aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
7(6) 562–571

DOI: 10.1177/1745691612457576
http://pps.sagepub.com
SAGE

Science or Art? How Aesthetic Standards Grease the Way Through the Publication Bottleneck but Undermine Science

Roger Giner-Sorolla
University of Kent, United Kingdom

Abstract

The current crisis in psychological research involves issues of fraud, replication, publication bias, and false positive results. I argue that this crisis follows the failure of widely adopted solutions to psychology's similar crisis of the 1970s. The untouched root cause is an information-economic one: Too many studies divided by too few publication outlets equals a bottleneck. Articles cannot pass through just by showing theoretical meaning and methodological rigor; their results must appear to support the hypothesis perfectly. Consequently, psychologists must master the art of presenting perfect-looking results just to survive in the profession. This favors aesthetic criteria of presentation in a way that harms science's search for truth. Shallow standards of statistical perfection distort analyses and undermine the accuracy of cumulative data; narrative expectations encourage dishonesty about the relationship between results and hypotheses; criteria of novelty suppress replication attempts. Concerns about truth in research are emerging in other sciences and may eventually descend on our heads in the form of difficult and insensitive regulations. I suggest a more palatable solution: to open the bottleneck, putting structures in place to reward broader forms of information sharing beyond the exquisite art of present-day journal publication.

Keywords

methodology, statistics, publishing

Imagine that two colleagues in psychology each show you a manuscript. One has two studies with roughly equal numbers of participants. Both studies support the hypothesis, each with a significant key result at $p = .04$. The other paper has three studies, also supporting the hypothesis, but the last two studies' individual results are only near significant: $p = .02$, $p = .07$, and that most annoying figure, $p = .11$.

After a quick calculation, you realize that you can actually have more statistical confidence in the three-study paper. The joint probability that those results could be found under the null hypothesis, using Fisher's method, is .0074; the joint null probability for the two-study paper is .012. The three-study paper thus has a lower overall p value, speaking more strongly against the null hypothesis. It also has one more study, so that if there are meaningful differences between the studies' methods, this more strongly establishes the generality of the effect.

But does this correspond to your intuitive assessment of the two papers? Probably not. Assuming everything else about the paper is good, your advice to the colleague with two significant results would be to submit to a well-regarded journal. But if you are like most academic psychologists, you would give different advice to the colleague with two marginal results: maybe drop the third study, run more participants in the second, and try different analytic techniques to "get them significant." If you are a savvy methodologist, you might suggest ending the three-study paper with a small meta-analysis pointing out the overall significant effect. But most likely, you will give that advice ruefully. The meta-analysis cannot hide that the individual results, though scientifically more reliable on the whole, are "messy," "ugly," and look "cobbled together." The three-study paper will have a harder time getting published.

Why do we allow aesthetic judgments, such as the p-value threshold of individual studies, to overshadow scientific judgments, such as the actual statistical evidence for a hypothesis? Are we practicing a science or an art? Indeed, artfully pleasing and clear presentations help us to communicate with each other and the public. But these standards should not mean that the uglier truth is completely suppressed. Yet under increasingly tight information economics of publication, the appearance of research has become a vital criterion for what is accepted as true in psychology and other sciences—to the detriment of truth seeking.

Corresponding Author:
Roger Giner-Sorolla, University of Kent, Keynes College, Canterbury, CT2 7NP, UK
E-mail: r.s.giner-sorolla@kent.ac.uk

Two Crises

> [I]t is a truly gross ethical violation for a researcher to suppress reporting of difficult-to-explain or embarrassing data in order to present a neat and attractive package to a journal editor. (Greenwald, 1975, p. 19)

We are currently confronting a crisis of confidence in research across scientific disciplines (Ioannidis, 2005; Sarewitz, 2012). In psychology, high-profile data fraud cases have recently given these concerns a special importance. But crisis is nothing new in psychology. "Crises" of existing practices and ideas in psychology have been declared regularly at least since the time of Wilhelm Wundt (for an overview, see Sturm & Mülberger, 2012, and articles in the associated special issue). Especially relevant to today's worries is the crisis that peaked about 40 years ago. The 1970s crisis had many facets. For example, in social psychology, mainstays of the field, such as the attitude concept and reliance on lab experiments, fell under question (Rosenthal & Rosnow, 1969; Wicker, 1969). However, other issues concerned all areas of psychology: limitations of null-hypothesis significance testing, bias toward positive results in publication, and the resulting lack of credibility of the standard research article (Elms, 1975; Greenwald, 1975).

Revisiting the 1970s methods crisis gives a certain sense of déjà vu. One key article, by David T. Lykken, appeared in *Psychological Bulletin* in 1968. It focused on an example pulled arbitrarily from the personality literature. A single study found that eating disorder patients were significantly more likely than others to see frogs in a Rorschach test, which the author interpreted as showing unconscious fear of oral impregnation and anal birth (Sapolsky, 1964). Lykken dissected the frog hypothesis in a wickedly amusing way. But his main point, supported by a survey of colleagues, was that the significant result was not enough to increase their acceptance of the hypothesis. If our research articles give no confidence, Lykken argued, our standards of evidence must be flawed.

Recent critiques of methodology resonate with Lykken's approach. Most notably, some critiques of Bem's (2011) precognition studies took their appearance in a top-ranked psychology journal as suggestive of flawed standards of evidence (LeBel & Peters, 2011; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011). Simmons, Nelson, and Simonsohn (2011) ran intentionally preposterous experiments to support their argument that even false hypotheses often appear true when we selectively use data analysis to ensure positive results. In one experiment involving the Beatles rather than frogs, participants reported significantly lower calendar ages after listening to "When I'm Sixty-Four."

In another resonance with today, the 1970s debate also questioned the weakness of current practices in the face of outright fraud. In the middle of that decade, Cyril Burt's findings on the heritability of IQ came under question (Gillie, 1977). Whatever the merits of accusations of fraud against Burt, which have proved controversial across the decades (Mackintosh, 1995; Samelson, 1997), the case led to reflection on how bias against publishing replications weakens the field's ability to detect fraud (Samelson, 1980; Wong, 1981). A number of writers have recently expressed similar concerns in the face of less controversial examples of fraud and, more generally, implausible or unreliable results (e.g., Ritchie, Wiseman, & French, 2012; Roediger, 2012).

Evidently, the measures taken to solve the issues of the 1970s have not been enough to keep them from popping up again. One such measure was to enforce multistudy criteria at the highest levels of publishing. "Ideally, all experiments would be replicated before publication but this goal is impractical," wrote Lykken (1968, p. 159; see also Elms, 1975). Despite Lykken's doubts, just such a change in standards happened postcrisis. In fields of psychology where data are easily collected, standards at the most respected journals shifted to practically require multiple studies. From 1976 to 1996, for example, studies per article in two top social and personality psychology journals, the *Journal of Personality and Social Psychology* (JPSP) and *Personality and Social Psychology Bulletin* (PSPB), increased by about 50% (Quiñones-Vidal, López-García, Peñaranda-Ortega, & Tortosa-Gil, 2004). Since then, the mean number of studies per article in JPSP has shot up even faster, from 2.20 in 2000 (Quiñones-Vidal et al., 2004) to 3.33 in 2008 (Witte & Brandt, 2011).[1]

But today, the multistudy solution shows holes. Disgraced social psychologist Diederik Stapel found no problem in fabricating data to meet the needs of four- and five-study JPSP articles. Bem's (2011) controversial JPSP article met and exceeded expectations of internal conceptual replication, presenting nine studies in support of precognition. Simmons et al. (2011) showed that taking analytic liberties seen as legitimate in psychology can lead to a 60% false positive rate. If all significant results are taken at face value, then regardless of a hypothesis's truth, only the most unlucky, uncreative, or poorly resourced researchers will fail to scrape together 3.33 studies in support of it.

The 1970s crisis, like today's, also forced reevaluation of the all-or-none Neyman-Pearson significance test as the gold standard of scientific truth (Hurlbert & Lombardi, 2009). After the 1970s, it slowly became acceptable to interpret "marginally significant" results at $p < .10$, to report exact p values, and to take into account statistical power and effect size (Cohen, 1994; Wilkinson & Task Force on Statistical Inference, 1999).[2] Statistical techniques of meta-analysis were also developed and used, in line with postcrisis pleas for more aggregation of results across studies (Epstein, 1980; Miller & Pollock, 1994b). Making the final word in psychology depend on the outcomes of many labs, instead of just one, is a safeguard against outright fraud. Better yet, it protects against the much more common false-positive biases that arise when positive results are disproportionately rewarded in publishing (Sterling, Rosenbaum, & Weinkam, 1995). To carry out its watchdog role effectively, an aggregate test should include all

attempts and all results, be they positive, negative, or inconclusive.

But although aggregate tests are now a part of the research landscape, they do not yet dominate it. Running a meta-analysis is long and painstaking. Although meta-analyses are often rewarded with a slot in a high-impact journal, their absence is rarely seen as a flaw in a midcareer curriculum vitae. Indeed, it might be smarter and faster to focus on making a name by publishing one's own research. Likewise, meta-analytic validation is not seen as necessary to proclaim an effect reliable. Textbooks, press reports, and narrative reviews often rest conclusions on single influential articles, rather than insisting on a replication across independent labs and multiple contexts. In this climate, it is hard to tell exactly how much evidence there is for the main point made by some well-cited classics. Finally, because the field does not disseminate or evaluate negative results from good-faith replication efforts, meta-analysis can tackle publication bias only indirectly, relying on the good will of researchers to share unpublished data (Rothstein, Sutton, & Borenstein, 2006). Elsewhere in this issue, Bakker, van Dijk, and Wicherts (2012) show that the steps taken by contemporary meta-analyses to gather studies from the "file drawer" are still not enough to defend against the impact of publication bias.

As all these problems arise again, we hear of the same kind of solutions and projects that were proposed after the 1970s crisis. In 1979, for example, the journal *Replications in Social Psychology* began publishing, its mission evident from its title. It put out three volumes before folding. *Representative Research in Social Psychology* was founded in 1970 and run by graduate students at the University of North Carolina at Chapel Hill, with the aim of publishing studies with good methodology regardless of results (Chamberlin, 2000). It had a longer run, but its last articles seem to have been published in 2006. Today, the online *Journal of Articles in Support of the Null Hypothesis*, founded in 2002, still lives but publishes only one to seven articles a year. Looking at the discouraging record of replication-focused journals, it is not clear whether the easier interface of a recently founded Web site dedicated to reporting replication attempts (psychfiledrawer.org) will be enough to attract and sustain contributions.

Responding to the current crisis, some writers have argued that journal editors should be more accepting of imperfect results (Kaiser, 2012; Simmons et al., 2011). This should sound familiar to those who have read Paul T. Wong's (1981) postcrisis critique of implicit editorial policies. To quote Wong: "It is my plea that editors and reviewers soften their insistence on perfection in paradigms or procedures. Their obsession with faultfinding may not only have discouraged many talented investigators from further research but also encouraged various questionable practices in reporting" (p. 691). According to Google Scholar, at the time I write this, Wong's brief commentary in the *American Psychologist* had received only six citations, none in English more recently than 1992. The case that journals have softened their criteria since then, from all accounts, is prima facie implausible.

These projects and pleas did not solve the 1970s crisis because they never addressed its root cause. When students and academics face tight constraints on time and resources, only a fool would spend effort trying to report mistakes rather than burying them or repeating someone else's work rather than promoting one's own. Even if they do get submitted, replications and articles with imperfect data will have a fatal disadvantage in a tight market. Nor will outlets flourish if they choose to relax their standards. As a graduate student in the 1990s, I was warned against trying to publish in certain journals because their appearance on my publications list would be seen as low or even negative in value. Perfectionism, it seems, also applies to the art of the career profile. Refusing to release findings that you have already worked on, just because their most likely outlet has low standards, is like throwing all pennies out of your house because having them around make you look cheap. It is a self-presentational, aesthetic argument, not a scientific one. But as with the aesthetics of data, it has to be recognized for what it is.

The Aesthetics of Scientific Results

> It is more important to have beauty in one's equations than to have them fit experiment. . . . If there is not complete agreement between the results of one's work and experiment, one should not allow oneself to be too discouraged, because the discrepancy may well be due to minor features that are not properly taken into account and that will get cleared up with further developments of the theory. (Dirac, 1963, p. 47)

Much has been written on the joy of an elegant theory in science. But as the physicist Paul Dirac acknowledged, beauty in a theory is not always matched by beauty in the data. The way in which we talk about data being "beautiful" and "neat" as opposed to "ugly" and "messy" shows that their content and presentation carry aesthetic value. Reality, however, should limit the influence of aesthetics on science (Engler, 1990). Dirac only said that a theory's beauty should encourage persistence in its testing. If empirical results consistently speak against it, it is the theory, not the results, that must be rejected or revised.

A highly selective publication market, with no credible alternate outlets for results, puts this standard in jeopardy. Science values a theory that is authentically supported by pleasing, strong, and consistent results, and rightly so. But what if only the most valuable of findings are allowed to be known? What if only scientists who can reliably present such findings are allowed to make a living from science? We can only expect that scientists under the gun will indulge in selective presentation to increase the apparent consistency of their results, even if most resist the temptation of outright fraud. Then, even the most gorgeous looking results become suspect, because the checks and balances that ensure their truth have failed.

In my opening thought experiment, I showed how apparent perfection in results can influence scientific evaluation. Early

European scientists accepted that illustrations could be prepared artistically to show ideal rather than real cases, whereas later scientists came to share a more modern idea of objectivity through direct reproduction (Daston & Galison, 1992). Still, a preference for aesthetically perfect results persists. Perhaps it is precisely because we trust scientific results to be objective that we are bowled over when they look perfect. Just as symmetry is diagnostic in human beauty because it reflects biological fitness (Rhodes, 2006), so a strong, clear, and unqualified statistical effect is diagnostic of useful information. Perfect-looking results are also easy to understand, and the resulting feelings of fluency in processing underlie aesthetic preference in a wide variety of judgments (Reber, Schwarz, & Winkielman, 2004). In this process, the $p < .05$ statistic, as Gigerenzer (2004) argued, has become a ritually applied hallmark in psychology, symbolizing the perfection of any given result.

If an investigator is not lucky enough to obtain naturally perfect results, there are ways to create their appearance, short of blatant fabrication. These tricks have been described in detail by Simmons et al. (2011): drop measures and whole studies that are not themselves significant, even if they go in the right direction and contribute to the overall trend; try out many different statistical analyses, covariates, and moderators and report only those that "work"; run more participants after the fact, hoping to reach the magic number of significance. These practices are now often accepted in psychology, but according to Simmons et al.'s (2011) simulations, they can inflate the false-positive rate of a study to the 60% range. In psychology and other fields, this leaves a trail of evidence. Reports that do not significantly support the hypothesis are underreported, compared with the rate of nonsignificant results to be expected under actual levels of experimental power, even if each and every hypothesis were true (Francis, 2012; Ioannidis & Trikalinos, 2007).

As when plastic surgery stretches an aging film star's face into an unappealing, taxidermic simulacrum of youth, there are ways of getting to $p < .05$ that are themselves ugly to the discriminating eye. Such telltales include shifting methods of analysis between studies; unexplained exclusion and transformation criteria (Simmons et al., 2011); and the still-popular practice of applying one-tailed tests (Lombardi & Hurlbert, 2009). Indeed, the one-tailed tests I have seen in psychology manuscripts rarely produce p values less than .025, rarely are applied throughout the manuscript, and never, ever lead to a refusal to interpret results because their direction is opposite to the one expected a priori. But even without obvious flimflam, a nagging doubt hangs over good-looking findings: Do they look perfect because the phenomenon is robust or because an unknown number of not-so-pretty studies have been socked away in the research attic like the portrait of Dorian Gray? It is this suspicion that Young, Ioannidis, and Al-Ubaydli (2008) describe as the "winner's curse" of scientific publishing. Fierce competition and no oversight on the completeness of reporting lead the appearance of perfect results to be mistaken for the reality of a robust effect.

A second aesthetic criterion follows how well scientific papers conform to an easily processed, self-promoting narrative format. Narrative has been studied extensively as a form of argument (e.g., Pennington & Hastie, 1991; Voss, Wiley, & Sandak, 1999), with the general finding that accounts conforming to narrative expectations are more persuasive.

In achieving a beautiful fit between hypotheses and data, one particular narrative temptation arises: to represent one's hypotheses as coming prior to results, when in fact they came after, adjusted to fit. Kerr (1998) proposes an acronym for this practice—HARKing, or "hypothesizing after [the] results [are] known"—and shows its increasing acceptability over the years in psychology, despite its many negative implications.

Admitting that you are wrong is part of science. But somehow the belief has taken hold that making such admissions in a research paper is a sign of weakness that muddies the story, eats up journal pages, and confuses the reader. Even being honest about an initial lack of theory or reporting a midcourse correction on the basis of a pilot study can be taken as a fatal flaw. I cannot add much to Kerr's (1998) original observation that HARKing is a response to aesthetic and presentational pressures, except to agree with the author that this practice continues undiminished in the years since the article was published (Kerr, personal communication, February 2012). One recent article has argued, tongue in cheek, that a priori scientific hypothesizing is the most reliable form of precognition because so few psychology papers state hypotheses that turn out to be disconfirmed (Bones, 2012).

If you have ever gotten a journal rejection letter because your findings were merely "incremental," you have fallen foul of a third aesthetic criterion: novelty. If you have ever then wondered, "But isn't science supposed to be incremental?," you have found one of its main problems. Research on aesthetics tells us that novelty has an inverted quadratic effect on preference. Unfamiliar things are distrusted and hard to process, overly familiar things are boring, and the perfect object of beauty lies somewhere in between (Sluckin, Hargreaves, & Colman, 1983). The familiar comes as standard equipment in every empirical paper: scientific report structure, well-known statistical techniques, established methods. In fact, the form of a research article is so standardized that it is in danger of becoming deathly dull. So the burden is on the author to provide content and ideas that will knock the reader's socks off—at least if the reader is one of the dozen or so potential reviewers in that sub-subspecialty.

Without novelty, science will stagnate. Intrinsically, a new idea is more exciting than an old one. We would all like to dine on what is new and interesting in the field, skipping over the dull and necessary. But when novelty determines whether results are disseminated at all, this creates a moral hazard. For one, it discourages due diligence in citing the literature, leading to a psychological theory that frequently "reinvents the

wheel" (cf. Miller & Pollock, 1994a). Although well-read editors and reviewers should be able to catch false representations of novelty, a more insidious by-product of the novelty criterion is its chilling effect on the communication and assessment of replication attempts. These are implicitly kept out of the top journals, whose reputation depends on the novelty of their articles, as Ritchie et al. (2012) found out when they tried to publish a nonreplication of Bem's (2011) precognition studies in the same prestigious outlet that published them. What is less defensible is the difficulty of replication attempts to get any form of review and distribution. All journals, no matter how small, have their own aspirations and receive enough research manuscripts on original topics to give replication manuscripts a very hard time.

Replication across independent labs is a crucial check on the validity of science (Collins, 1985). But the unglamorous nature of replication work, confronted with the narrow publishing bottleneck, makes much of it unpublishable, and therefore not worth starting, in a world of precarious careers and limited resources. In psychology, some rigorous and relevant nonreplications have gotten published (e.g., Bower & Mayer, 1985; Grice & Seely, 2000; Klauer & Musch, 2001). But even nonreplications have some novelty value, because they suggest boundary conditions (or, tantalizingly, misrepresentation) in the original experiment. What of the successful replications from independent labs that are also needed in order to establish an effect as genuine? The least novel kind of paper, indeed, may be the one that straightforwardly replicates the results of its original. Excluding these results from scrutiny is just as harmful to the truth as excluding nonreplications.

How is it that aesthetic benchmarks are allowed to influence the official conclusions of our field? Often, the decision rests with journal editors. They may cite page space in requesting revisions or rejecting papers outright or may just directly refer to aesthetic rather than scientific logic (e.g., "this study with weak results detracts from the paper"—even though it makes the results more certain by a factor of two or so). As editors' preferences quickly become known, authors conform in anticipation. Journal readers are also stakeholders, with a desire for clear and easily processed information content. One common response I have heard to suggestions for publication process reform goes along these lines: "Well, it's important to be honest of course, but nobody wants to read a bunch of half-baked results and replications in a journal." This may be true, but those results need to be made available somewhere, so we can have confidence in the fully baked conclusions that do get published. Responsibility for the encroachment of art into science, in sum, appears to be shared—and therefore diffused—among authors, readers, and editors.

I should also mention that none of these aesthetic standards is absolute in the realm of art. There can be aesthetics of imperfection, as opposed to perfection (e.g., the key Japanese aesthetic concept of *wabi-sabi*, which values flaws as reminders of the impermanence of reality); aesthetics of familiarity, as opposed to novelty (e.g., in the postmodern aesthetics of repetition and reproduction; Eco, 1985); and aesthetics of schema-incongruent narrative (as in Antonioni's film *L'avventura*, which intentionally never delivers the resolution of its central mystery). These examples make clear that aesthetic criteria do not have to distort scientific judgment. Perhaps, over time, we can learn to develop our professional aesthetics away from formalism and toward a greater realism. Currently, however, the criteria that undermine scientific realism resemble those that keep audiences coming back to Hollywood blockbusters. There are big promises of novelty and spectacle (Cowboys fight aliens!) but with a tried-and-true, scriptwriterly narrative arc that leaves no room to show the often tedious, unclear, and imperfect nature of real life—or real science.

The Bottleneck

In most fields of psychological research across the decades, the number of peer-reviewed outlets for publication has not kept up with a parallel increase in the amount of research being done. This phenomenon is described at length by Judson (2004) across all fields of science and is the topic of an economic analysis by Young et al. (2008), focusing on bioscience. Although a precise accounting of the narrowing bottleneck in psychology remains to be done, a good estimate of the rise in research-active people in my subfield comes from attendance at the annual Society for Personality and Social Psychology (SPSP) meeting. From an unexpectedly high figure of 812 at the first meeting in 2000, attendance reached roughly 1,500 in 2003 and 3,500 in 2010, with no sign of reaching a plateau yet, as nearly 4,000 attended the 2012 meeting (SPSP Dialogue, 2012). Most SPSP attendees present research posters or talks that they want to publish. A useful if rough figure might therefore be the ratio of the number of articles published in social and personality psychology journals (ISI Web of Knowledge, 2012; category: "PSYCHOLOGY, SOCIAL") to SPSP attendees. Just from 2003 to 2010, this ratio has dropped from 1.32 to 0.90 article spaces per head.

Nor must the bottleneck show a narrowing trend over time; the typical journal submission has also evolved through constant selective pressure. Analyses of the aforementioned social–personality psychology journals, JPSP and PSPB, found that their rejection rates remained fairly stable, and upward of 70%, across some 20 years, from 1976 to 1996. But at the same time, the number of pages per article increased by a factor of 2 to 4, and as already noted, the number of studies per article increased (Reis & Stiller, 1992; Sherman, Buddie, Dragan, End, & Finney, 1999). This time span coincides with the development of the main response to the first crisis, the requirement of more studies to confirm initially significant results.

Bornmann and Marx (2012) reviewed empirical studies of scientific peer review that lend support to an "Anna Karenina principle" named after Tolstoy's observation that happy

families are all alike. When resources supporting proposals are scarce, conjunction rather than sum rules are used in decision making. In effect, this means that the proposal with nothing wrong with it, rather than the proposal highest in overall excellence, is most likely to succeed. In the peer review process, there eventually comes a time when a journal editor implementing an 85% rejection rate has already discarded all the fatally methodologically flawed manuscripts and still has to choose among a number exceeding the available space. It is here that the "artistic" criteria of novelty and perfection of results can enter in.

In a head-to-head competition between papers, the paper testing ideas that are new will be preferred over the paper that confirms—or fails to confirm—existing ideas. Likewise, the paper with results that are all significant and consistent will be preferred over the equally well-conducted paper that reports the outcome, warts and all, to reach a more qualified conclusion. A perfect-looking paper might even be preferred over a paper that looks imperfect but reaches the same substantive conclusion. Perhaps the less pretty paper reports a principled reason for excluding one dependent variable that did not work out as planned or reports one or two nonsignificant results that nonetheless support the overall trend.

The aesthetic criteria of novelty, narrative facility, and perfection may also owe their influence to being relatively clear and straightforward to apply. A paper's ability to advance the field is a highly subjective judgment, while flaws in novelty, narrative, and perfection are easier to find and justify. Playing by "Anna Karenina" rules, an editor or grant panel will take the superficially novel account with perfect results over the possibly groundbreaking account with strong but aesthetically flawed evidence. This state of affairs should give us pause. Do we insist that only perfect results deserve to be declared as true, without considering the big picture? Then we are no better than the know-nothing who insists that every winter must be warm rather than believe in global warming. Science helps us overcome biases in testing the reality of our aesthetically appealing ideas by considering the entirety of the data, no matter how ugly or difficult they are to process intuitively.

Carrot or Stick?

> Individual scientists have to work and survive in the system as it exists. Without systemic, structural changes, individual, principled choices . . . may be futile and professionally destructive. (Kerr, 1998, p. 213)

What can be done in the face of all these problems? One solution might be to keep the publication bottleneck as it is but with smarter criteria for acceptance. Knee-jerk reliance on the $p < .05$ standard, study by study, would be replaced by a consideration of evidence across multiple sources of replication. Decisions would be based on the reality of support for hypotheses, rather than on the appearance of perfection in data. This would discourage the kind of statistical convolutions that are needed to reach the magic significance number. Also, the standard expectations about our narrative might be replaced by a more sophisticated appreciation of the place of exploration in research, as Kerr (1998) suggests. These expectations, perhaps, would be communicated across the field by prominently placed recommendations aimed at all who take part in the editorial process (cf. Wilkinson & Task Force on Statistical Inference, 1999).

However, smarter criteria will be hard to establish and maintain in the face of an entrenched status quo of interpretation. Recommendations often falter in the face of established procedure. For example, picking up an issue of *Psychological Science* at random from 2008, I was easily able to find at least two articles whose editors had not required the reporting of effect size for statistical tests, contrary to the recommendation of Wilkinson and Task Force on Statistical Inference (1999) almost ten years earlier.[3] What is more, just using smarter criteria for acceptance would still leave latitude for suppressing inconvenient results and still would leave no place for replication attempts. This solution falls far short of what is needed.

Merely expanding the bottleneck by creating more journal outlets for "messy" results also will not work. The failures of post-1970s-crisis outlets show that this kind of solution tends to wither and die in a competitive career environment. The kudos, grants, and jobs will continue to go to those who publish clear-looking results in top journals. Without incentive to do otherwise, other forms of dissemination will be seen as not worthwhile. They might even acquire negative value for individuals' careers and journals' reputations.

Journal readers, too, do not have the time or inclination to plow through articles full of hard-to-interpret but honest findings, unless they concern the reader's own particular research interests. The link between fluency and aesthetic value also means that what is ugly is difficult to process. Scientists and laypeople, according to psychology's own dual-process models of information processing, prefer easy-to-read accounts in areas where they are not personally interested or expert (Chaiken, 1980; Petty & Cacioppo, 1986). All the same, instead of privileging easily processed research articles that are based on perfect-looking conceptual replications from a single lab, we need to give greater priority to the kind of easily processed research article that has a more solid meta-analytic basis. To achieve this, we need clear career incentives for disseminating and assessing research attempts regardless of results. Only this way can these summaries be based on a complete evaluation of the research actually done. As I see it, we can either do this through our own institutions, or eventually—this crisis, or next crisis, or the one after that—our regulators and paymasters will force it on us.

Sharing ugly data as normal voluntary service

If you tell me it is impossible to get academics to do things that do not directly advance their own ideas, I will ask you if you are a journal editor and when was the last time you wrote a

review. We engage in peer review activities without substantial pay, partly because of social norms, partly because of indirect benefits (such as keeping an eye on other people's research), but ultimately because science requires that results pass methodological scrutiny, and someone has to provide that service. Beyond just the vetting of method and theory that we get from peer review, science needs a similar effort to look at the entirety of good-faith attempts at a question, regardless of results, for its conclusions to be accurate.

As Nosek and Bar-Anan (2012) argue, psychology's current journal publication system is already strained and inefficient, with barely enough capacity to deal with existing production of research. It is unrealistic to ask that more articles be processed through this system or to relax the system's standards to accept any article with a good idea and a sound methodology. However, some aspects of the journal system are needed to make information useful. Outside the journal gates, a shadow economy of information circulates in the form of unpublished manuscripts and conference posters. But as with contraband goods, the complete lack of review procedures means that these findings and ideas are of unknown quality and vulnerable to theft (a finding is currently only "real" when published).

What we need are professionally recognized arenas where all kinds of research results, not just perfect ones, can be shared and evaluated. Such arenas would have to be very specialized, focusing perhaps on one topic of research; researchers, like parents of small babies, have a high tolerance for dealing with mess as long as it is within the family. Nosek and Bar-Anan (2012) have outlined in some detail how an online structure would work in which review of results happens after they are published, with contributors and evaluators all evaluated by each other. This structure would exist in parallel to the more traditional outlets of publication. Moreover, it would not restrict dissemination of findings because they are messy, replicative, or inconclusive. As with editorial work, participating in this system would not be a primary consideration for career advancement, but a good record would act as a tiebreaker in close cases. And even more so than editorial work, the indirect benefits to one's recognition, reputation, and research timeliness would give good reasons to take part.

Sharing ugly data as imposed obligation

For an unpleasant taste of what could happen to psychology if we do not take the initiative in data sharing, keep an eye on developments in biomedical research in the coming years. That field is currently facing a similar crisis of methods but with much more intense rhetoric and consequences. It was in that field that Ioannidis (2005) declared that "most published research findings are false" because of incentives to publish only positive results. Confidence in the applied usefulness of the publication process has taken a hit, as clinical trials have demonstrated unusually low success rates when trying to reproduce academic results (Begley & Ellis, 2012; Prinz, Schlange, & Asadullah, 2011; Turner, Matthews, et al., 2008). Editorial recommendations are taking on an urgent tone (Casadevall & Fang, 2012).

It is likely that funding agencies will have to do something. They may well decide that the only way to solve the false-positive problem, if scientists are too shortsighted to regulate themselves effectively, is by top-down regulation. One easily conceivable step would be to make clear that failure to disseminate the full results of funded research for some kind of peer evaluation is fraud. Although the creaky machinery of the peer-reviewed journal has made such a demand impossible until recently, technology makes available a broader bandwidth and faster turnaround time. The requirement of full dissemination could then easily be extended to all institution-supported research, by adding it as a condition of institutional review board (IRB) approval. Failure to report results in a definite time frame would result in suspension of further approvals. In the United States, social science researchers have good cause to doubt IRB procedures, which often are applied inflexibly, bureaucratically, and with requirements more appropriate to biomedical research (Carpenter, 2006). If we do not redefine honesty on our own initiative, the prospects for a sensible regulatory regime from above do not look good.

Our research, one way or another, will have to ensure that we are more forthcoming with ugly data if we want to stay credible. Try talking to an educated layperson—or even to an academic in another discipline—and justifying the practices so many of us have learned as a way to get ahead. "If it doesn't come out significant, try a different statistical analysis and maybe it will." "If it doesn't come out significant, keep running participants (or studies) until you get something that works." "Leave out those not-quite significant measures from your report; they'll weaken the paper." The usual excuses rooted in the aesthetic preferences of the publication process—"nobody wants to read ugly data and failed pilot studies"—fail to pass a basic ethical smell test. There must be principled reasons to change analyses or drop participants and principled arguments describing why a failed study did not adequately test hypotheses, beyond just "it didn't confirm them" (LeBel & Peters, 2011). Although we may still want to publish streamlined reports for general consumption, basic norms of honesty demand that our claims be backed up with a fuller account, available for inspection one way or another.

The strongest reason why we clutch on to aesthetic criteria, perhaps, touches on our own lives and livelihoods. An artist maintains complete control over his or her production, but a scientist has to confront the unknown. Two scientists may start with equally brilliant and insightful ideas, but these ideas must then meet the empirical world. By happenstance, one idea might turn out to be perfectly supported by facts, while the other does not work at all. Now, if both scientists can continue to make a reasonable living from science, it is entirely fair that the lucky scientist gets the Nobel and other accolades, while the unlucky one passes into obscurity. But when the bottleneck becomes narrow enough, then in order to beat the competition

and move on to the next career stage, a scientist has to present not just good work but good work that has consistently proven its hypotheses.

This situation does not seem fair. Its unfairness, I think, gives us moral justification for whatever means people take to clean up messy results, as long as they are not just making up the data. Here, sympathy and collective self-interest among psychologists may very well overcome principled research practices. Anyone who stands on principle, unless very lucky in results, will fail to compete effectively. If we really want to be more honest scientists, we will have to let go of this lower-level moralization of science as art and recognize the personal dangers that beginning scientists take when they stake everything on the honestly presented outcome of research. Simply put, we will have to accept that research progresses more slowly under closer scrutiny. Careers will have to value following up established results as well as making a name with bold but risky ideas. This would make for a less spectacular field, certainly, but a more reliable one.

Well-meaning pleas are not enough. The current criteria that have led to fraud and falsehood have deep roots in the economics of information, which drives the economics of careers and lives. We need to create the kind of conditions that do not put us in moral hazard as scientists. A miraculous investment would be needed to reduce competition and uncertainty in scientific careers; but it is not likely to come. More feasibly, reform will have to rely on widening the bottleneck of communication about our work, so that research however artfully presented can be effectively evaluated for factual merit. It is likely that the labor-intensive world of traditional journal publishing lacks the bandwidth for this. Parties interested in uprooting fraud and inaccuracy should therefore support alternate methods of communication, backed up by clear system-level incentives, not just hand-wringing editorials or patchwork solutions. The decisions that our organizations make today in the face of the latest crisis will determine whether we, or someone else, will ultimately control the way we do research.

Acknowledgments

I would like to thank Anna Brown, Martha Carey, David Corfield, Geoffrey Francis, Norb Kerr, Brian Nosek, Hal Pashler, Murray Smith, Giovanni Travaglino, and Eric-Jan Wagenmakers for their insights and references along the way.

Declaration of Conflicting Interests

The author declared no conflicts of interest with respect to the authorship or the publication of this article.

Notes

1. Indeed, the growing demand for internal replication and extension of effects, with accompanying increases in article and review process length, has produced a recent reaction: the rise of brief report formats in psychology journals. However, this development has been criticized (e.g., Ledgerwood & Sherman, 2012). Among other reasons, it abandons the implicit deal struck after the 1970s crisis, in which a preference for multistudy papers would mitigate the unreliable nature of single-study findings.
2. Gigerenzer (2004, p. 589) called attention to the deletion of the following sentences between the 1974 and 1983 editions of the American Psychological Association's *Publication Manual*: "Caution: Do not infer trends from data that fail by a small margin to meet the usual levels of significance. Such results are best interpreted as caused by chance and are best reported as such. Treat the result section like an income tax return. Take what's coming to you, but no more."
3. These articles were by Sayette, Loewenstein, Griffin, and Black (2008) and Zhong, Dijksterhuis, and Galinsky (2008); but these authors deserve no blame for deviating from a standard that the review process apparently does not explicitly enforce.

References

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*, 543–554.

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature*, *483*, 531–533.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407–425.

Bones, A. K. (2012). We knew the future all along: Scientific hypothesizing is much more accurate than other forms of precognition—A satire in one part. *Perspectives on Psychological Science*, *7*, 307–309.

Bornmann, L., & Marx, W. (2012). The Anna Karenina principle: A way of thinking about success in science. *Journal of the American Society for Information Science and Technology*, *63*, 2037–2051.

Bower, G. H., & Mayer, J. D. (1985). Failure to replicate mood-dependent retrieval. *Bulletin of the Psychonomic Society*, *23*, 39–42.

Casadevall, A., Fang, F., & Morrison, R. P. (2012). Reforming science: Methodological and cultural reforms. *Infection and Immunity*, *80*, 891–896.

Carpenter, D. (2007). Institutional review boards, regulatory incentives, and some modest proposals for reform. *Northwestern University Law Review*, *101*, 687–706.

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology*, *39*, 752–766.

Chamberlin, J. (2000). A student publishing tradition. *APA Monitor*, *31*, 36.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.

Collins, H. M. (1985). *Changing order: Replication and induction in scientific practice*. London, England: Sage.

Daston, L., & Galison, P. (1992). The image of objectivity. *Representations*, *40*, 81–128.

Dirac, P. A. M. (1963). The evolution of the physicist's picture of nature. *Scientific American*, *208*, 45–53.

Eco, U. (1985). Innovation and repetition: Between modern and postmodern aesthetics. *Daedalus*, *114*, 161–184.

Elms, A. C. (1975). The crisis of confidence in social psychology. *American Psychologist*, *30*, 967–976.

Engler, G. (1990). Aesthetics in science and in art. *British Journal of Aesthetics*, *30*, 24–34.

Epstein, S. (1980). The stability of behavior: II. Implications for psychological research. *American Psychologist*, *35*, 790–806.

Francis, G. (2012). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin & Review*, *19*, 1–6.

Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics*, *33*, 587–606.

Gillie, O. (1977). Did Sir Cyril Burt fake his research on heritability of intelligence? Part I. *The Phi Delta Kappan*, *58*, 469–471.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*, 1–20.

Grice, J. W., & Seely, E. (2000). The evolution of sex differences in jealousy: Failure to replicate previous results. *Journal of Research in Personality*, *34*, 348–356.

Hurlbert, S. H., & Lombardi, C. M. (2009). Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. *Annales Zoologici Fennici*, *46*, 311–349.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, e124.

Ioannidis, J. P. A., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials*, *4*, 245–253.

ISI Web of Knowledge. (2012). *Journal citation reports*. Retrieved from http://admin-apps.webofknowledge.com/JCR/JCR?PointOfEntry=Home&;SID=R2H17OMe9G5oN@26iil

Judson, H. F. (2004). *The great betrayal: Fraud in science*. London, England: Harcourt.

Kaiser, C. R. (2012). Campaign for real data. *Dialogue*, *26*, 8–10.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*, 196–217.

Klauer, K. C., & Musch, J. (2001). Does sunshine prime loyal? Affective priming in the naming task. *Quarterly Journal of Experimental Psychology: Section A*, *54*, 727–751.

LeBel, E. P., & Peters, K. R. (2011). Fearing the future of empirical psychology: Bem's (2011) evidence of psi as a case study of deficiencies in modal research practice. *Review of General Psychology*, *15*, 371–379.

Ledgerwood, A., & Sherman, J. W. (2012). Short, sweet, and problematic? The rise of the short report in psychological science. *Perspectives on Psychological Science*, *7*, 60–66.

Lombardi, C. M., & Hurlbert, S. H. (2009). Misprescription and misuse of one-tailed tests. *Austral Ecology*, *34*, 447–468.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin*, *70*, 151–159.

Mackintosh, N. J. (Ed.). (1995). *Cyril Burt: Fraud or framed?* Oxford, England: Oxford University Press.

Miller, N., & Pollock, V. E. (1994a). Meta-analysis and some science-compromising problems of social psychology. In W. R. Shadish & S. Fuller (Eds.), *The social psychology of science* (pp. 230–261). New York, NY: The Guilford Press.

Miller, N., & Pollock, V. E. (1994b). Meta-analytic synthesis for theory development. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 457–483). New York: Russell Sage Foundation.

Nosek, B. A., & Bar-Anan, Y. (2012). Scientific Utopia I: Opening scientific communication. *Psychological Inquiry*, *23*, 217–243.

Pennington, N., & Hastie, R. (1991). A cognitive theory of juror decision making: The Story Model. *Cardozo Law Review*, *13*, 519–557.

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology*, *19*, 123–205.

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews: Drug Discovery*, *10*, 712.

Quiñones-Vidal, E., López-García, J. J., Peñaranda-Ortega, M., & Tortosa-Gil, F. (2004). The nature of social and personality psychology as reflected in JPSP: 1965-2000. *Journal of Personality and Social Psychology*, *86*, 435–452.

Reber, R., Schwarz, N., & Winkielman, P. (2004). Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? *Personality and Social Psychology Review*, *8*, 364–382.

Reis, H. T., & Stiller, J. (1992). Publication trends in JPSP: A three-decade review. *Personality and Social Psychology Bulletin*, *18*, 465–472.

Rhodes, G. (2006). The evolutionary psychology of facial beauty. *Annual Review of Psychology*, *57*, 199–226.

Ritchie, S. J., Wiseman, R., & French, C. C. (2012). Replication, replication, replication. *The Psychologist*, *25*, 346–348.

Roediger, H. L., III. (2012). Psychology's woes and a partial cure: The value of replication. *Observer*, *25*. Retrieved from http://www.psychologicalscience.org/index.php/publications/observer/2012/february-12/psychologys-woes-and-a-partial-cure-the-value-of-replication.html

Rosenthal, R., & Rosnow, R. L. (1969). *Artifact in behavioral research*. New York, NY: Academic Press.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2006). *Publication bias in meta-analysis*. New York, NY: John Wiley & Sons.

Samelson, F. (1980). JB Watson's Little Albert, Cyril Burt's twins, and the need for a critical science. *American Psychologist*, *35*, 619–625.

Samelson, F. (1997). What to do about fraud charges in science; or, will the Burt affair ever end? *Genetica*, *99*, 145–151.

Sapolsky, A. (1964). An effort at studying Rorschach content symbolism: The frog response. *Journal of Consulting Psychology*, *28*, 469–472.

Sarewitz, D. (2012). Beware the creeping cracks of bias. *Nature*, *485*, 149.

Sayette, M. A., Loewenstein, G., Griffin, K. M., & Black, J. J. (2008). Exploring the cold-to-hot empathy gap in smokers. *Psychological Science*, *19*, 926–932.

Sherman, R. C., Buddie, A. M., Dragan, K. L., End, C. M., & Finney, L. J. (1999). Twenty years of PSPB: Trends in content, design, and analysis. *Personality and Social Psychology Bulletin*, *25*, 177–187.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.

Sluckin, W., Hargreaves, D. J., & Colman, A. M. (1983). Novelty and human aesthetic preferences. In J. Archer & L. Birke (Eds.), *Exploration in animals and humans* (pp. 245–269). Wokingham, England: Van Nostrand Reinhold.

SPSP Dialogue. (2012). Retrieved from http://www.spsp.org/?page=dialogue

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *American Statistician*, *49*, 108–112.

Sturm, T., & Mülberger, A. (2012). Crisis discussions in psychology: New historical and philosophical perspectives. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*, 425–433.

Turner, E. H., Matthews, A. M., Linardatos, E., Tell, R. A., & Rosenthal, R. (2008). Selective publication of antidepressant trials and its influence on apparent efficacy. *New England Journal of Medicine*, *358*, 252–260.

Voss, J. F., Wiley, J., & Sandak, R. (1999). On the use of narrative as argument. In S. R. Goldman, A. C. Graesser, & P. W. van den Broek (Eds.), *Narrative comprehension, causality, and coherence: Essays in honor of Tom Trabasso* (pp. 235–252). Mahwah, NJ: Erlbaum.

Wagenmakers, E. J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, *100*, 426–432.

Wicker, A. W. (1969). Attitudes versus actions: The relationship of verbal and overt behavioral responses to attitude objects. *Journal of Social Issues*, *25*, 41–78.

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.

Witte, E. H., & Brandt, V. (2011). *Social psychological research: The comparison of four journals*. Retrieved from http://psydok.sulb.uni-saarland.de/volltexte/2011/2743/

Wong, P. T. (1981). Implicit editorial policies and the integrity of psychology as an empirical science. *American Psychologist*, *36*, 690–691.

Young, N. S., Ioannidis, J. P. A., & Al-Ubaydli, O. (2008). Why current publication practices may distort science. *PLoS Medicine*, *5*, 1418–1422.

Zhong, C. B., Dijksterhuis, A., & Galinsky, A. D. (2008). The merits of unconscious thought in creativity. *Psychological Science*, *19*, 912–918.