). However, in the wake of the Stapel investigation, three other cases of fabrication by social psychologists came under scrutiny, and these cases were prompted by outside methodologists, who instigated suspicion by showing that statistical patterns in the investigators' reported results were highly unlikely (Simonsohn 2013; Boorsboom et al. 2014; van der Heijden et al. 2014). For example, in one case, the means across different conditions in four different experiments varied considerably, and yet the standard deviations remained too similar statistically for what might be plausibly expected (Simonsohn 2013). Put simply, all three cases turned on some variety of *implausible consistency*: a pattern in results was "too good to be true" given the expected natural fluctuations of real data.

Forensic demonstrations use inferential statistics to provide compelling judgments about the plausibility of different scenarios about how data were generated. The logic of forensic demonstration is not intrinsically limited to detecting fraud, however. Instead, it encompasses a family of methods that may be invoked to raise doubts about the production of sets of numbers. Indeed, for the developments we describe in this paper, the use of forensic demonstration to reveal fraud is secondary. In psychology, biomedicine, and elsewhere, forensic demonstrations provide compelling statistical tools for evaluating the plausibility that a literature is infested with "false positives." The true disruptive force of forensic demonstrations comes from their ability to evaluate the plausibility that a large share of the published effects in a literature may be greatly overstated, if not simply wrong.

One important tool in such work is a funnel plot, shown in Figure 2. Each dot in a funnel plot represents the results of one study. If a set of experiments all estimate the same effect, effect sizes should be symmetrically distributed around the average effect size (the dashed lines in Figure 2). However, estimates should narrow as the statistical uncertainty of experiments decrease (usually by increasing the number of experimental subjects), yielding what looks like a "funnel" (the bottom plot in Figure 2). If the set of studies available in the published record is biased because only statistically significant findings are being published, on the other hand, larger studies will have systematically smaller effect sizes, leading to a greater concentration of studies in the bottom right and upper left quadrants of the funnel (the top plot). Consequently, even though the top plot would appear to depict a set of studies with consistent, positive results in favor of a hypothesis, the funnel plot may be taken to demonstrate that the literature in question is biased. In fact, as the bottom plot shows, the pattern of published results shown in the top plot is consistent with a scenario in which no true effect exists at all.

FIGURE 2 ABOUT HERE

Another example of forensic demonstrations applied to evaluating the possibility of "false positive" results in the "*p*-curve" method illustrated in Figure 3 (Simonsohn et al. 2014). A p-curve looks at the relative frequency of different p-values below a conventional threshold, which in psychology is $p < .05$. If the set of studies are estimating a true non-zero effect, the curve will slope downward. In a set of "false positive" studies in which the true effect is actually zero, on the other hand, the p-curve may be flat or even upward sloping, and the extent to which it is upward sloping might indicate the presence of dubious analytic practices known in as “p-hacking” (described below).[2]

FIGURE 3 ABOUT HERE

Emphatically, then, even though concern about fraud in science is far greater than it once was -- so much so that Merton’s extolment of "the virtual absence of fraud in the annals of science" nowadays reads as quaint (van Noorden 2011) -- the power of funnel plots, p-curves, and related types of forensic demonstration in science extends far beyond merely detecting individual cases of fraud. Instead, the more frequent--and ultimately more disruptive--use of forensic demonstration has been to advance claims that the collective properties of published literatures do not accord with any account in which published effect sizes can be taken as unbiased estimates of true effect sizes. This can happen for any number of reasons, many of which would not fall under recognized definitions of research misconduct. For this reason, epistemic activists for various reforms we describe below routinely stress the pernicious effects of widely-accepted practices rather than focusing on obviously unethical behavior (e.g., LeBel et al. 2013; Simmons et al. 2011; Wagenmakers et al. 2012).[3]

Two features typical of such forensic demonstrations are worth highlighting. First, forensic demonstrations may be used juxtapose the results of actual literatures and statistical expectations in clear, visual terms. The rhetorical strength of public demonstrations has been posited as one of the pillars of modern science (Ashmore 2005; Shapin & Shaffer 1985) and data visualizations remain a central tool of scientific persuasion (Burri & Dumit 2008; Latour 1990). What makes these demonstrations especially potent is that they are based upon purely formal statistics. There can be little debate regarding what literatures *should* look like and serious deviations from these expected patterns can be revealed in striking visualizations.

[2] A third test by Ioannidis and Trikalinos (2007) evaluates whether there are an unlikely amount of statistically significant findings in a set of studies given their average effect size. This last test has been used extensively in psychology raise questions about sets of findings and even entire journals (Francis 2014; Francis, Tanzman, & Matthews 2014).

[3] Bad collective properties revealed by these assessments may have various causes, but what makes them “bad” is a particular shared consequence: the exaggeration of effects, which in turn implies weaker prospects for successful replication. Replicability, in this respect, makes different types of flaws commensurable and provides a simple and rhetorically powerful means by which bad collective properties can be used to sound a general alarm.

Second, forensic demonstrations analyze and produce conclusions about *collections* of numbers. These tools are most powerful for demonstrating general “fishiness.” That is, they reveal questionable collective properties which can raise doubt and prompt investigations. Yet, analytics alone can rarely establish the specific cause of the problem. Even when analytics point to extreme fishiness, any particular number in the collection could be fully legitimate. However, even though the statements of forensic statistics are intrinsically probabilistic, they can articulate anomalies so plainly, with associated probabilities so tiny, that they demand explanation.

Forensic demonstrations produce a new plane on which the integrity of a collection of numbers may be scrutinized. When the credibility of an individual number is considered on its own, judgments of its objectivity focus on *how* the number was produced. Once aggregated into a collection of numbers, however, these collections may be expected to exhibit particular statistical properties if they are to sustain the interpretation that they are collectively credible. When individually credible numbers do not have credible collective properties, doubt can pervade a literature. Individual credibility is threatened even if one cannot identify any specific problem in how any specific number was produced. That is, forensic demonstrations make possible that a collection of studies, which previously appeared impressively consistent in their findings and impeccable in methods, might be instead shown to be consistent with a “crisis of false positives,” in which the true effect is either radically smaller than what had been reported, or even potentially non-existent.

One might view the challenge of collective assessment as simply adding to the “trials of strength” (Latour 1987) that papers must withstand in order to be published. But this misses the more fundamental disruption posed by forensic demonstration: the introduction of a new and separate assessment of studies, faced by sets of published studies collectively. Failures in the collective assessments that forensic demonstrations reveal can provoke new skepticism and doubt about individual studies. Moreover, as we will show, some of the same features of individual studies that strengthen prospects for initial publication later serve as liabilities for collective assessment. Consequently, some reforms that epistemic activists propose to reduce bad collective properties imply, as a side consequence, reducing the rhetorical forcefulness of individual studies. As we discuss next, these reforms follow from a particular understanding of the systemic causes of false positives which stems from the tension between the cogency of individual papers and the integrity of collected literatures.

ECONOMIC REASONING AND SCIENTIFIC SELVES

Forensic demonstrations can be used to ascribe a non-specific, yet compelling, fishiness to collections of studies, but this determines neither how the problem is understood nor how potential solutions for it are posed. As Daston and Galison (2007:36-37) argue, new objectivities are typically posited as solutions for "a certain kind of willful self, one perceived as endangering scientific knowledge." In the present case, subjective influence could be viewed as a *moral* failure of individuals to resist temptations, or as a *socialization* failure by epistemic cultures. Some arguments have been made to each effect. Yet, in both psychology and science more broadly, what is striking about discussions of the causes and potential solutions of the "replicability crisis" (Pashler and Harris 2012) is how thoroughly dominated they are by an *economic* view of the self.

By "economic," we mean a view of self that emphasizes responsiveness to incentives provided by institutions rather than one driven by morals or socialization to scientific norms. Epistemic activists locate the root cause of biased literatures as a "dysfunctional reward structure" for scientific selves (Miguel et al. 2013). In this view, individual researchers are capable of doing better; indeed, some may full well see the problem and *yearn* to do better; and some may even be willing to forego rewards to do better as a moral stand. Ultimately, however, idealism and moral exhortation are depicted as insufficient to overcome incentives that reward shoddy or unethical research. Instead, the problem can only be addressed effectively by measures that "realign scholarly incentives with scholarly values" (Miguel et al 2013: 30).

Below we outline how epistemic activists have (1) framed the problem of questionable literatures as an issue of misaligned incentives and (2) argued that production of implausible findings fostered by misaligned incentives leads to further deterioration of a field over time. These combine to allow articulation of the problem as a variety of social dilemma, in which the individualistic pursuit of gain results in damage to the community as a whole.

The market for getting it wrong

One incisive description of the incentive problem frames it as a conflict between "getting it published" and "getting it right" (Nosek et al. 2012). This disconnect is highlighted by the title of perhaps the most influential paper prompting concerns about false positives: "Why Most

Published Findings Are False," by Ioannidis (2005). Mixing statistical theory and rational-choice style reasoning, the paper presents its provocative title as a logical inevitability given prevailing incentives and standards. Specifically, Ioannidis argues that the proportion of false positive findings in a literature depends upon (1) the likeliness of the hypotheses that researchers pursue and (2) the strength of the evidence required to publish findings as positive, while the extent to which false positive findings remain unrefuted in a literature depends on (3) the strength of mechanisms of self-correction. In recent challenges to social psychology, failures in all three aspects have been asserted.

Unlikely hypotheses. Ioannidis does not denigrate the pursuit of daring hypotheses, as novelty and discovery are, of course, vital to scientific progress. Social psychology, however, has often been criticized for overvaluing highly counterintuitive findings. Counterintuitive hypotheses are understood as having particular popular appeal, and popular interest is rewarded in the field in many ways. Activists argue that this "Gladwellization" of the field promotes both the pursuit of counterintuitive findings by authors and a preference for them among editors (Nosek et al. 2012; Posner 2014). Critics contend this not only directs attention toward hypothesis that are likely false, but, alluding to Kuhn (1962), also reduces incentives to produce the "normal science" that is more incremental but has a greater chance of enduring.

Although epistemic activists have criticized many unlikely hypotheses in social psychology, one study in particular has been significant in galvanizing opposition: a 2011 publication in social psychology's leading journal, the *Journal of Personality and Social Psychology* (*JPSP*), in which Cornell psychologist Daryl Bem (2011) presented experimental evidence of precognition. His paper supposedly demonstrated that subject behavior was influenced by randomly-assigned *future* event (in one trial, the event was viewing an erotic image). Media interest in the study was of course high, including Bem appearing on the *Colbert Report* in a segment that contemplated possible implications for "time-travel porn." If the work had been conducted by an unfamiliar investigator, perhaps fabrication may have been suspected but Bem was a high profile psychologist with a long history of contributions to psychological science. For those unwilling to entertain paranormal claims, this took the pursuit of unlikely hypotheses to its logical extreme: a hypothesis with zero chance of being true. That a false hypothesis could be published with experimental evidence that, if anything, exceeded prevailing evidential standards in the field provided an obvious prompt for reflection.

Weak evidentiary safeguards. Ioannidis (2005) strongly targets a standard that has long served as the primary gatekeeper in many behavioral and biomedical fields: null hypothesis significance testing (NHST), and particularly the reliance on a threshold of $p < .05$ to present results as providing positive support for a hypothesis. While the idea that NHST provided an "objective" method to evaluate results was key to its rise in psychology and social science (Danziger 1990; Porter 1995), it has since come under growing criticism because "flexibility" in how experimental analyses are done provide the potential to manufacture significant results in cases where no true relationship exists. That is, whatever objectivity *p*-values may have is undermined by a lack of restraints on the subjective decisions that researchers are allowed. The principal axes of this flexibility are summarized on Table 1.

TABLE 1 HERE

The undisclosed use of various flexible practices described in Table 1 have recently come to be called "questionable research practices" (QRPs) (Leahey 2008; Swazey et al. 1993). Anonymized surveys of psychologists suggest that QRPs are widespread (John et al. 2012). Although such practices have long inhablited gray areas, activists have encouraged the reinterpretation of QRPs as "soft fraud" (Chambers 2014).

Weak self-correction. As one epistemic activist bluntly stated, "There is no cost to getting things wrong" (The Economist 2013). In his classic discussion of science as a self-correcting enterprise, Merton (1973:276) describes scientists as "subject to rigorous policing, to a degree perhaps unparalleled in any other field of activity." However, Stapel's (2014) memoir of his fraud presents a different picture:

> It was very, very easy... Nobody ever checked my work; everyone trusted me... I did it all myself, with a big cookie jar right next to me... and nobody watching. and next to me was a big jar of cookies... with nobody even near. I could take whatever I wanted.

For Merton, the fear of losing prestige was vital to maintaining scientists' discipline, and "rigorous policing" connected prestige to quality of work. In contrast, a persistent complaint in social psychology is that the types of replications that might identify false positive studies are infrequently undertaken and even more rarely published (Brandt et al. 2014; Makel et al. 2012; Peters et al. 2012). The journal that published Bem's precognition study refused to review a failed replication of Bem's findings, citing a (since-changed) policy against publishing replication studies (Aldhous 2011).

As already noted, the common thread linking published findings that are false positives- that is, wrong- is nonreplicability. Concerns have been raised about poor incentives for replication work throughout science (e.g., Makel and Plucker 2013; Hamermesh 2007; Nosek et al. 2012). Science studies has shown that, in many fields, researchers seldom conduct studies directly for the purpose of replication, due to (1) their deliberate lack of originality and (2) the difficulty of establishing persuasively that failed replications are not simply the result of experimenter error (Collins 1985).

Yet, for several reasons, incentives for replication in social psychology may be especially low. First, compared to many "bench" sciences, the ability to extend social psychological findings with a new experiment is less contingent on being able to replicate the prior experiment. Second, repeating experiments provide very little opportunity for displays of technical virtuosity. Compared to research requiring "good hands" (Doing 2004), and given the field's historical emphasis on "creativity"- as manifested in its recurrent criticism for too much "cute" work (e.g., Baer 1987; Zwaan 2013)- repeating experiments is easily derided as time-wasting and diagnostic of a lack of ideas of one's own. Third, in contrast to many "applied" biomedical sciences, the low external stakes regarding whether a given claim is true or false increases the interpretability of attempts at replication as personal attacks, even as "bullying" (Schnall 2014). Consequently, at least until recent developments, even some results regarded as "classics" –spawning whole literatures—had no published record of anyone simply trying to repeat the original experiment as closely as possible (Klein et al. 2014).

Runaway expectations

Ioannidis (2005:700) raises the gloomy prospect that some areas of science could be "null fields," in which all the positive findings comprising the literature are simply reflections of the potential bias in their incentives and standards. If published findings shape what scientists subsequently regard as plausible, then false positive findings can inspire and beget other false positive findings. This may especially true whenever there is weak gatekeeping and few consequences to publishing studies that cannot be replicated.

Similarly, some problems in published articles may change expectations in ways that spur subsequent publications that are even more problematic. For example, high-profile publications shape author and reviewer expectations of what high-status publications should look like. The publication of clear, consistent findings increases the imperative for similar non-ambiguity for subsequent work. "To publish in high-impact journals," writes Stroebe and colleagues (2012: 681), "data have to provide strong, unambiguous support for the hypothesis." Social psychologists report that reviewers and editors sometimes instruct authors to remove results that weaken or qualify findings because they detract from a study's effectiveness (e.g., Schimmack 2014).

In these ways, problems originating in bad incentives can have a runaway character, in which problematic practices raise expectations, and the push to meet those expectations beget even more problematic practices. The use of questionable research practices by psychologists to make an individual study's finding more compelling has been likened to the use of performance-enhancing drugs in sports (John et al. 2012:524). In both cases, the level of competition is artificially raised, putting fair competitors at a disadvantage. And, like performance-enhancing drugs, these practices can produce outcomes that appear increasingly dubious to outsiders. Some vocal recent critics of social psychology, like the statistician and political scientist Gelman (Gelman & Carlin 2013; Gelman & Loken 2014), for example, have focused attention on large published effects that outside audiences might find implausible on their face, like a study finding that women's ovulatory cycles have large effects on their approval of Barack Obama (Durante et al. 2013).

The need to report clear findings of counterintuitive hypotheses presents an obvious moral dilemma for those who believe QRPs are unethical. One social psychologist explains that, "anyone who stands on principle, unless very lucky in results, will fail to compete effectively" (Giner-Sorolla 2012). Another decries this feature of publication incentives as "sending a clear message to graduate students and assistant professors that they must compromise their own integrity in order to succeed in our field" (Roberts 2014).

Willingness to engage in questionable practices greatly increases the extent to which a given experiment can be, in one way or another, published as a positive finding. One social psychologist that Peterson interviewed during fieldwork (see Appendix) has become a strong advocate for reforming practices that he had used himself and regarded as standard practice earlier in his career. He said that, earlier, “80%, 85% of the experiments that I ran generated a result that was significant and publishable,” but after adopting more stringent standards, he acknowledged that his rate of successful studies “is massively lower. A tenth.” Although he felt justified, the professor expressed concern for one of his graduate students going on the job market: “she’s going to have fewer papers than if she had worked in nearly any other lab in the country.” “Getting it right” often means *not* “getting it published” and this can mean not getting a job or not getting tenure.

False positives as a social dilemma

Articles aggregated together and shown to have bad collective properties raise doubts for the entire population of studies, with no precise indication of which articles or how many are responsible for the problem. One major consequence of runaway literatures, then, is that articles acquire new, negative externalities for others’ work. Doubts raised through forensic demonstrations may even extend beyond the literature analyzed. Nobel laureate Kahneman (2012) sent an open e-mail to researchers in one area warning that mounting criticisms portended a “trainwreck looming,” asserting that "your field is now the poster child for doubts about the integrity of psychological research."

Thus, as forensic demonstrations reveal hidden problems in literatures, researchers in those fields may be under considerable pressure to address them. But what to do? An economic understanding of the problem yields a straightforward social dilemma. Individual researchers have an incentive to produce studies that are compelling as possible. Yet, the cumulative consequence is a literature that cannot withstand forensic analysis. This, in turn, raises doubts about the whole field, regardless of whether specific methodological flaws are apparent. As with other social dilemmas, when the pursuit of individual interest is insufficiently constrained, the long-term welfare of the group as a whole suffers. Moreover, moral exhortations may be regarded as insufficient to produce change in the absence of a realignment of institutional incentives. Consequently, solutions involving deep, structural changes are favored—perhaps even regarded as necessary for meaningful progress—and, in the second half of the article, we outline reforms advocated by epistemic activists which seek to change the incentives provided by science institutions.

FORENSIC OBJECTIVITY, I: PROJECTIVE DISCLOSURE

Goffman (1959:112) famously argued that social performances often require hidden, "back stages" where "illusions and impressions are openly constructed." This behind-the-scenes work is necessary for performances, yet needs to be obscured because it would undermine the desired impression. Laboratory ethnographies have long made the point that the actual practice that takes place on the "backstage" of scientific labs is messier and more interpretive than what is presented to audiences in journal articles (Gilbert & Mulkay 1984; Holton 1978; Knorr Cetina 1983;1995; Woolgar 1982). The discrepancy between backstage practice and frontstage presentation might be regarded as ultimately benign, even if sociologically interesting, under the premise that the science ultimately "works." When forensic demonstrations cast doubt on whether the science actually does work, however, they raise the prospect that the tidying process is not merely sparing readers unnecessary details, but instead is obscuring systematic subjective bias.

Both mechanical and regulatory objectivity seek to mitigate the damage of researcher subjectivity by constraining discretion in research practice. Mechanical objectivity does so by automating processes, while regulatory objectivity does so by enforcing the standardization of research objects and processes. Forensic objectivity, on the other hand, makes no direct effort to constrain the role of expert's subjective judgments in producing findings. Instead, its epistemic activists pursue a policy of "front-staging" in which practices and decisions that were previously allowed to remain in the backstage of scientific practice are brought into public view where they become available for inspection by others. Projective disclosure is, thus, a part of the broader movement toward new surveillance technologies like police body cameras (Lyon 2001). In both instances, practices designed to increase transparency produce possible material for some unknown future investigation and, perhaps more importantly, the mere existence of the record can be enough to change behavior.

Thus a watchword of forensic objectivity is "open": open data, open materials, open practice, open science.[4] Advocates argue that openness addresses threats to objectivity in three ways. First openness increases the *verifiability* of findings because it reduces the extent to which

[4] The open-source software movement and the open-access research publishing movement might not have a straightforward epistemological connection, but, both genealogically and rhetorically, all signal collective projects conducted as decentralized and public affairs seeking to displace traditional practices more privately coordinated and held (Willinsky 2005).

readers need to take an author's claims on faith. In periods when there is trust in experts, a lack of disclosure may not be perceived as a problem. Once rising cynicism and doubt become seen as threats to the broader credibility of fields, however, explicit verifiability becomes available as an obvious mechanism for enhancing credibility and earning trust. Second, openness improves the quality of individual papers by discouraging questionable or incautious practices by introducing the risk of revelation and its reputational cost. Third, openness enhances what can be detected and learned from analyses of collections of studies. Making more details of studies available leads to better forensic analytics.

Forensic objectivity provokes two movements toward open practice: standardization about what research details are expected to be explicitly reported within a journal article and increasing expectations about the extensiveness of supplementing materials that are made publicly available as part of publication, but are not part of the article itself.

Standardized reporting

Transparent reporting practice promotes explicit expectations about which details of data collection and analysis will be reported. Elaborate guidelines have emerged in recent years in biomedical domains, most notably the CONSORT guidelines for reporting results of randomized clinical trials (Schulz et al. 2010; Simera et al. 2010). As part of the changes that *Psychological Science* has recently implemented, researchers who submit manuscripts are now required to complete a checklist. It requires authors to affirm that, for each experiment, their paper accurately and explicitly reports how the sample size of the experiment was determined, how many observations were excluded from analysis and why, all independent variables or manipulations ("whether successful or failed"), and all outcomes that were analyzed (Eich 2014).

Two features of this checklist bear emphasis. First, the checklist transforms what were backstage decisions into explicit moments of potential misconduct by mandating an occasion for truth-telling or lying where before there was the possibility of strategically ambiguous silence. The checklist, thus, "draws lines" about permissible conduct (Frow 2012). Second, in contrast to strategies of regulatory objectivity, the checklist does not directly regulate what researchers do with respect to any of these four areas. They are permitted full use of their judgment in designing experiments and analyzing data. Disclosure may require articulating details that might make the paper less credible or compelling for readers, and anticipation of such reactions of readers might influence how data are collected and analyzed. However, unlike regulatory objectivity, any implications for practice are only indirect.

However, some methodologists do not regard disclosing study details as sufficient to align "getting it published" and "getting it right." One suggestion has been to introduce explicit mechanisms that allow researchers to certify virtuous research practices. Epistemic activists hope that these certifications will be interpreted as signals of quality, and, perhaps eventually, will become sufficiently normative that papers without these certifications will be seen as deficient for not engaging in the virtuous practice (COS 2013).

Psychological Science, along with a half-dozen other psychological journals, has agreed to publish *badges* for articles that meet particular guidelines (Eich 2014). These are displayed as colorful icons appearing just below the title of articles, as well as in listing of articles on their website and serve as mechanisms for signaling open practices. For instance, a "Preregistration" badge is available for articles that include an experiment for which the experimental design and details of planned analysis were deposited in an independent archive prior to the data being collected (COS 2013).[5] Because it constrains experimental and analytic choices, pre-registration is intended to eliminate the possibility that significant results are due to flexibility analysis practices the data (see Table 1). Of course, before badges, nothing prevented researchers from saying that data collection and analyses followed a plan specified in advance, but public pre-registration allows a systematic mechanism for providing objective evidence for such reports. Also, without directly regulating the specifics of research practice, the badge nevertheless displays the journal's endorsement of a particular practice as virtuous, and it makes available a specific set of guidelines for what fulfilling this virtuous practice entails.

Supplementing materials

Earlier we suggested that social and technological developments can motivate changes to epistemic virtues. Advancements in the field of information technology have been especially significant for the development of forensic objectivity. Information technologies have radically altered what might be asked of researchers to share about the "backstage" of their work. Types of disclosure that would have been impracticable when fields like psychology developed can now be accomplished easily.

[5] This is similar to the strictest version of registration requirements for clinical trials in medicine.

This is apparent in the rise of online "supplemental materials" to journal articles. Articles have traditionally offered additional results or materials as "available upon request" and many professional ethical codes state explicitly what authors are expected to provide upon request by others. However, studies report abysmal success rates for such requests in practice (LeBel et al. 2013).[6] Epistemic activists have pressed to replace vague ethical expectations about how researchers should respond to requests after publication with explicit incentives to post all relevant information publicly online at the time of publication. Posting supplemental materials is already commonplace for journals like *Science*. In some cases, the supplemental materials may be far longer than the actual article (e.g., Rietveld et al. [2013] is a 3-page article with a 172-page supplement). *Psychological Science* now allows articles to display an "Open Materials" badge if they publicly share sufficient material about their experimental procedures to permit other researchers to attempt to replicate the article's findings by collecting new data for new subjects.

The push for additional materials has also included calls for public disclosure of the quantitative raw data on which findings are based. In Simonsohn's paper (2013) describing the forensic demonstrations that eventuated in the resignation of two psychologists, he notes that he had also found similar irregularities in work of a third, unnamed psychologist, but pursuit had been stymied by the author simply claiming to have lost the original data. Internet advances have greatly simplified the technical possibility of data being made publicly available at the time of publication, and so enables the possibility of requiring data availability as a condition of publication.

At present, journals that have changed policy in response to pressure for greater data "openness" evince the same three levels of reform articulated earlier, allowing us to review these as summary here. First is *requiring explicit disclosure about whatever is done*: some have introduced checklist-style forms requiring authors to affirm that the paper discloses explicitly all relevant experimental and analytic decisions. Second is *certification and endorsement of a virtuous disclosure practice*, to provide a non-compulsory incentive toward adoption of the practice. *Psychological Science* offers researchers the opportunity to display an "Open Data" badge certifying that raw data have been deposited in an independent, public archive at the time of publication, along with any code needed to reproduce reported results. Third is *mandating*

[6] This is another problem activists have framed in economic terms: because replying to requests is costly and the benefits of publication have already been attained, poor response to requests simply follows as the natural consequence to an absence of incentive to behave differently.

virtuous disclosure practices: some journals have taken the step of requiring data deposit as a condition of publication, unless the editor approves the authors' rationale for why this is not done.

FORENSIC OBJECTIVITY, II: CULTIVATING POPULATIONS OF STUDIES

Forensic demonstrations make threats to the objectivity of literatures visible through the aggregate analysis of multiple studies. In response, institutionalized practices of increased, standardized disclosure provide strategies for improving confidence in individual studies. At the same time, increasing and standardizing the information available about a study also increases the capacity for recasting individual studies as mere *data points* in larger and potentially more authoritative datasets. Rather than a rhetorical object with arguments and conclusions, the study is recast as simply a set of quantitative inputs from which, once aggregated, more objective conclusions may be derived. Forensic objectivity seeks to replace the logic of the "crucial experiment" with a logic of ongoing accumulation and assessment.

This transformation toward thinking in terms of populations of studies rather than individual findings has been a growing concern the area of "evidence-based medicine" (EBM) (Lambert 2006; Timmermans & Berg 2003; Mykhalovskiy & Weir 2004). Advocates of EBM have argued that medical decisions should be based upon syntheses of the literature based upon a "hierarchy of evidence" in which unsystematic methods like case reports are given relatively little weight compared to more systematic methods like randomized control trials (Knaapen 2013). At the top of the hierarchy, however, are meta-analyses of randomized control trials, a method that aggregates multiple studies on the same topic into a single dataset.

This demotion of findings to mere data points is manifested in two closely related developments: (1) growing calls for replication studies that follow practices of original studies as closely as possible, and (2) a reorientation toward "cumulative estimation" in which researchers attempt to draw defensibly objective conclusions from populations of studies using the conceptual and methodological tools of meta-analysis.

Mechanical replication

Grounding objectivity in populations of studies requires, first and foremost, that a population of studies exists. Cultivating this population of studies thus entails "replication" studies. As Collins (1985) made clear, however, whenever results of an intended replication diverge from those of an original study, an essential interpretive ambiguity is posed: should the divergence be understood as evidence against the credibility of the original study, or should the divergence be explained by the differences in how the two studies were conducted?

Some psychologists use the term "conceptual replication" to refer to a study that employs a *deliberately dissimilar* research design to address the same hypothesis. Successful conceptual replications can be interpreted as strengthening an initial result by showing it to be robust to alternative operationalizations. Devising compelling conceptual replications are valued as creative scientific achievements in their own right. Yet, for purposes of cumulative estimation, "conceptual replications" are problematic because a "failed conceptual replication" is a practical oxymoron: since study practices were deliberately intended to be dissimilar, any difference in outcome can easily be attributed to those dissimilarities. "Conceptual replications" can thus be dismissed by critics as intrinsically incapable of speaking to the credibility of the original study.

Thus, the possibility of cumulative estimation requires replications that are as similar to the original study as can be logistically achieved. These have been referred to as "exact," "direct," or "close" replications, reflecting different levels of authorial optimism about the level of similarity (Finkel et al. 2014). We remain agnostic and call such studies *mechanical replications*, to highlight their grounding in the basic logic of mechanical objectivity. The key principle is that researchers subordinate their own judgments to those of the authors of the original study, being as self-consciously *non-creative* as possible, so that whatever differences do exist only minimally reflect researchers' "willful" selves. By maximizing similarity, the study maximizes its commensurability for a cumulative estimation, and thus also the extent to which results of the second study may be used to adjudge the credibility of the first.

The deliberate lack of creativity in mechanical replications presents an especially acute incentive problem in fields that prize novelty. Many esteemed psychology journals have simply refused to consider direct replication studies (Aldous 2014), making incentives for mechanical replication very low. One study found a 1% rate of replication in psychological research since the year 1900 (Makel et al. 2012).

The lack of incentives for mechanical replications also reduces their credibility by increasing the plausibility of the interpretations for failed replications that focus on the competence or motivation of investigators. Mechanical replications are often posed as training exercises for students who work in a lab, but, when results diverge, their inexperience provides an easy rejoinder. For example, social psychologist Dijksterhuis (2013) attributed the null results of a replication study of his findings as flawed by inclusion of "student projects" replete with "beginners' mistakes."

Thus, although forensic demonstrations provoke a push for more mechanical replications, this push confronts several incentive problems. In order to increase the benefits to doing replications, activists have successfully pressed several psychology journals to change policies and entertain submission of mechanical replication studies. For reducing the cost to doing replications, activists have encouraged authors of original studies to publicly deposit materials at the time of publication. For increasing the credibility of replications that are done, numerous strategies have been offered, two of which we highlight here in order to illustrate how activists have conceptualized the problem and its solution in terms of configurations of incentives.

1. When Kahneman (2012) warned investigators in an especially controversial subfield of psychology that there was "trainwreck looming" in regards to their replicability, he recommended that they set up a "daisychain" system in which each participating lab would commit resources to conducting mechanical replications of the original findings of another participating lab. That lab would, in turn, have its own studies subject to mechanical replication by a third participating lab. As envisioned, this would reduce the output of original findings by every participating lab. However, the hope is that such a system would ultimately strengthen the credibility of those findings by providing a record of their independent mechanical replication by another lab that had demonstrated expertise in conducting the type of experiment in question.

2. The journal *Social Psychology* produced a special issue comprising "registered replications" (Nosek & Lakens 2014). Investigators provided pre-registered proposals that detailed data collection and analysis plans for the mechanical replication of an important published finding. These proposals were then peer-reviewed, including, when possible, one of the authors of the original studies. Studies were then conditionally accepted for publication based on these proposals *before* any data were collected and, thus, irrespective of their results. This initiative recognized that skeptical investigators who embark on mechanical replications may have an incentive for null results, and the initiative sought to address this problem by requiring advance specification, by improving attention to detail through the *a priori* review of proposals

by an original author, and by reducing the incentive to produce a particular result by guaranteeing publication regardless of outcome.

Meta-analytic fundamentalism

"Cumulative science patiently awaits the meta-analysis," write Moffitt et al. (2006).[7] Because replication attempts so regularly yield inconsistent results, epistemic activists have strongly urged against placing much confidence in new studies before success in replication studies is demonstrated. As one writer explains, "The problem isn't that many studies fail to replicate. It's that we believe in them before they've been thoroughly vetted" (Adler 2014).

Literature reviews cultivated by experts are the traditional means of performing this "vetting," but, of course, such expert judgments are subject to the same feared threats of subjectivity that often prompt more formal quantitative research designs in the first place (Hunt 1997; Light & Pillemer 1984). "Meta-analysis" encompasses various quantitative techniques that seek to draw objective conclusions about real-world relationships by combining results of multiple studies. One study found that the prevalence of meta-analyses in MEDLINE increased 12-fold between 1986 and 1999 (Egger, Davey Smith, & O-Rourke 2001).

Meta-analysis is neither new nor new to psychology. In fact, some locate the founding of meta-analysis in the effort to derive objective conclusions from the highly inconsistent record of published findings in parapsychology experiments and the term "meta-analysis" originated in efforts to draw objective conclusions from a similarly highly inconsistent record of published findings about the efficacy of psychotherapy (Chalmers et al. 2002; O'Rourke 2007; Pratt et al. 1940; Smith & Glass 1977). Yet, the prospective shift toward "meta-analytic thinking" is novel enough to serve as a cornerstone of what has been called psychology's "new statistics" (Cumming 2013). What is putatively new about "new statistics" is not the formal tools of meta-analysis but rather its broader reconceptualization of how results from individual studies are to be understood. This reconceptualization changes the locus of objectivity to the collective analysis of multiple studies, and, in doing so, seeks to alter the reporting of individual studies so that they may be brought in line with the ultimate authority of meta-analysis.

[7] Somewhat ironically, the first two authors were the primary researchers behind one of the most influential paper in psychiatric genetics of the early 2000s, which is now regarded by many as having since been revealed by meta-analysis to be a false positive (Tabery 2014).

Meta-analysis is made more powerful by the changes described above, especially (1) by standardizing analysis details reported by studies and (2) by facilitating mechanical replication by making study materials publicly available upon publication. In its approach to data analysis, "meta-analytic thinking" promotes changes that displace the traditional emphasis on null hypothesis testing with a Bayesian-style approach to evidence. Basic ideas of Bayesian statistics are old—Bayes' theorem dates to 1763– but awareness of Bayesian methods has exploded in the past two decades as computational advances have greatly increased their practical availability (e.g., Kruschke 2014).

Regardless of whether researchers make explicit use of Bayesian methods, a philosophically Bayesian style of reasoning motivates this emergent interpretation of individual experiments. Rather than each study providing a "finding", results merely increase or decrease the likelihood of some hypotheses being true versus others, with stronger evidence changing these likelihoods more. Meta-analysis can then be understood as simply extending this principle, articulating the likelihood of hypotheses being true and updating for the separate contributions of each study that is included.

Once conceived as such, meta-analysis becomes the apex of objectivity (Stegenga 2011). By combining three separate, escalating virtues, meta-analysis is argued to be the most rational conclusion that may be drawn given available evidence. First, meta-analytic conclusions are necessarily based on more information than are the conclusions of any single study that the meta-analysis contains. Second, since they aggregate studies from different investigators, meta-analyses can be seen as transcending idiosyncratic intrusions that may afflict particular investigations. Third—and bringing us back to the logic of forensic analytics presented at the outset—meta-analysis allows for the possibility that collective analysis of results may produce evidence of distortions or biases, such as the "file drawer problem" of unreported studies (described in Table 1), and affords attempted adjustments that seek to "correct" these problems. For example, there are methods that attempt to correct for the exaggeration of effect sizes revealed by biased funnel plots (Stanley and Doucouliagos 2013).

Putting matters together, then, the meta-gauntlet introduces the prospect of shared accountability for literatures, in which bad collective properties may undermine credibility even without revealing specific flaws of specific studies. Once this threat is posed, meta-analysis may be seen as the fundamental tool by which literatures can be collectively assessed, interpreted, and perhaps even rescued.

DISCUSSION

Above we describe a set of developments that together constitute a newly coherent epistemic virtue that is emerging in various scientific fields. The different aspects of forensic objectivity offer researchers a potent package that includes a set of scientific tools, a philosophy of science, and an ethic of scientific disclosure. However, as explained above, the elements are not only complementary but mutually reinforcing, in a specific way that we summarize in Figure 4.

FIGURE 4 HERE

Forensic demonstrations reveal problems of bias, cast doubt on entire literatures, and suggest the need for new epistemic virtues. An economic understanding of the problem then motivates efforts to adjust institutional incentives. Adjusted institutional incentives promote greater disclosure of experimental and analytic detail which makes it easier for other researchers to conduct replication studies and improves the information available to include in aggregations of studies (e.g., meta-analysis). The possibility of meta-analyses, in turn, affords the view that studies are inherently tentative without it, and so promotes adjusting incentives further to increase the potential scope and power of meta-analytic methods. The reinforcing character of forensic objectivity is significant because it implies that half-measures anywhere can result in weakness throughout. Thus, the call for a wide-ranging overhaul of prevailing research and reporting practices is a frequent refrain. Everything from graduate training to journal reviewing is implicated.

Forensic objectivity presents both challenges and possibilities for social scientists. Ethnographers and historians of science have long highlighted the "contingent," "messy," or "social" aspects of lab research (Knorr Cetina 1983; Woolgar 1982) and have been careful to suspend judgment when research practices deviate from classical theories of scientific method. However, forensic demonstrations raise the possibility that these descriptions have been masking research practices that can, in fact, be reabsorbed into a positivist theory of science through the investigation of collective bias. What appear to be routine interpretive choices at the level of the individual study become visible as bias when studies are aggregated.

Rather than precluding sociological investigation, however, the move by epistemic activists to a population level of analysis suggests opportunities for a corresponding move on the part of the science studies researcher. Forensic objectivity attempts to overcome subjectivity by aggregating individual knowledge claims. However, this only moves the locus of subjectivity from the individual lab to the ongoing translation between the individual and larger epistemic culture- that is, choices made during replications, the local interpretation of disclosure guidelines, and the design of meta-analyses. Here, we find the epistemic tensions between objectivity and expertise reappear on this new, collective plane.

For instance, although meta-analysis is touted for its ability to objectively synthesize literatures (Hunt 1999; Light & Pillemer 1984), the goal of overcoming expertise has, perhaps inevitably, reproduced the vacuum of interpretation that characterizes mechanical objectivity. In perhaps the earliest example, when a meta-analysis called into doubt work by Hans Eysenck (Smith & Glass 1977), he lashed out in familiar terms against what he labeled "meta-silliness":

> [Smith and Glass] advocate and practice the abandonment of critical judgments of any kind. A mass of reports- good, bad, and indifferent- are fed into the computer in the hope that people will cease caring about the quality of the material on which the conclusions are based. If their abandonment of scholarship were to be taken seriously... it would mark the beginning of a passage into the dark age of scientific psychology. (517)

In seeking to avoid bias by including all relevant studies, Smith and Glass opened themselves to the critique that they were abdicating their role as experts to evaluate the quality of studies.

We find the debate between objectivity and expertise recapitulated even more fully in more recent controversies. A high-profile example from psychology concerns a finding that a specific genetic variant moderates the relationship between stressful life events and depression (Caspi et al. 2003). Subsequent replication attempts were a confusing mix of successes, partial successes, and failures. When a research team reported a meta-analysis that yielded no evidence for the influence of gene on depression (Risch et al. 2009), the primary authors of the original study responded with their own meta-analysis arguing that the null finding was the result of overly selective criteria regarding which studies ought to be included (Caspi et al 2010). Additional meta-analyses supporting both positions, strengthening and undermining the original finding, were reported by others (Karg et al. 2011; Duncan and Keller 2011). More recently, in what may be considered an attempt to settle these arguments through a move to regulatory objectivity, experts were brought together to formulate a consensus document about how an authoritative meta-analysis would proceed (Culverhouse et al. 2013). This was immediately criticized as biased from the outset by the original authors (Moffitt and Caspi 2014).

In other words, meta-analysis is not a foolproof method of marginalizing subjectivity because—among other things—populations of studies do not build themselves. Impactful choices are made regarding study similarity ("Are these studies actually testing the same hypothesis?") and quality ("Should higher quality studies count more? Should some studies be omitted entirely?") (Eysenck 1994; Knaapen 2013; Moreira 2007; Stegenga 2011; Will 2009). Yet these decisions involve "meta-expertise" (Collins & Evans 2007) which cannot be adjudicated through purely objective criteria. This has led to the phenomenon of "dueling meta-analyses" in which experts, using different selection or analytic criteria, produce meta-analyses of the same body of studies that arrive at opposing conclusions. In his summary of the dueling meta-analyses generated around the Caspi et al. (2003) findings, Tabery notes (2014), "The meta-analyses were supposed to provide the meta-solution, but instead they only elevated it to a meta-problem" (87).

Such developments highlight how objectivity is like an ouroboros, the snake that eats its own tail. Forensic objectivity neither summarizes nor supersedes earlier developments in the history of objectivity. Rather, it creates a new level on which prior developments may repeat themselves, only in the analysis of sets of studies rather than the analysis of primary data. This is not to say that products of these inquiries do not represent progress over their predecessors. They may, but the logic behind these products, and the methods used to produce them, invoke strategies of disciplining subjectivity that are at once novel and familiar. Looking forward, three ways in which forensic objectivity transforms these familiar debates deserve special emphasis.

First, mechanical replication was developed in order to reduce the interpretive dilemma Collins (1985) famously highlighted. More broadly, however, mechanical replication in service of forensic objectivity presents a novel scenario in which replication is not a rarely conducted test of proof but, rather, a common occurrence in an ongoing system which works to increasingly decenter individual studies. Rather than debates between two research groups regarding the explanation of a failed replication, the *population* of studies is a new context for understanding replication which has not been significantly investigated.

Second, of all phases of research, the period of "data-editing" has been shrouded in the most secrecy (Leahey 2008). How will rank-and-file researchers respond to increasing encouragements (if not demands) for a level of transparency that may seem unnecessary or invasive? Can the ethic of projective disclosure be integrated into everyday research or will it be viewed as another layer of bureaucracy, to be completed with perfunctory formalism (Smith-Doerr & Vardi 2014; Zimmerman 2008)?

Finally, and perhaps most importantly, it is yet to be seen how the challenge of forensic objectivity will be met in different fields. Policies encouraging disclosure, replication, and meta-analysis have been increasingly embraced by medical research, psychology, behavioral genetics, political science, and others. The emergence of forensic objectivity indicates that the pendulum of scientific credibility may be swinging away from expertise. However, it remains to be seen how actors in different sorts of types of scientific fields will respond. For instance, evaluating the success of replication is far more complex in fields which require high levels of local or embodied knowledge (e.g., Doing 2004) and "transparency" may be especially difficult to enforce in fields in which a single article may involve bringing together the work of a hundred or more co-authors.

REFERENCES

Adler, J. 2014. “The Reformation: Can Social Sciences Save Themselves?” *Pacific Standard*. Retrieved on 1/6/15 from http://www.psmag.com/navigation/health-and-behavior/can-social-scientists-save-themselves-human-behavior-78858/

Aldous, Peter. 2011. "Journal rejects studies contradicting precognition." *New Scientist*. http://www.newscientist.com/article/dn20447-journal-rejects-studies-contradicting-precognition.html

Ashmore, Malcolm, Brown, Steven D, and Macmillan, Katie. 2005. "Lost in the Mall with Mesmer and Wundt: Demarcations and Demonstrations in the Psychologies." *Science, Technology & Human Values*, 30(1), 76-110

Baer, D. M. 1987. Do We Really Want the Unification of Psychology? A Response to Krantz.” *New Ideas in Psychology*, 5(3), 355-359.

Baker, Monya. 2015. "First Results from Psychology's Largest Reproducibility Test." *Nature*. doi:10.1038/nature.2015.17433

Bakker, M., van Dijk, A., & Wicherts, J. M. 2012. “The Rules of the Game Called Psychological Science.” *Perspectives on Psychological Science*, 7(6), 543-554.

Begley, C. G., & Ellis, L. M. 2012. “Drug Development: Raise Standards for Preclinical Cancer Research.” *Nature*, 483(7391), 531-533.

Bem, Daryl J. 2011. "Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect." *Journal of Personal and Social Psychology* 100:407-25.

Benford, F. 1938. “The Law of Anomalous Numbers.” *Proceedings of the American Philosophical Society*, 78, 551-572.

Berg, M., Horstman, K., Plass, S., & van Heusden, M. 2000. “Guidelines, Professionals and the Production of Objectivity: Standardisation and the Professionalization of Insurance Medicine.” *Sociology of Health & Illness*, 22(6), 765-791.

Borsboom, Denny, Han van der Mass, and Eric-Jan Wagenmakers. 2014 "Questions and Answers about the Förster Case." Blog post at: http://osc.centerforopenscience.org/2014/05/29/forster-case/

Brandt, Mark J. IJzerman, Hans, Dijksterjuis, AP, Farach, Frank J., Geller, Jason, Giner-Sorolla, Roger, Grange, James A., Perugini, Marco, Spies, Jeffrey, Veer, Anna V. 2014. “The Replication Recipe: What Makes for a Convincing Replication?” *Journal of Experimental Social Psychology*, 50, 217-224.

Burri, R. V., & Dumit, J. 2008. “Social Studies of Scientific Imaging and Visualization.” In (Eds.) E. J. Hackett, O. Amsterdamska, M. Lynch, & J. Wajcman, *The Handbook of Science and Technology Studies (3rd Ed.)*, 297-317. Cambridge, MA: The MIT Press

Cambrosio, A., Keating, P., Schlich, T., & Weisz, G. 2006. "Regulatory objectivity and the generation and management of evidence in medicine." *Social Science & Medicine*, 63(1), 189-199.

Cambrosio, A., Keating, P., Schlich, T., & Weisz, G. 2009. "Biomedical Conventions and Regulatory Objectivity A Few Introductory Remarks." *Social Studies of Science*, 39(5), 651-664.

Caspi, A., Hariri A. R., Holmes, A., Uher, R., Moffitt, T.E. 2010. “Genetic Sensitivity to the Environment: The Case of the Serotonin Transporter Gene and Its Implications for Studying Complex Diseases and Traits.” *American Journal of Psychiatry* 167: 509-27.

Caspi, Avshalom, Karen Sugden, Terrie E. Moffitt, Alan Taylor, Ian W. Craig, HonaLee Harrington, Joseph McClay, Jonathan Mill, Judy Martin, Antony Braithwaite and Richie Poulton. 2003. "Influence of Life Stress on Depression: Moderation by a Polymorphism in the 5-Htt Gene." *Science* 301:386-89.

Castel, P. 2009. “What’s Behind a Guideline? Authority, Competition and Collaboration in the French Oncology Sector.” *Social Studies of Science*, 39(5), 743-764.

Chalmers, I., Hedges, L., & Cooper, H. 2002. “A Brief History of Research Synthesis.” *Evaluation and the Health Professions*, 25(1), 12-37.

Chambers, C. 2014. “Physics Envy: Do ‘Hard’ Sciences Hold the Solution to the Replication Crisis in Psychology?” *TheGuardian.com.* Retrieved 2/12/15 from http://www.theguardian.com/science/head-quarters/2014/jun/10/physics-envy-do-hard-sciences-hold-the-solution-to-the-replication-crisis-in-psychology

Cole, S. 1998. “Witnessing Identification: Latent Fingerprinting Evidence and Expert Knowledge.” *Social Studies of Science*, 28(5-6), 687-712.

Collins, Harry M. 1981. "The Place of the Core-Set in Modern Science: Social Contingency with Methodological Propriety." *History of Science*, 19(1), 6-19.

Collins, Harry M. 1985. *Changing Order: Replication and Induction in Scientific Practice*. London: Sage Publications.

Collins, Harry M., & Evans, Robert. 2007. *Rethinking Expertise*. Chicago: University of Chicago Press.

COS (Center for Open Science). 2013. “Badges to Acknowledge Open Practices.” *Centerforopenscience.org.* Retrieved on 1/6/15 from https://osf.io/tvyxz/wiki/1.%20View%20the%20Badges/

Culverhouse R.C., Bowes L., Breslau N., Nurnberger J.I., Burbeister M., Fergusson D.M., Munafo M., Saccone N.L., Beirut L.J.. 2013. “Protocol for a collaborative meta-analysis of 5-HTTLPR, stress, and depression.” *BMC Psychiatry,* 13(304), 1-11.

Cumming, G. 2013. “The New Statistics: Why and How.” *Psychological Science*, 25(1), 7-29.

Danziger, Kurt. 1990. *Constructing the Subject: Historical Origins of Psychological Research*. Cambridge: Cambridge University Press.

Daston, L. 1992. "Objectivity and the Escape from Perspective." *Social Studies of Science,* 22(4), 597-618.

Daston, L. 1998. “Fear and Loathing of the Imagination in Science.” Daedalus, 127(1), 73-95.

Daston, L., & Galison, P. 1992. “The Image of Objectivity.” Representations, 40, 81-128.

Daston, L., & Galison, P. 2010. *Objectivity*. New York: Zone Books.

Dijksterhuis, A. 2013. “Replication Crisis or Crisis in Replication A Reinterpretation of Shanks et al.” Retrieved on 1/6/15 from http://www.plosone.org/annotation/listThread.action?root=64751.

Doing, P. 2004. "‘Lab Hands’ and the ‘Scarlet O’Epistemic Politics and (Scientific) Labor." *Social Studies of Science*, 34(3), 299-323.

Duncan, Laramie E. and Matthew C. Keller. 2011. "A Critical Review of the First 10 Years of Candidate Gene-by-Environment Interaction Research in Psychiatry." *American Journal of Psychiatry* 168:1041-49.

Durante, K. M., Rae, A., & Griskevicius, V. 2013. “The Fluctuating Female Vote: Politics, Religion, and the Ovulatory Cycle.” *Psychological Science*, 24(6), 1007-1016.

Durtschi, C., Hillison, W., Pachini, C. 2004. “The Effective Use of Benford’s Law to Assist in Detecting Fraud in Accounting Data.” *Journal of Forensic Accounting*, Vol. V., 17-34.

Edwards, P., Mayernik, M. S., Bowker, G., & Borgman, C. 2011. “Science Friction: Data, Metadata, and Collaboration.” *Social Studies of Science*, 41(5), 667-690.

Eich, Eric. 2014. "Business Not As Usual." *Psychological Science* 25(1):3-6.

Egger, Matthias, Davey Smith, George, Altman, Douglas, G. 2001. “Rationale, Potential, and Promise of Systematic Reviews.” In M. Egger, Davey Smith, G. & Altman, D. G. (Eds.), *Systematic Reviews in Health Care: Meta-Analysis in Context,* pp. 3-22. London: BMJ Publishing.

Epstein, Steven. 1996. *Impure Science: AIDS, Activism, and the Politics of Knowledge*. Berkeley, CA: University of California Press.

Evans, J. A., & Foster, J. G. 2011. “Metaknowledge.” *Science*, 331, 721-725.

Eysenck, H. J. 1978. “An Exercise in Mega-Silliness.” *American Psychologist*, 33(5), 517.

Eysenck, H. J. 1994. “Meta-Analysis and its Problems.” *BMJ*, 309(6957), 789-792.

Francis, G. 2014. “The Frequency of Excess Success for Articles in Psychological Science.” *Psychonomic Bulletin & Review*, 21(5), 1180-1187.

Francis, G., Tanzman, J., & Matthews, W. J. 2014. “Excess Success for Psychology Articles in the Journal Science.” *PLoS ONE*, 9(12), e114255.doi:10.1371/journal.pone.0114255

Finkel, E. J., P.W. Eastwick & H. T. Reis. 2015. “Best research practices in psychology: Illustrating epistemological and pragmatic considerations with the case of relationship science.” *Journal of Personality and Social Psychology* 108: 275-297.

Frickel, S., & Gross, N. 2005. “A General Theory of Scientific/Intellectual Movements.” *American Sociological Review*, 70(2), 204-232.

Frow, E. K. 2012. "Drawing a line: Setting guidelines for digital image processing in scientific journal articles." *Social Studies of Science,* 42(3), 369-392.

Gelman, A., & Carlin, J. 2014. “Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors.” Perspectives on Psychological Science, 9(6), 641-651.

Gelman, A., & Loken, E. 2014. “The Statistical Crisis in Science.” *American Scientist*, 102(6), 460-465.

Gilbert, G. N., & Mulkay, M. 1984. *Opening Pandora's Box: A Sociological Analysis of Scientists' Discourse*. Cambridge: Cambridge University Press.

Giner-Sorolla, Roger. 2012. “Science or Art? How Aesthetic Standards Grease the Way Through the Publication Bottleneck but Undermine Science.” *Perspectives on Psychological Science*, 7(6), 562-571.

Goffman, E. 1959. *The Presentation of Self in Everyday Life*. New York: Anchor Books.

Holton, G. 1978. “Subelectrons, Presuppositions, and the Millikan-Ehrenhaft Dispute.” *Historical Studies in the Physical Sciences*, 9, 161-224.

Hunt, M. 1997. *How Science Takes Stock: The Story of Meta-Analysis*. New York: Russell Sage Foundation.

Ioannidis, J. P. A. 2005. “Why Most Published Research Findings Are False.” *PLoS Medicine*, 2(8): e124.

Ioannidis, J. P. A., & Trikalinos, I. A. 2007. “An Exploratory Test for an Excess of Significant Finding.” *Clinical Trials*, 4, 245-253.

John, L. K., Loewenstein, G. & Prelec, D. 2012. "Measuring the prevalence of questionable research practices with incentives for truth telling." *Psychological Science*, 23(5), 524-532.

Kahneman, Daniel. 2012. "A Proposal to Deal With Questions About Priming Effects." http://www.nature.com/polopoly_fs/7.6716.1349271308!/suppinfoFile/Kahneman%20Letter.pdf

Karg, K., Burmeister, M., Shedden, K., Sen, S. 2011. "The Serotonin Transporter Variant (5-HTTLPR), Stress, and Depression Meta-Analysis Revisited: Evidence of Genetic Moderation." *JAMA* Psychiatry, 68(5), 444-454.

Klein, R. A. et al. 2014. "Investigating Variation in Replicability: A 'Many Labs' Replication Project. *Social Psychology,* 45(3), 142-152.

Knaapen, L. 2013. "Being 'Evidence-Based' in the Absence of Evidence: The Management of Non-Evidence in Guideline Development." *Social Studies of Science*, 43(5), 681-706.

Knorr-Cetina, K. 1983. "The Ethnographic Study of Scientific Work: Towards a Constructivist Interpretation of Science." In K. Knorr-Centina (Ed.) *Science Observed: Perspectives on the Social Studies of Science*, 115-40. London: Sage.

Knorr Cetina, Karen. 1995. "Laboratory Studies: The Cultural Approach to the Study of Science." In S. Jasanoff, G. E. Markle, J. C. Petersen, and T. Pinch (Eds.), *The Handbook of Science and Technology Studies*, 140-167. Sage: Thousand Oaks, CA.

Kruse, C. 2012. "The Bayesian Approach to Forensic Evidence: Evaluating, Communicating, and Distributing Responsibility." *Social Studies of Science*, 43(5), 657-680.

Kruschke, J. K. 2014. *Doing Bayesian Data Analyses: A Tutorial with R, JAGS, and Stan* (Second Edition). Academic Press.

Kuhn, T. S. 1962. *The Structure of Scientific Revolutions.* Chicago, University of Chicago Press.

Lambert, H. 2006. "Accounting for EBM: Notions of Evidence in Medicine." *Social Science & Medicine*, 62(11), 2633-2645.

Latour, B. 1987. *Science in Action: How to Follow Scientists and Engineers through Society*. Cambridge, MA: Harvard University Press.

Latour, B. 1990. Drawing Things Together. In M. Lynch and S. Woolgar (Eds), *Representation in Scientific Practice*, 19-68. Cambridge, MA: MIT Press.

Leahey, E. 2008. "Overseeing Research Practice: The Case of Data Editing." *Science, Technology & Human Values*, 33(5), 605-630.

LeBel, E. P., Borsboom, E., Giner-Sorolla, R., Hasselman, F., Peters, K. R., Ratliff, K. A., & Smith, C.

T. 2013. “PsychDisclosure.org: Grassroots Support for Reforming Reporting Standards in Psychology.” *Perspectives on Psychological Science*, 8(4), 424-432.

Levelt Committee. 2012. *Flawed Science: The Fraudulent Research Practices of Social Psychologist Diederik Stapel*. Retrieved 1/6/15 from https://www.commissielevelt.nl/wp-content/uploads_per_blog/commissielevelt/2013/01/finalreportLevelt1.pdf

Light, R. J., & Pillemer, D. B. 1984. *Summing Up: The Science of Reviewing Research*. Cambridge, MA: Harvard University Press.

Lynch, M., Cole, S. A., McNally, R., & Jordan, K. 2008. *Truth Machine: The Contentious History of DNA Fingerprinting*. Chicago: University of Chicago Press.

Lyon, D. 2001. Surveillance Society: Monitoring Everyday Life. Buckingham, UK: Open University Press.

Makel, M. C., Plucker, J. A., & Hegarty, G. 2012. “Replications in Psychology Research: How Often do They Really Occur?” Perspectives on Psychological Science, 7(6), 537-542.

Merton, R. K. 1973. *The sociology of science: Theoretical and empirical investigations*. Chicago: University of Chicago press.

Miguel, E., et al. 2014. “Promoting Transparency in Social Science Research.” *Science*, 343(6166), 30-31.

Moffitt, T.E. & Caspi, A. 2014. “Bias in a protocol for a meta-analysis of 5-HTTLPR, stress, and depression.” *BMC Psychiatry*. 14: 179

Moffitt, Terrie E., Caspi, Avshalom, & Rutter, Michael. 2006. “Measured Gene-Environment Interactions in Psychopathology: Concepts, Research Strategies, and Implications for Research, Intervention, and Public Understanding of Genetics.” *Perspectives on Psychological Science* 1(1):5-27.

Montgomery, K., & Oliver, A. L. 2009. “Shifts in Guidelines for Ethical Scientific Conduct: Now Public and Private Research Integrity.” *Social Studies of Science*, 39(1), 137-155.

Moreira, T. 2007. “Entangled Evidence: Knowledge Making in Systematic Reviews in Healthcare.” *Sociology of Health & Illness*, 29(2), 180-197.

Mykhalovskiy, E., & Weir, L. 2004. “The Problem of Evidence-Based Medicine: Directions for Social Science.” *Social Science & Medicine*, 59(5), 1059-1069.

Nigrini, M. J., & Mittermaier, L. J. 1997. “The Use of Benford’s Law as an Aid in Analytical Procedures.” *Auditing: A Journal of Practice and Theory*,” 16(2), 52-67.

Nosek, B., & Lakens, D. 2014. "Registered Reports: A Method to Increase the Credibility of Published Results." *Social Psychology*, 45(3), 137-141.

Nosek, B. A., Spies, J. R., Motyl, M. 2012. "Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth of Publishability." *Perspectives on Psychological Science*, 7(6), 615-631.

O'Rourke, K. 2007. "An Historical Perspective on Meta-Analysis: Dealing Quantitatively with Varying Study Results." *Journal of the Royal Society of Medicine*, 100(12), 579-582.

Pashler, H. & C. R. Harris. 2012. "Is the replicability crisis overblown? Three arguments examined" *Perspectives on Psychological Science* 7: 531-536.

Peters, G. J. Y., Abraham, C., & Crutzen, R. 2012. "Full Disclosure Doing Behavioral Science Necessitates Sharing." *European Health Psychologist*, 14(4), 77-84.

Peterson, D. Forthcoming. "All that is Solid: Bench-Building at the Frontiers of Two Experimental Sciences." *American Sociological Review*.

Porter, T. M. 1995. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton, NJ: Princeton University Press.

Posner, E. 2014. "The Lewisization/Gladwellization of Social Science." Ericposner.com. Retrieved on 1/5/15 from http://ericposner.com/the-lewisizationgladwellization-of-social-science/

Pratt, J. G., Rhine, J. B., Smith, B. M., Stuart, C. E., Greenwood, J. A. 1940. *Extra-Sensory Perception after Sixty Years: A Critical Appraisal of the Research in Extra-Sensory Perception.*" New York: Henry Holt and Company.

Prinz, F., Schlange, T., & Asadullah, K. 2011. "Believe it or Not: How Much Can We Rely on Published Data on Potential Drug Targets?" Nature.com. Retrieved 1/6/15 from http://www.nature.com/nrd/journal/v10/n9/full/nrd3439-c1.html

Rietveld, C. A. et al. 2013. "GWAS of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment." *Science*, 340(6139), 1467-1471.

Risch, N., Gerrell, R., Lehner, T., Liang, K. Y., Eaves, L., Hoh, J., Griem, A., Kovacs, M., Ott, J., & Merikangas, K. R. 2009. "Interaction Between the Serotonin Transporter Gene (5-HTTLPR), Stressful Life Events, and Risk of Depression." *JAMA*, 301(23), 2462-2471.

Roberts, Brent W. 2014. "The Deathly Hallows of Psychological Science." Blog post available at http://osc.centerforopenscience.org/2014/04/02/deathly-hallows/

Schimmack, Ulrich. 2014. "Roy Baumeister's R-Index." http://replicationindex.wordpress.com/2014/12/01/roy-baumeisters-r-index/

Schnall, S. 2014. “Social Media and the Crowd-Sourcing of Social Psychology.” Retrieved on 1/5/2015 from http://www.psychol.cam.ac.uk/cece/blog

Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 Statement: Updated Guidelines for Reporting Parallel Group Randomised Trials. PLoS Med 2010;7(3): e1000251. doi:10.1371/journal.pmed.1000251

Shapin, S., & Shaffer, S. 1985. *Leviathan and the Air Pump*. Princeton, NJ. Princeton University Press.

Simera, I., D. Moher, J. Hoey, K. F. Schulz, and D.G. Altman. 2010. "A catalogue of reporting guidelines for health research." *European Journal of Clinical Investigation*, 40(1), 35-53.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. 2011. "False-Positive Psychology Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science*, 22(11), 1359-1366.

Simonsohn, U. 2013. “Just Post It: The Lesson From Two Cases of Fabricated Data Detected by Statistics Alone.” *Psychological Science*, 24(10), 1875-1888.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. 2014. “*P*-Curve: A Key to the File-Drawer Problem.” *Journal of Experimental Psychology: General*, 143(2), 534-547.

Smith, M. L., & Glass G. V. 1977. “Meta-Analysis of Psychotherapy Outcome Studies.” *American Psychologist*, 32(9), 752-760.

Smith-Doerr, L., & Vardi, I. 2015. “Mind the Gap: Formal Ethics Policies and Chemical Scientists’ Everyday Practices in Academia and Industry.” *Science, Technology, and Human Values*, 40(2), 176-198.

Stanley, T. D. and H. Doucouliagos. “Meta-regression approximations to reduce publication selection bias.” *Research Synthesis Methods* 5: 60-78.

Stapel, Diederik. 2014. *Faking Science: A True Story of Academic Fraud.* (Translated by Nicholas J. L. Brown) [need to find URL]

Stegenga, J. 2011. “Is Meta-Analysis the Platinum Standard of Evidence?” *Studies in History and Philosophy of Biological and Biomedical Sciences*, 42(4), 497-507.

Stroebe, W., Postmes, T., & Spears, R. 2012. “Scientific Misconduct and the Myth of Self-Correction in Science.” *Perspectives on Psychological Science*, 7(6), 670-688.

Swazey, J. P., Anderson, M. S., & Lewis, K. S. 1993. “Ethical Problems in Academic Research.” *American Scientist*, 81(6), 542-553.

Tabery, J. 2014. *Beyond Versus: The Struggle to Understand the Interaction of Nature and Nuture.*

Cambridge, MA: The MIT Press.

The Economist. 2013. “Trouble at the Lab.” *Economist.com*. Retrieved 1/5/15 from http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble

Timmermans, S., & Berg, M. 2003. *The Gold Standard: The Challenge of Evidence-Based Medicine and Standardization in Health Care*. Philadelphia, PA: Temple University Press.

van der Heijden, A. J., P. J. F. Groenen, R. Zeelenberg. 2014. “Report of the Smeesters Follow-up Investigation Committee.” Erasmus University. Available online at: http://www.eur.nl/fileadmin/ASSETS/press/2014/maart/Report_Smeesters_follow-up_investigation_committee.final.pdf

van Noorden, R. 2011. “The Trouble with Retractions.” *Nature*, 478, 26-28.

Wagenmakers, E. J., Wetzels, R., Borsboom, van der Maas, H. L. J., & Kievit, R. A. 2012. “An Agenda for Purely Confirmatory Research.” *Perspectives on Psychological Science*, 7(6), 632-638.

Will, C. M. 2009. “Effectiveness in ‘The Old Old’: Principles and Values in the Age of Clinical Trials.” *Science, Technology, & Human Values*, 34(5), 607-628.

Wilson, Tim. 2014. “Is there a Crisis of False Negatives in Psychology?” Available at: https://timwilsonredirect.wordpress.com/2014/06/15/is-there-a-crisis-of-false-negatives-in-psychology/

Willinsky, John. 2005. "The unacknowledged convergence of open source, open access, and open science." *First Monday* 10(8):online only.

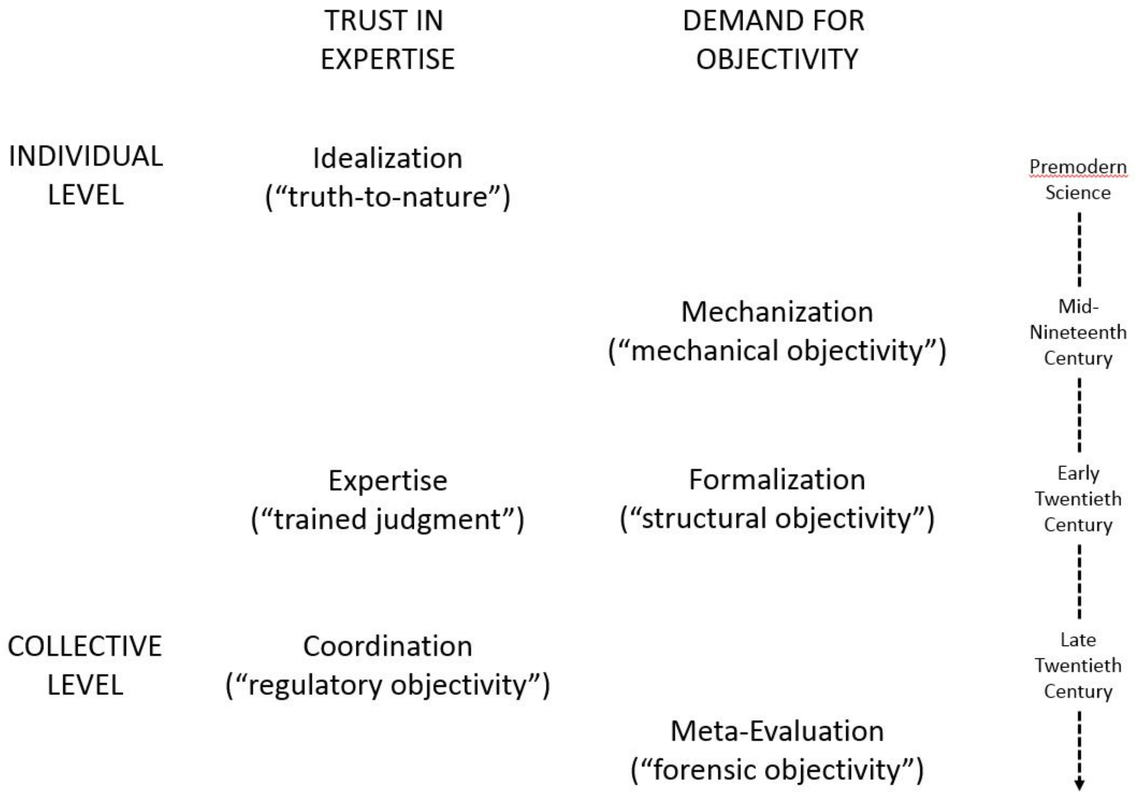
Woogar, S. 1982. “Laboratory Studies: A Comment on the State of the Art.” *Social Studies of Science*, 12(4), 481-98.

Zimmerman, A. S. 2008. “New Knowledge from Old Data: The Role of Standards in the Sharing and Reuse of Ecological Data.” *Science, Technology & Human Values*, 33(5), 631-652.

Zwaan, R. 2013. “Amusing Titles in Psychological Science.” *Zeistgeist: Psychological Experimentation, Cognition, Language, and Academia*. Retrieved 2/12/15 from http://rolfzwaan.blogspot.com/2013/01/normal.html

FIGURES AND TABLES

Figure 1: Movements in Expertise and Objectivity



	TRUST IN EXPERTISE	DEMAND FOR OBJECTIVITY	
INDIVIDUAL LEVEL	Idealization (“truth-to-nature”)		Premodern Science
		Mechanization (“mechanical objectivity”)	Mid-Nineteenth Century
	Expertise (“trained judgment”)	Formalization (“structural objectivity”)	Early Twentieth Century
COLLECTIVE LEVEL	Coordination (“regulatory objectivity”)		Late Twentieth Century
		Meta-Evaluation (“forensic objectivity”)	

Figure 2: Funnel Plots

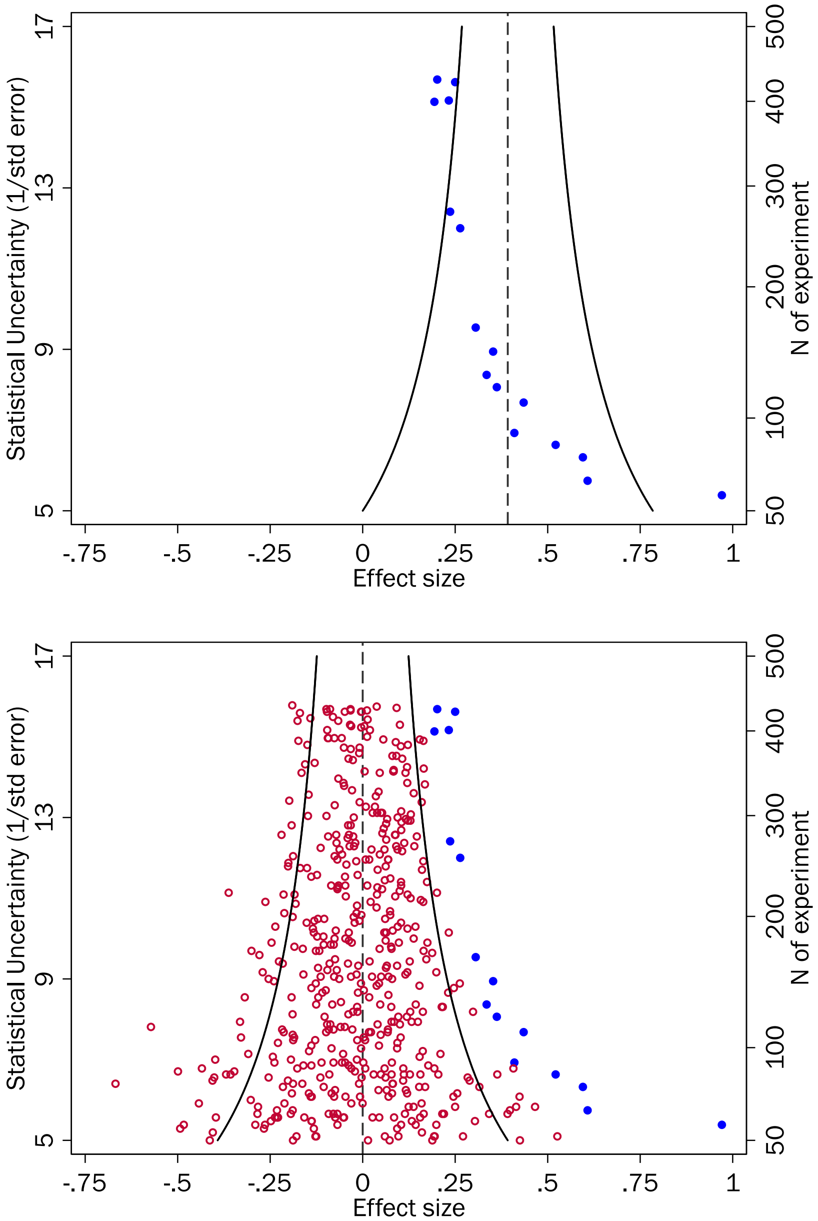


Figure 2. Funnel plots. Each dot represents a simulated study estimating a true effect size of zero (see supplemental material for simulation design). The top panel is a collection of studies that report positive, statistically significant findings. Bias in the collection is evident from the negative association between the observed effect size (*x*-axis) and its statistical uncertainty (*y*-axis). The bottom panel includes the effect sizes from all the simulated studies that were not statistically significant in the predicted direction (hollow circles). Note both that (1) only in the bottom panel do results correspond to the expected funnel shape, and (2) the average effect size in the biased collection (dashed line) diverges sharply from the average of zero in the unbiased collection.

Figure 3: P-curves

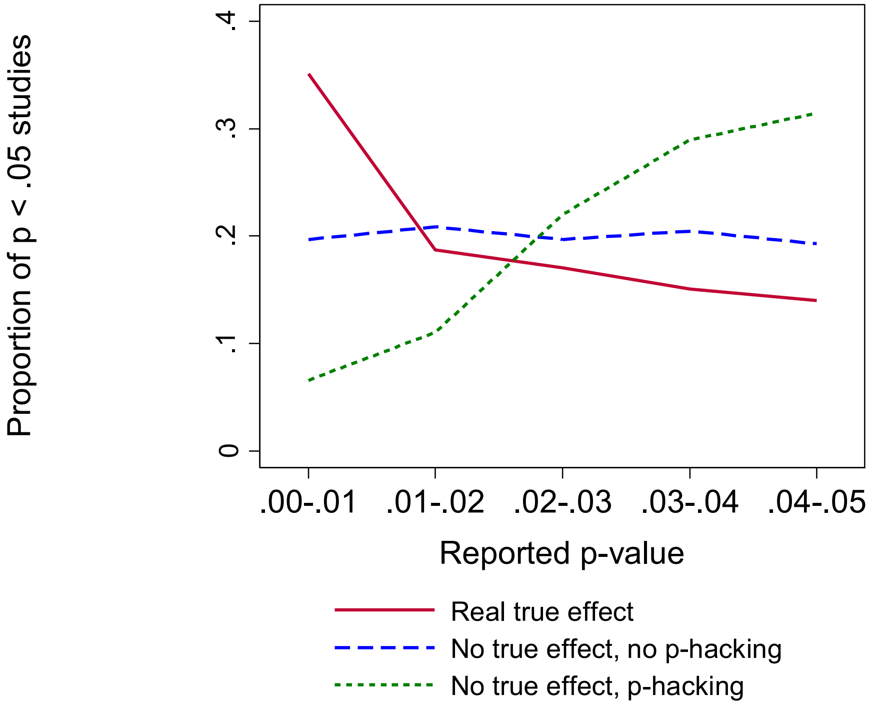


Figure 3. P-curves. Three different scenarios under which statistically significant findings are generated have different implications for the distribution of p-values reported in those studies. In a set of studies estimating a parameter that is not zero, the p-curve will slope downward (see supplemental material for simulation design). If the parameter actually is zero, significant findings are "false positives." The expected distribution of a set of false positives will be flat (the line with longer dashes), but various dubious analytic practices collectively known as "*p*-hacking" may produce a distribution of p-values that slopes upward (the line with shorter dashes).

Figure 4: The Reinforcing Character of Forensic Objectivity

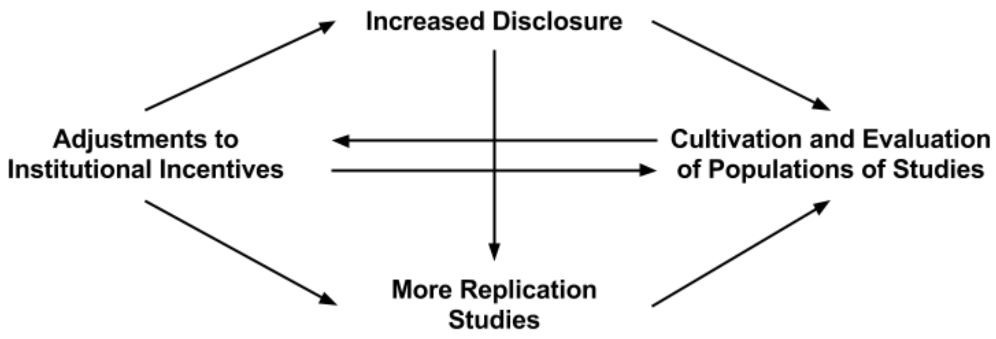

Increased Disclosure
Adjustments to Institutional Incentives
Cultivation and Evaluation of Populations of Studies
More Replication Studies

Table 1. Problems with *p*.

<table>
<tr><td colspan="2">The standard interpretation of a p-value in an experiment is the probability of observing the difference between treatment and control groups if the manipulation had no actual effect. This interpretation assumes a given hypothesis test is the only test. The following practices all make p fictive in ways that inflate the probability of obtaining a publishable p-value when the null hypothesis is true.</td></tr>
<tr><td>File drawer problem</td><td>The investigator conducts many experiments of many hypotheses but selectively reports experiments based on whether p is significant.</td></tr>
<tr><td>Dropping studies</td><td>The investigator conducts multiple experiments that test a hypothesis but selects which to report based on whether p is significant.</td></tr>
<tr><td>Data peeking</td><td>The investigator computes p as data are being collected, deciding to stop collecting data if results are significant but continuing otherwise.</td></tr>
<tr><td>p-hacking</td><td>The investigator tests the hypothesis by analyzing the data in various ways and determining which analyses to present based on whether the results are significant.</td></tr>
<tr><td>HARKing</td><td>The investigator conducts exploratory analyses, devises a post hoc explanation for an analyses for which a significant result is found, and then interprets the result as if it were an <i>a priori</i> prediction.</td></tr>
</table>

Appendix A. Materials and Warrant.

The authorship of this paper represents an unusual collaboration that is important for understanding the scientific warrant (Katz 1997) for the arguments we make. The first author is a sociologist who identifies as both a quantitative social psychologist and social science methodologist who has served in elected offices for the sections corresponding to both identities of the American Sociological Association. The first author has published empirical research in psychology and social psychology journals and has served as principal investigator of a federal grant that has involved supervising the fielding of hundreds of social science experiments embedded in population-based surveys, including many experiments by social psychologists. Part of the first author's methodological work has involved participating in various "open science" initiatives, including as co-author a recent paper on journal guidelines published in *Science*.

The second author is a science studies scholar who has conducted a three-year ethnography of psychology laboratories covering 10 sites. These include two social psychology labs as well as labs in adult and developmental cognition, cross-cultural psychology, emotional development, and cognitive neuroscience. More detail is provided in (Peterson forthcoming). Fieldwork entailed both detailed observation as well as 52 interviews with psychologists, including professors, postdocs, graduate students, and lab managers. The fieldwork also included a six month participation in a weekly, methodologically-focused "journal club" hosted by a psychologist who is a prominent epistemic activist, which was attended by 15-20 faculty members, postdocs, and graduate students. The second author has also done extensive archival research, including collecting and coding over 1,000 e-mails sent over the 14-month formation period of one epistemic activist organization, the Open Science Collaboration.

Over the course of writing the paper, the authors have amassed an archive of the unfolding arguments about the "crisis" in social psychology that have occurred across various social media platforms and in hundreds of newspaper, magazine, and website articles. Considering these materials together has allowed us an extensive integration of our separate standpoints as authentic participant and deliberate, systematic observer. Our goal has been to achieve together a de-centered variant of what Collins (1998) has labeled "participant comprehension." In articulating our arguments for this paper, we draw upon disparate examples from these source materials. It bears emphasis that, however, that while the cited materials provide extensive in-text support for our contentions, these data are mere illustrations and should be understood within the larger framework of our respective native experience and ethnographic fieldwork.

APPENDIX B. Description and Stata code for simulations used for drawing illustrative funnel and p-curve plots.

Funnel plot: The set of simulated studies varied in sample size from 50 to 500, by selecting a random number *i* from a uniform distribution and setting the N for the treatment and control groups as = 25 + 10^{i}. The effect sizes for treatment and control groups were simulated as random draws from a normal distribution, and standard errors are this effect size divided by the square root of the sample size. A critical value of $z > 1.96$ (corresponding to two-tailed $p < .05$) was used to indicate statistical significance.

```
clear all

set seed 867530
set obs 500

local sd = 1
local floor_n = 25 // each condition, so true N is times 2

gen random = runiform()
gen N = int(`floor_n' * 10^(random))
gen se = `sd' / sqrt(N)
gen invse = 1/se
gen graphy = ((_n-1)/(_N-1))*225 + 25
gen graphse = `sd' / sqrt(graphy)
gen graphinvse = 1/graphse
gen graphci = 1.96 * graphse
gen mean1 = rnormal(0, se)
gen mean2 = rnormal(0, se)
gen size = mean2 - mean1
gen p = .

local rows = _N
forvalues i = 1(1)`rows' {

       local N = N[`i']
       local mean1 = mean1[`i']
       local mean2 = mean2[`i']
       qui ttesti `N' `mean1' `sd' `N' `mean2' `sd'
       replace p = r(p) in `i'

}

gen sig = (p < .05)

local if "if mean2 > mean1 & sig == 1"


qui su size `if'
local yline = r(mean)
gen graphu = `yline' + graphci
gen graphl = `yline' - graphci

local mcolor2 = "blue"

twoway ///
       (scatter invse size `if', msize(small) mcolor(blue)) ///
       (line graphinvse graphu, lcolor(black) yaxis(2)) ///
       (line graphinvse graphl, lcolor(black) yaxis(2)) ///
       , ///
       xlabel(-.5(.25)1, labsize(medlarge)) ///
       ylabel(5(4)17, axis(1) labsize(medlarge)) ///
       ylabel(15.81 "500" 14.14 "400" 12.25 "300" 10 "200" 7.07 "100" 5 "50", ///
              axis(2) labsize(medlarge)) ///
       ytitle("Statistical Uncertainty (1/std error)", axis(1) size(medlarge)) ///
       ytitle("N of experiment", axis(2) size(medlarge)) ///
       xtitle("Effect size", size(medlarge)) ///
       xline(`yline', lp(dash)) ///
       legend(off) ///
       scheme(s1mono) ///

local yline = 0
drop graphu
gen graphu = `yline' + graphci
drop graphl
gen graphl = `yline' - graphci

twoway ///
       (scatter invse size if sig == 1 & mean2 > mean1, msize(small) mcolor(blue)
msymbol(circle)) ///
       (scatter invse size if (sig == 1 & mean2 > mean1)==0, msize(small)
mcolor(cranberry) msymbol(circle_hollow)) ///
       (line graphinvse graphu, lcolor(black) yaxis(2)) ///
       (line graphinvse graphl, lcolor(black) yaxis(2)) ///
       , ///
       xlabel(-.5(.25)1, labsize(medlarge)) ///
       ylabel(5(4)17, axis(1) labsize(medlarge)) ///
       ylabel(15.81 "500" 14.14 "400" 12.25 "300" 10 "200" 7.07 "100" 5 "50", ///
       axis(2) labsize(medlarge)) ///
       ytitle("Statistical Uncertainty (1/std error)", axis(1) size(medlarge)) ///
       ytitle("N of experiment", axis(2) size(medlarge)) ///
       xtitle("Effect size", size(medlarge)) ///
       xline(0, lp(dash)) ///
       legend(off) ///
       scheme(s1mono)
```

***p*-Curve plot:** The “non-zero true effect” p-curve was simulated using a scenario in which the true effect increased *z* by .8 of a standard deviation from an otherwise null-effect distribution. The “*p*-hacked” curve was generated by simulating a scenario in which a researcher who obtained non-significant results had available two post-hoc analytic decisions to “nudge” the z-score of the null-effect toward statistical significance, each of which increased the z-score by a random value drawn from a uniform distribution ranging from 0 to .5.

```
clear all
set seed 8675309
set obs 10000

gen z_null = rnormal(0,1)
gen z_true = z_null + .8
gen z_phack = abs(z_null)

forvalues i = 1(1)2 {
       replace z_phack = z_phack + (runiform()/2) if z_phack < 1.96
}

foreach name in null true phack {
       gen p_`name' = 2*(1 - normal(abs(z_`name')))
```

```
        gen bin_`name' = .
        replace bin_`name' = 1 if p_`name' < .01
        replace bin_`name' = 2 if p_`name' < .02 & p_`name' > .01
        replace bin_`name' = 3 if p_`name' < .03 & p_`name' > .02
        replace bin_`name' = 4 if p_`name' < .04 & p_`name' > .03
        replace bin_`name' = 5 if p_`name' < .05 & p_`name' > .04
}

gen bin = _n in 1/5

foreach name in null true phack {

        gen pct_`name' = .
        count if bin_`name' != .
        local N = r(N)
        forvalues i = 1(1)5 {
                count if bin_`name' == `i'
                local binN = r(N)
                replace pct_`name' = `binN' / `N' in `i'
        }
}

twoway ///
        (line pct_true bin, lp(solid) lw(medthick) lc(cranberry)) ///
        (line pct_null bin, lp(dash) lw(medthick) lc(blue)) ///
        (line pct_phack bin, lp(shortdash) lw(medthick) lc(green)) ///
        , ///
        scheme(s1mono) ///
        xlabel(1 ".00-.01" 2 ".01-.02" 3 ".02-.03" 4 ".03-.04" 5 ".04-.05",
labsize(medlarge)) ///
        ytitle("Proportion of p < .05 studies", size(medlarge)) ///
        xtitle("Reported p-value", size(medlarge)) xscale(r(.75 5.25) titlegap(3)) ///
        legend(label(1 "Non-zero true effect") ///
                label(2 "No true effect, no p-hacking") ///
                label(3 "No true effect, p-hacking") ///
                rows(3) region(lcolor(none))) ///
        aspectratio(.75)

exit
```