15

Blind Analysis as a Correction for Confirmatory Bias in Physics and in Psychology

Robert J. MacCoun and Saul Perlmutter

The perception that scientific psychology is in a state of "crisis" results from a perfect storm of coinciding developments.[1] First, there has been a steady stream of new cases of fraudulent data fabrication (and subsequent article retractions), triggered in part by new statistical methods of forensic re-analysis of published results (see Fang, Steen, & Casadevall, 2012; Simonsohn, 2013). Second, researchers have reported failures to replicate various prominent research studies (see Pashler & Wagenmakers, 2012; Yong, 2012; see Chapters 1 and 2). And third, new analyses and studies are demonstrating that research in psychology (and other social and behavioral sciences) is vulnerable to "*p*-hacking," "data-snooping," and "HARKing" (hypothesizing after the results are known) – a variety of questionable practices designed to obtain statistically significant results (Fanelli & Ioannidis, 2013; Ioannidis, 2012; Ioannidis & Trikalinos, 2007; John, Loewenstein, & Prelec, 2012; Kerr, 1998; Simmons, Nelson, & Simonsohn, 2011; Vul, Harris, Winkielman, & Pashler, 2009; see Chapter 5).

In fairness, many psychologists contend that the crisis is overstated, or that the proposed cures (discussed later) might be worse than the disease. Some argue that the obsession with Type I (false positive) errors distracts us from a more serious problem of pervasive Type II (false negative) errors (Braver, Thoemmes, & Rosenthal, 2014; Fiedler, Kutzner, & Krueger, 2012; see Chapter 4). Others are reassured that an ambitious "Many Labs" pilot replication project was able to reproduce 10 of 13 published effects using 36 independent samples (Klein et al., 2014; see Chapter 1). And statisticians have offered both frequentist (Sagarin, Ambler, & Lee, 2014) and Bayesian (Wagenmakers, 2007; see Chapter 8) perspectives in which disciplined data-snooping is both defensible and reasonable.

Psychological Science Under Scrutiny: Recent Challenges and Proposed Solutions, First Edition.
Edited by Scott O. Lilienfeld and Irwin D. Waldman.

And anyway, is not science ultimately self-correcting? Given enough research on a topic, one might expect biased studies to eventually cancel each other out. But this cannot happen when a research community's biases are homogeneous (MacCoun, 1998). Indeed, psychologists are fairly homogeneous in many respects – their training, their demographics (disproportionately European–American), and their politics (disproportionately left of center; see Duarte, Crawford, Stern, Jussim, & Tetlock, 2015; Gross & Fosse, 2012; Redding, 2012). But while we have difficulty seeing our shared biases, they seem more glaring to citizens outside our research community, making it easier for them to dismiss our findings (MacCoun & Paletz, 2009). This is particularly problematic for psychologists working on politically charged topics such as gender, race, ethnicity, cognitive ability testing, sexuality, parenting, or moral reasoning.

If psychology is in the midst of a crisis, we take the optimistic perspective that it is a healthy opportunity to strengthen the scientific study of mind and behavior. In all the sciences, we must constantly be re-inventing and improving our methods, as we learn new ways that we, very human scientists, can fool ourselves, and psychology is no different. Indeed, in the history of science, past epistemological crises are often seen as vital opportunities that led to improved methods and theories.

In this chapter, we consider the various forms of bias that contribute to the crisis, and then examine methods of *blind analysis* (MacCoun & Perlmutter, 2015) that physicists have developed to cope with similar inferential problems, and we sketch out various ways in which such methods might be adapted to canonical data analysis situations in psychology.

Biases in the Research Process

There are many forms of bias that can distort the selection and interpretation of research evidence. Here, we focus on two types of bias – *confirmation bias* and *disconfirmation bias*.

Confirmation biases occur when the analysis is conducted in a way that favors one hypothesis or result over others, irrespective of the actual direction of the evidence (see also Chapter 9). The literature on confirmation bias is now quite large, and it has developed from many different disciplinary and theoretical streams (e.g., Bruner & Potter, 1964; Klayman & Ha, 1987; Lord, Ross, & Lepper, 1979; Mahoney, 1977; Nickerson, 1998; Platt, 1964; Rabin & Schrag, 1999; Snyder, 1984; Wason, 1960). In fact, the term "confirmation bias" encompasses many distinct variants. They are all biases that involve a process that favors one conclusion more than justified by either logic or empirical reality. But the varieties of confirmation biases differ with respect to modes of inference – whether they involve deduction (logic) vs. induction (evidence); and, if inductive, whether they involve evidence gathering vs. evidence interpretation. Our chapter will primarily focus on evidence interpretation.

Somewhat confusingly, a particularly important form of confirmation bias is known as *disconfirmation bias* (Ditto & Lopez, 1992; Edwards & Smith, 1996). Despite the name, this is not the opposite of confirmation bias; it is simply an

asymmetric bias *against* one conclusion rather than (or in addition to) a bias in favor of a different conclusion. Thus, congenial or expected results are scrutinized in a lax manner, but facts that run counter to one's preferences or expectations are scrutinized in a more rigorous fashion.

Disconfirmation bias is hardly unique to psychology; the phenomenon is very familiar to physicists. For example, it probably explains some suspicious patterns in historical plots of the estimates of various key physical parameters over time. Figure 15.1 shows four such plots. Several features are apparent. First, in all four plots, the estimates eventually stabilize on a specific value. Second, the confidence intervals shrink over time. Both of these features match what one would hope to see in a successfully cumulative science. However, a closer inspection suggests that something is amiss. The new estimates tend to be strongly tethered to the running average of recent estimates in the past. This "serial autocorrelation" is obviously unrelated to any actual changes in the physical constants. Rather, it suggests that most of the estimates are influenced by previous studies. One might expect some

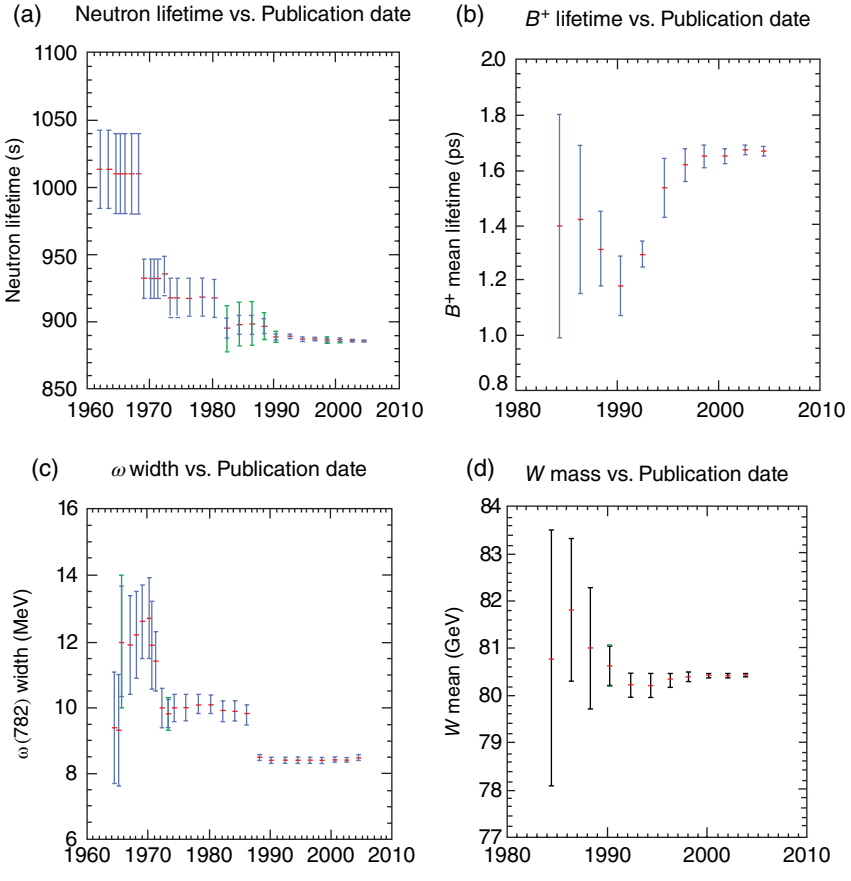

Figure 15.1 Reported estimates of various physical parameters by year of publication (Source: K. A. Olive et al., 2014).

temporal overlap due to common instrumentation and methods, but that cannot be the whole story here. Note that these are "one-sigma" error bars, implying a 68% confidence interval (rather than the 95% confidence intervals that are conventional in psychology). If valid, any one of these confidence intervals would lead one to expect that almost a third of future estimates would fall outside the confidence region. Instead, successive estimates almost completely overlap.[2] Indeed, these time series look strikingly similar to what is seen in experimental demonstrations of the intergenerational transmission of arbitrary cultural norms (Jacobs & Campbell, 1961; Kashima, 2014).

Feynman (1985, p. 342) offers an account of why scientists took so long to correct the first reported estimate of the electron's charge:

> It's a thing that scientists are ashamed of – this history – because it's apparent that people did things like this: When they got a number that was too high above Millikan's, they thought something must be wrong – and they would look for and find a reason why something might be wrong. When they got a number close to Millikan's value they didn't look so hard. And so they eliminated the numbers that were too far off, and did other things like that.

A related but conceptually distinct family of biases involve our susceptibility to be "fooled by randomness" (Taleb, 2001). Psychologists are familiar with this family under the pejorative labels "capitalization on chance" (Humphreys, Ilgen, McGrath, & Montanelli, 1969), "fishing expeditions" (Payne, 1974), and "data dredging" (Tukey, 1991) (see also Chapter 5).

For example, discoveries in particle physics often take the form of a histogram showing a peak – a large number of observations occurring at a particular point on a spectrum. Such inferences run the risk of capitalizing on fluctuations that are likely to appear somewhere in the data solely by chance. Physicists sometimes refer to a "look elsewhere effect" (Lyons, 2008), in which the investigator fails to properly discount for the number of possibilities examined when searching for an anomalous fluctuation: for example, if a particular location for a peak in a spectrum is not specified ahead of time, then any of the (perhaps thousand) bins in the spectrum might reveal a peak.

It can be difficult to completely distinguish confirmation biases from biases involving capitalization on chance, but one difference involves their time course. In confirmation bias, one conclusion is favored at the outset, whereas in capitalization on chance, an attractive conclusion seems to emerge from inspection of the data.

The variety of research biases can be classified with respect to motivation (does the investigator want this result?), intention (does the investigator intend to be biased?), and normative justification (is there an epistemological stance that justifies the bias?), suggesting five bias prototypes (MacCoun, 1998). *Fraud* is motivated, intentional, and normatively proscribed under any model of truth seeking. *Advocacy* involves intentional bias (selective emphasis on congenial evidence), but can be normatively defensible in some contexts (particularly when all parties understand that one is operating as an advocate). *Skeptical* processing occurs when one uses

unbiased methods to assess the diagnosticity of the evidence (the Bayesian likelihood ratio), but either integrates them with a very low prior probability estimate, or applies a very stringent standard of proof. An example might be an editor's scrutiny of an article purporting to support an extravagant claim such as extrasensory perception or extraterrestrial contact.[3] *Hot bias*es are unintentional but motivated; the evaluator wants and hopes to support a particular result. *Cold biases* are neither motivated nor intentional; they occur when we use faulty sampling or procedures that skew the results – possibly against our preferred result.

Note that confirmatory biases can vary from cold to hot. "Cold" confirmatory biases occur when we unwittingly use an inferential procedure skewed to favor a particular conclusion. The classic example is "positive test strategy" (Klayman & Ha, 1987), which disproportionately focuses on evidence consistent with a hypothesis (call it the H1+ cell), to the neglect of evidence inconsistent with the hypothesis (the H1– cell), evidence consistent with the alternative hypothesis (the H0+ cell), or evidence inconsistent with the alternative hypothesis (the H0– cell). There are situations in which the positive test strategy is normatively defensible or efficient (Klayman & Ha, 1987; Navarro & Perfors, 2011), but people clearly use it in situations in which it is likely to produce errors (e.g., Snyder, 1984). Cold confirmation biases are surely common in scientific practice. "Discoveries" are often notable precisely because the investigator shows that the H1+ cell is not empty – that the phenomenon of interest actually exists. Only later do researchers begin to flesh out its frequency and the necessary and/or sufficient conditions for its existence. And the pervasive lack of statistical power in social science studies shows that scientists routinely deploy methods biased against the hypothesis they are interested in (see Braver et al., 2014; Cohen, 1988) – although this bias is offset by others in the opposite direction (Ioannidis & Trikalinos, 2007; Simmons et al., 2011).

"Hot" confirmation biases occur when we prefer one conclusion over other possible candidates, even when we have no intention to be biased. This "motivated cognition" (Kunda, 1990) can take different forms, depending on the extent to which we are motivated to approach one conclusion vs. avoiding another one, and the extent to which we feel compelled to settle on a conclusion at all (Kruglanski & Webster, 1996). The stereotypic image of the scientist as a cool, dispassionate, objective technician is belied by countless scientific biographies and tales of scientific discovery – most famously Watson's (1968) *The Double Helix.* Still, it is important to distinguish these hot biases from outright fraud. Kunda (1990) reviewed evidence that motivated cognition is perhaps better characterized as "warm" because people are rarely completely impervious to or rejecting of uncongenial facts.

Fishing expeditions (a form of capitalization on chance) also range from cold to hot. Many ephemeral "discoveries" of the dustbowl empiricist era of early factor analysis were made by investigators operating in good faith who had not yet recognized the conceptual risks inherent in large sets of pairwise significance tests.[4] But where confirmation biases often involve a "need for specific closure" (a need for one particular answer), capitalization on chance often involves a "need for non-specific closure" – a desire to find *something* interesting, whatever it may be (Kruglanski & Webster, 1996).

In either case, the motivations can involve a mix of theoretical and professional considerations. Sometimes we prefer a result because we favor a theory that predicts it; sometimes we prefer a result because we think we can publish it or get the *New York Times* to report it. The motives need not be selfish or nefarious; we often simply want to help our graduate students find something interesting that they can present at a conference.

Corrective Practices in Psychology

There are a variety of traditional practices intended to minimize confirmation biases (see review and bibliography in MacCoun, 1998). Indeed, textbooks on research methodology and statistical analysis are primarily concerned with the reduction of bias, especially confirmation bias. (Reducing noise and increasing generalizability are other key goals.) Replication, peer review, and meta-analysis are essential tools in the debiasing toolbox, but, as discussed at the outset, they are clearly insufficient, and they arguably perform far less well than conventionally assumed.

There are less conventional practices and proposals. In Platt's (1964) "strong inference" scheme, the investigator tests the fit of data to each of many competing hypotheses, rather than testing for the support of any single candidate. New Bayesian methods provide a disciplined way that this might be implemented (e.g., Wagenmakers, 2007). The "destructive hypothesis testing" approach (Anderson & Anderson, 1996) requires the investigator to apply *disconfirmation* bias to one's preferred hypothesis, vigorously attempting to either falsify it or establish its boundary conditions. These approaches seem easy to implement, and, to some extent, each is already part of good scientific training.

More controversially, in Kahneman's (e.g., Kahneman & Klein, 2009) "adversarial collaboration" method, advocates for competing hypotheses collaborate in the design and conduct of a study, and then each participates in the analysis and interpretation. There are successful examples (e.g., Kahneman & Klein, 2009) but also some unpleasant failures to collaborate (Jost et al., 2009). Proposals to institutionalize routine replicability testing across labs (see Nosek, 2014; see also Chapter 1) have met the "proof-of-concept" test (Klein et al., 2014), but conducting fair and accurate replications is quite expensive in terms of labor costs, opportunity costs, and political costs.

Finally, there are proposals to institutionalize complete transparency via public registries of materials, data, and planned analyses and hypothesis tests (Miguel et al., 2014; Nosek, 2014; see also Chapter 5). Although registration of datasets is becoming routine in many fields, the proposal to register hypothesis tests in advance of data collection is somewhat problematic. First, as with institutionalized replication, registries pose labor costs and opportunity costs, especially for junior researchers who are already understaffed, underfunded, and overburdened in meeting the daunting publication standards of contemporary tenure review. Second, it is not inconceivable that "hypothesis trolls" could flood registries with

proposals as a low-cost means of discouraging others from researching those topics, similar to web domain squatters and so-called "patent trolls." But third, and more subtly, we see some risk that pre-registered analysis undermines some of the fun and excitement and openness to discovery that motivate scientific careers and leads to genuinely new insights.

Ideally, our corrective procedures should serve two different goals:

1 *Discourage biased evidence search and evidence assessment*
2 *Encourage active problem-solving and discovery*

This first goal is at the heart of the procedures we have discussed so far. It places a priority on the scientific values of honesty, objectivity, and rigor. Feynman (1985, p. 341) described "… a kind of scientific integrity, a principle of scientific thought that corresponds to a kind of utter honesty – a kind of leaning over backwards. For example, if you're doing an experiment, you should report everything that you think might make it invalid – not only what you think is right about it: other causes that could possibly explain your results; and things you thought of that you've eliminated by some other experiment, and how they worked – to make sure the other fellow can tell they have been eliminated."

Psychologically, the first goal is *prevention-focused*, oriented toward avoiding proscribed behaviors and bad decisions (Higgins, 1998). This kind of orientation piggybacks on our proclivity for "cheater detection" (Cosmides, 1989).

But it is important to preserve the second, more *promotion-focused* (Higgins, 1998) goal of exploration and discovery. According to John Tukey, one of the foremost statisticians of the twentieth century:

> Data analysis needs to be both exploratory and confirmatory. In exploratory data analysis, there can be no substitute for flexibility, for adapting what is calculated – and, we hope, plotted – both to the needs of the situation and the clues that the data have already provided. In this mode, data analysis is detective work – almost an ideal example of seeking what might be relevant. (Tukey, 1969, p. 90)

In the remainder of this chapter, we discuss the *blind analysis* family of methods increasingly used by physicists to counteract confirmation bias. We think these approaches, adapted to the distinctive needs of psychological science, can help to serve both epistemic goals.

Blind Analysis in Physics

We introduce this examination of methods in physics not because we want to encourage "physics envy," or because physics is "the queen of the sciences." Rather, we discuss physics because physicists have explored a family of methods called *blind analysis*, methods that seem potentially useful for psychological research, if suitably

adapted to the very different instrumentation and subject matter. Though far from universal, blind analysis methods are increasingly common in physics, especially particle physics and cosmology. Klein and Roodman (2005) provide a clear and authoritative review.

Blind analysis was apparently introduced into physics by Luis Alvarez, in research attempting to identify quarks. According to Lyons (2008, p. 907):

> A potential problem was that large corrections had to be applied to the raw data in order to extract the final result for the charge. The suspicion was that maybe the experimenters were (subconsciously) applying corrections until the value turned out to be "satisfactory."

To circumvent this problem, Alvarez and colleagues added random numbers to their raw estimates before analysis. This noise was only removed after the experimentalists were confident that they had made all appropriate corrections to the data. In this case, blinding prevented the researchers from publishing what was probably a spurious "discovery" of quarks where none were actually detected (Lyons, 2008).

Klein and Roodman (2005, p. 147) defined blind analysis as "a method that hides some aspect of the data or result to prevent experimenter's bias. There is no single blind analysis technique, nor is each technique appropriate for all measurements. Instead, the blind analysis method must carefully match the experiment, both to prevent experimenter's bias and to allow the measurement to be made unimpeded by the method." They described a wide array of blinding techniques, depending on what is being hidden (the signal being measured, the result of an analysis, the number of target events that have occurred), and how it is being hidden (through removal, through perturbation with noise, through a biasing offset).

According to Klein and Roodman (2005, p. 148), "it is crucial that the blind analysis technique be designed as simply and narrowly as possible. A good method, appropriately used, minimizes delays or difficulties in the data analysis." They caution that "[b]lind analyses solve only one problem, the influence of experimenter's bias on the measurement."

It is important to describe what blind analysis *is not*. It is similar in spirit and in logic to single- and double-blind methods used in clinical trials (see Schulz & Grimes, 2002; Stolberg, 2008), forensic science (Saks & Koehler, 2005), and even orchestra auditions (to reduce gender and race discrimination; Goldin & Rouse, 2000). But those methods tend to conduct blinding during data collection; blind analysis, as the name implies, applies blinding to the *data analysis process*. Obviously, the two approaches are complementary rather than mutually exclusive. Mathematically, it is perhaps closer to a literature addressing an entirely different problem – computer science work on methods to protect data confidentiality (see Fung, Wang, Chen, & Yu, 2010; Sweeney, 2002). But the goal there is to enable the analyst to conduct conventional analyses while protecting identifying information, so the "blind" is not intended to be lifted.

A case study

To illustrate blind analysis, we will sketch out how it was used in a paper by the second author and his colleagues in the Supernova Cosmology Project (SCP; Conley et al., 2006). By examining the brightness and spectra of high-redshift Type Ia supernovae, a series of papers by the SCP and the competing High-*z* Supernova Search Team determined that the well-known expansion of the universe is actually accelerating. This result is consistent with Einstein's formerly discredited cosmological constant, and implies the existence of "dark energy." As Conley et al. (2006, p. 1) noted, "the implications of this result for the future fate of the universe and our understanding of fundamental physics are profound; therefore it is extremely important that it be verified by independent methods."

One such method, deployed by Conley et al. (2006), involves measurements of a specific brightness metric in color-magnitude diagrams of Type Ia supernovae. Although the details are complex, the data analysis requires the researchers to select a variety of "cuts" with respect to data quality (e.g., the maximum allowable error associated with various measured parameters) and analysis (e.g., minimum and maximum redshift cutoffs). These analytic decisions create a potential for confirmation bias.

To reduce this risk, the SCP team employed a blind-analysis method suitable for their task (Conley et al., 2006, pp. 10–11). Their study sought estimates of two key quantities: Ω_M, an index of the density of matter, and Ω_Λ, an index of the density of dark energy (for this study assumed to have the properties of Einstein's cosmological constant). It is difficult to exaggerate the profundity of what these quantities tell us. According to our best cosmological understanding, whether the universe will expand forever or eventually collapse ("the Big Crunch") depends on the balance of these quantities. Thus, to minimize the likelihood of choosing data cuts that would produce a particular verdict, the SCP team applied *offset values* to their measurements, and these offset values were kept hidden from the analysis team until they judged that the analysis was complete. Thus, the team literally did not know what their findings implied until the blind was lifted. As seen in Figure 15.2, the resulting analysis confirmed earlier results, supporting the continued and accelerating expansion scenario, as well as the existence of "dark energy." Though the details are beyond the scope of this chapter, it is worth noting that the particular method of blinding that was used allowed the authors to successfully conduct almost all necessary "debugging" tests. This is the goal of a well-designed blind, but whether it can be fully achieved will depend on the specific measuring procedures, instruments, and analyses required for the study.

Applying Blind Analysis to Psychology

There is a long tradition of psychologists looking to physics for inspiration or as a benchmark for assessing psychology's progress as a science (e.g., Furr, 2002; Hedges, 1987; Lewin, 1931). But there are important differences between the disciplines.

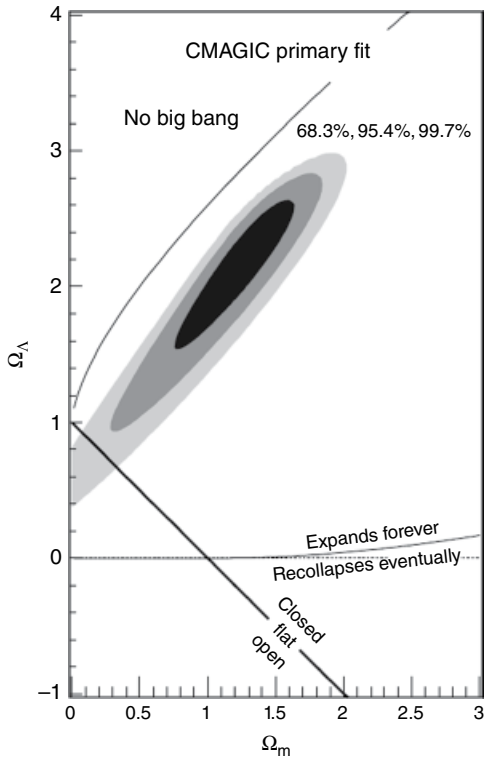

Figure 15.2 Results of Conley et al. (2006, Figure 6). Grey contours represent the 68.3%, 95.4%, and 99.7% confidence regions.

For example, physics generally has far greater measurement precision. And physics generally has strong formal theories that make precise quantitative predictions, so that an investigation's primary goal is often point estimation (rather than causal identification; see Meehl, 1967). Fanelli (2010) examined the frequency of rejections of the null hypothesis (a possible indication of publication bias) for 20 disciplines, finding that space science had the lowest rate (70.2%) and psychology/psychiatry had the highest rate (91.5%). Even so, it is clear that confirmation bias is a concern in both the social sciences and the natural sciences.

Psychology is an extremely heterogeneous field, both in its topics and in its methods. But in a recent content analysis of 155 studies in *Personality and Social Psychology Bulletin* (Kashy, Donnellan, Ackerman, & Russell, 2009), 52% reported an ANOVA or *t*-test, and 41% reported multiple regression techniques (including factor analysis, path analysis, and structural equation modeling). Thus, we will briefly sketch out a simulated example using ANOVA, followed by a more cursory discussion of possible blind methods for factor-analytic and path-analytic approaches.

An illustrative example

Consider the following situation, which is hypothetical, but not unlike many experiments in social psychology. (Any resemblance to specific studies in the literature is unintentional.) A psychologist is interested in the interactive effects of a source's expertise and conflict of interest on persuasion. An experiment is designed to examine the persuasiveness of an advertisement urging people to vote for a Massachusetts wetlands protection ballot initiative. The researcher conducts a 2×2 factorial experiment, with a complete crossing of two independent variables:

- *Source expertise*: 1 = low (source has a BA in biology from Harvard) vs. 2 = high (source is a Harvard PhD and a Harvard professor of biology)
- *Conflict of interest*: 1 = none vs. 2 = Harvard will get new wetlands science center if initiative passes

In our initial simulation, we randomly sampled 50 cases each from normal distributions with a standard deviation of 1 and cell means of 3 (low expert, no conflict), 3 (low expert, conflict), 4.5 (high expert, no conflict), and 2.5 (high expert, conflict) – consistent with the investigator's prediction that a conflict of interest will undermine the source's credibility, but only if the source is a Harvard professor rather than a Harvard alumnus.

A few comments are in order. First, assuming the study used seven-point Likert-type items, notice that our Monte Carlo procedure will produce some scores outside the 1–7 range. These can represent the types of rating and recording errors that occur in actual experiments. Second, note that the investigator has confounded expertise (BA vs. PhD) with current affiliation (alumnus vs. professor). This illustrates the kind of conceptual problems (e.g., construct validity) that data blinding is unlikely to correct. Finally, our investigator is to be commended for choosing a cell sample size of 50, making this better powered than the typical psychology experiment. (The Appendix also examines a case involving noisier data.)

In the Appendix, we compare means and F statistics for a number of different ways in which these data might be blinded. Here, we choose one we find particularly promising, which we call *cell scrambling* (MacCoun & Perlmutter, 2015).

In this method, the data for each of the four cells of the design are kept together, but the identities of the four cells are scrambled at random. For our four-cell design, there are $4! = 24$ possible orderings of the cells. Rather than sampling one such ordering, imagine that the investigator is given *a set* of, say, six of them. (Our intuition here is that there may be a cognitive sweet spot between providing only a single permutation and providing all 24; six seems like just enough to encourage hard thinking about the data.) Note that the true raw data have a chance (in this case, a one-in-four chance) of appearing in the ensemble of cell-scrambled datasets. By coincidence, in this run, the very first scrambled set is, in fact, the true data set, though of course the investigator should not know that until the blind is lifted.

As described in the Appendix, cell scrambling preserves the three F statistics, so the investigator will know whether there are significant effects, and how many. But he or she will not know *which* effects are significant, nor the patterns that the means actually take.

How might our clever and highly motivated investigator react to the ensemble of scrambled sets in Figure 15.3? In this ensemble, Sets 1 and 2 are likely to be very appealing; both support the same qualitative pattern that was predicted. Sets 3 and 4 are likely to appear tolerable, because each shows that experts are more persuasive and that conflicts reduce persuasion, but neither shows a particularly interesting interaction effect. Set 5 is quite different from the predictions, suggesting – counter-intuitively – that professors become more persuasive when they have a conflict of interest. Yet, after contemplating that pattern, it might occur to the investigator that what was intended to be a "conflict of interest" (a bad thing) might actually be seen as the professor's "skin-in-the-game" level of engagement in the state's ecological health (a good thing). Of the six, only the final set (in this run) is sufficiently implausible on its face that the investigator will probably dismiss it as a decoy. In principle, and perhaps even in practice, an investigator should be able to write up all six versions of the paper before the unblinding occurs.

One reason we find cell scrambling appealing is that it is so similar in spirit to one of the few consistently successful methods of "debiasing" many judgmental processes – the "consider the opposite" strategy, in which people are encouraged to systematically consider the opposite of whatever conclusion they are inclined to

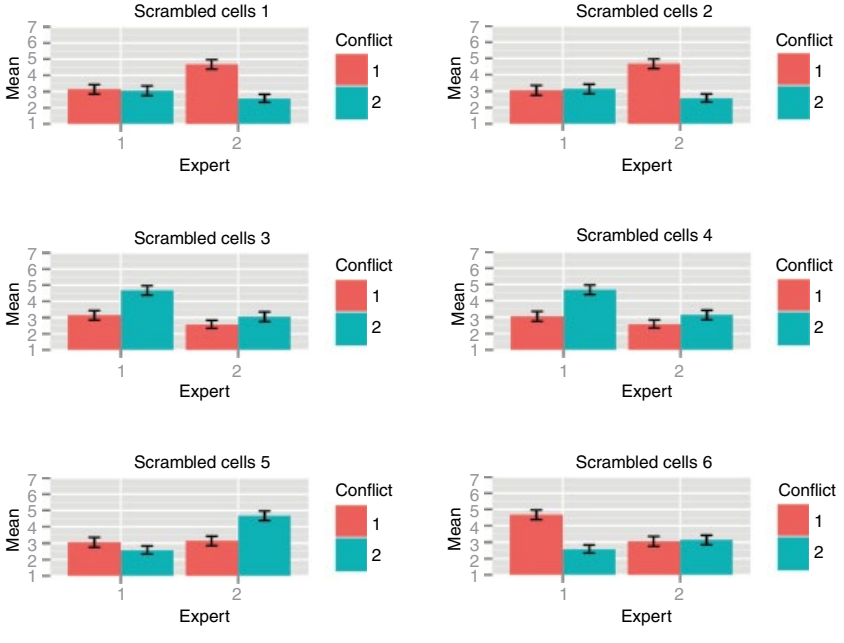

Figure 15.3 Six sets of blinded means perturbed by "cell scrambling." (By chance, Set 1 is identical to the unblinded raw data.)

reach (Lord, Lepper, & Preston, 1984). A drawback of cell scrambling, used alone, is that, while it does not reveal the nature of any significant result, it does show the investigator *whether* there is at least one significant result, and, as such, may fail to discourage some *p*-hacking practices. In the Appendix, we show that other methods blind the *p*-values but do a poorer job of blinding the substantive pattern of results. A hybrid approach might be to combine cell-scrambling with a method that perturbs the test statistics, although this might not obscure the likely significance of very large mean differences.

Blind analysis of correlational data

Much of empirical psychology is correlational; strictly speaking, experiments are correlational, but we use the term in its conventional sense of "non-experimental correlations" – that is, correlational statistics estimated in the absence of random assignment or strict experimental controls. Putting aside the serious problems of causal identification when interpreting non-experimental correlations, here our concern is with a different problem: the enormous risks of capitalization on chance in data sets that permit dozens or even hundreds of pairwise correlations to be estimated. This is a special concern in educational testing, neuroimaging (see Vul et al., 2009), and the so-called "big data" science (Marcus & Davis, 2014).

Analysts using multiple regression need to make many decisions about model specification: *What covariates should I include? Should I transform any of the variables? Which regression approach (i.e., link function) should I use – ordinary least squares? Logit? A multilevel model?* If the analysis includes multiple locations and/or time periods, there are additional choices to make: *Clustered standard errors? Fixed or random effects? What start year? What end year?*

Even in experimental psychology, correlational analyses play an important role. For example, many studies use some form of factor analysis to build a measurement model. Researchers want to know: *Do the data load on a single factor? Do the data fit my theory about measurement? Which items do I keep, and which should I throw out?*

And experimentalists often use some form of path analysis or structural equation modeling to ask: *Is the relationship between the manipulated variable (e.g., candidate name) and the measured dependent variable (e.g., voting) mediated by some hypothesized intervening variable (e.g., sexism or racism)?*

Although we do not develop them here, we can imagine many plausible ways of suitably blinding data for regression analysis, factor analysis, and path analysis.

One could apply noise, bias, or both to the individual data points (as in our methods 1, 2, 3, and 4) in the preceding text. Or one could apply noise + bias to the coefficients in the covariance matrix. Or one would simply scramble the identity of the items – that is, "coefficient-scrambling" rather than cell scrambling.

Discussion

How should blind analysis be implemented?

There are many procedural issues to consider. First, there are multiple ways to blind the data, and different methods will be appropriate for different situations. Choosing a blinding method requires some creativity, but making an informed choice will require serious mathematical analysis, Monte Carlo simulations, and empirical testing – well beyond any analyses offered here.

Second, once a blinding method has been selected, who should apply (and later undo) the blinding algorithm? A member of the team, or a neutral third party? When should the blind be lifted? Who enforces against peeking? Contemporary empirical physics often involves "big science." Physics data are sometimes sparse and difficult to obtain, requiring very large interdisciplinary teams. Psychology papers often have either a single author or a very small team (often consisting of one professor and his or her students). In theory, one might expect that the larger the team, the more likely that team members will object to any effort to cheat the blinding procedure (see Faia, 2000). But confirmation biases are often unconscious, and groups often amplify rather than attenuate shared biases (Kerr, MacCoun, & Kramer, 1996).

Third, are post-blind analyses permissible? According to Lyons (2008, p. 909): "A question that arises with blind analyses is whether it should be permitted to modify the analysis after the data had been unblinded. It is generally agreed that this should not be done … unless everyone would regard it as ridiculous not to do so." Conley et al. (2006, p. 10) pointed out that blind analysis is not mindlessly mechanical:

> A critical point is that these techniques do not seek to completely hide all information during the analysis. In fact, the goal is to hide as little information as possible while still acting against experimenter bias. Human judgment and scientific experience continue to play a critical role in a blind analysis. One does not mechanically carry out the steps of the analysis and then publish the results.

In some cases, an examination of the actual results may enable the team to recognize an overlooked error. Imagine, for example, finding out that unblinded data show that high school dropouts outperform college graduates in math problems; the implausibility of the result might help one discover that education levels were miscoded. But the important thing is to acknowledge any post-blind analyses and distinguish them from blind analyses in the write-up – much in the same way that psychologists are taught to report post-hoc tests separately from their main hypothesis tests.

Finally, is blind analysis voluntary, or should it be compulsory, and if so, who should be the enforcing agency? The university? The funding agency? A journal? Interestingly, in several areas of physics, blind analysis has emerged as a norm, and it is mostly self-enforced on research teams. As such, it has become an important

part of the socialization process; indeed, graduate students are often the most zealous about enforcing and protecting the blinding.

Should blind analysis be implemented?

These implementation questions are daunting but manageable. But readers might ask whether blind analysis is even worth the trouble.

Certainly, blind analysis is no panacea. According to Conley et al. (2006, p. 10):

> All that a blind analysis does is prevent unconscious misuse of particular types of information during the analysis process. The kind of data that are excluded from consideration (namely, the final answer derived from each option under consideration) is invariably that which no reasonable scientist would allow to consciously influence his or her decision-making process. However, subconscious effects are still present, and this is what this approach helps prevent.

In their survey of professional psychological researchers, John et al. (2012) asked about ten different "questionable research practices" (QRPs). The responses suggest that, contrary to what Conley et al. assume, many psychologists *do* let questionable considerations "consciously influence" their decision making.

We believe that a proper data blinding protocol, implemented honestly, would reduce or constrain three of these QRPs (all quoted bullet points are from John et al., 2012):

- Deciding whether to collect more data after looking to see whether the results were significant (58% self-admission rate under an incentivized honesty condition)
- Stopping collecting data earlier than planned because one found the result that one had been looking for (22.5%)
- Deciding whether to exclude data after looking at the impact of doing so on the results (43.4%)

But blind analysis, by itself, is no panacea. The three examples seem to involve direct confirmation bias, where blind analysis is most likely to be effective.

Four other QRPs involve capitalization on chance:

- In a paper, selectively reporting studies that "worked" (50%)
- In a paper, failing to report all of a study's dependent measures (66.5%)
- In a paper, failing to report all of a study's conditions (27.4%)
- In a paper, reporting an unexpected finding as having been predicted from the start (35%)

Blind analysis, by itself, is unlikely to prevent capitalization on chance, at least not in any mechanical way, but we believe the self-conscious cautiousness it produces reduces the likelihood of such practices.

But of course, blind analysis is unlikely to deter more blatantly fraudulent practices, such as:

- In a paper, "rounding off" a *p*-value (e.g., reporting that a *p* value of 0.054 is less than 0.05) (23.3%)
- In a paper, claiming that results are unaffected by demographic variables (e.g., gender) when one is actually unsure (or knows that they do) (4.5%)

Blind analysis is also unable to correct unreliable or invalid measurements, disentangle any confounded variables, improve causal identification of correlational evidence, or make a study more interesting or insightful. But blind analysis is just a valuable tool; it is not the whole toolbox.

Some Bayesians may feel that blind analysis is a "Band-Aid" solution where major surgery – abandoning null-hypothesis testing – is required (see Wagenmakers, 2007; see Chapter 8). Although Bayesian methods avoid some of the worst forms of *p*-hacking, Simmons et al. (2011, p. 1365) cautioned that "[a]lthough the Bayesian approach has many virtues, it actually increases researcher degrees of freedom. First, it offers a new set of analyses (in addition to all frequentist ones) that authors could flexibly try out on their data. Second, Bayesian statistics require making additional judgments (e.g., the prior distribution) on a case-by-case basis, providing yet more researcher degrees of freedom."

What do we want from blind analysis?

Although it is no panacea, blind analysis does offer certain strengths that replication studies and pre-registration do not. Unblinded replication studies run a risk of simply replicating shared biases (or introducing a new contrarian bias against the original findings). And, unlike pre-registration, blind analysis allows for an open-minded, exploratory frame of discovery. It motivates researchers to find *all* the errors, biases, and rival hypotheses in their study – not just the ones they do not like.

At its best, blind analysis is more than just an algorithm for data processing; it provides a disciplined habit of mind. As Feynman (1985, pp. 342–343) argued, "this long history of learning how to not fool ourselves – of having utter scientific integrity – is, I'm sorry to say, something that we haven't specifically included in any particular course that I know of. We just hope you've caught on by osmosis. The first principle is that you must not fool yourself – and you are the easiest person to fool" (see Chapter 9).

Appendix

There are many possible ways of blinding data, and different methods will be appropriate in different analytic situations, depending on the measurement and statistical properties of the data, the procedure by which they were obtained, the types of

experimental manipulations and controls that were deployed, and so on. In this Appendix, we compare cell scrambling (described in the preceding text) with four other potential methods of blinding data from the hypothetical 2 × 2 psychology experiment we describe in the main text – first in a simulation of a well-powered experiment (i.e., adequate sample size), and then in a simulation of a weakly powered experiment. We show that different blinding methods have different strengths and weaknesses with respect to correcting errors and discouraging biases. For example, some methods are more effective at blinding the substantive results (the cell means), while others are more effective at blinding the statistical significance of the results (the *p*-values). But investigators do not necessarily have to confront this tradeoff, because it is possible to combine two or more approaches. Our list of approaches is not exhaustive, and we hope others will explore and test additional methods of data blinding, tailored to the specific features of other research situations.

A 2 × 2 Factorial with moderate effects and good power

In Table 15.1, we show the F statistics for the Expert and Conflict main effects and the Expert × Conflict interaction for a single run of the simulation that is described in the main text; the raw means are plotted in the left panel of Figure 15.4. For present purposes, we limit ourselves to this single illustrative run and do not consider asymptotic properties or sensitivity analyses of the various parameters of the simulation. The first row shows the F statistics for the "raw data" – what the investigator would see if the data were unblinded. In this case, the three effects correspond to effect sizes of $\eta_2 = 0.05$, 0.19, and 0.16, where $\eta_2 = 0.01$, 0.06, and 0.14 are considered the benchmarks for "small," "medium," and "large" effects, respectively (Cohen, 1988).

Table 15.1 Simulation 1: F Statistics.*

	Expert		*Conflict*		*Expert × Conflict*	
Raw data	14.6	***	60.8	***	51.4	***
Raw + noise	0.1		16	***	7.5	**
Raw + cell bias	153.5	***	321.5	***	142	***
Raw + noise + cell bias	13.3	***	0		10.3	**
Row scrambling	0.2		0.4		0.2	
Cell scrambling						
Set 1	14.6	***	60.8	***	51.4	***
Set 2	14.6	***	51.4	***	60.8	***
Set 3	60.8	***	51.4	***	14.6	***
Set 4	51.4	***	60.8	***	14.6	***
Set 5	60.8	***	14.6	***	51.4	***
Set 6	14.6	***	51.4	***	60.8	***

* Each effect has 1 degree of freedom, and the error term has 196 degrees of freedom.

The remaining rows show the F statistics for these same raw data after various blinding methods have been used to transform the data.

Before presenting our blinding methods, what might we want blind analysis to achieve here? For a 2×2 experiment, we want to minimize biases in any of the following:

1 Data deletion
2 Data correction
3 Data transformation
4 Significance testing (crossing the $p < 0.05$ threshold)

Blinding method 1. Add noise. In our first blinding method, we perturb the raw data by averaging each of the 200 data points together with one of 200 random numbers sampled from a uniform (minimum = 1, maximum = 7) distribution: viz., $blind_i = average(raw_i, noise_i)$. As seen in Table 15.1 and the right panel of Figure 15.4, this has the regressive effect of weakening all the effects. Despite the fact that the random numbers were sampled uniformly from the full-scale range (1–7), the perturbed data are qualitatively similar (and the ordinal rankings are identical) to the raw data, at least for this scenario involving a strong "true" effect pattern. As such, in this case, this blinding method could actually backfire – encouraging the investigator to engage in more strenuous *p*-hacking to obtain statistical significance and/or

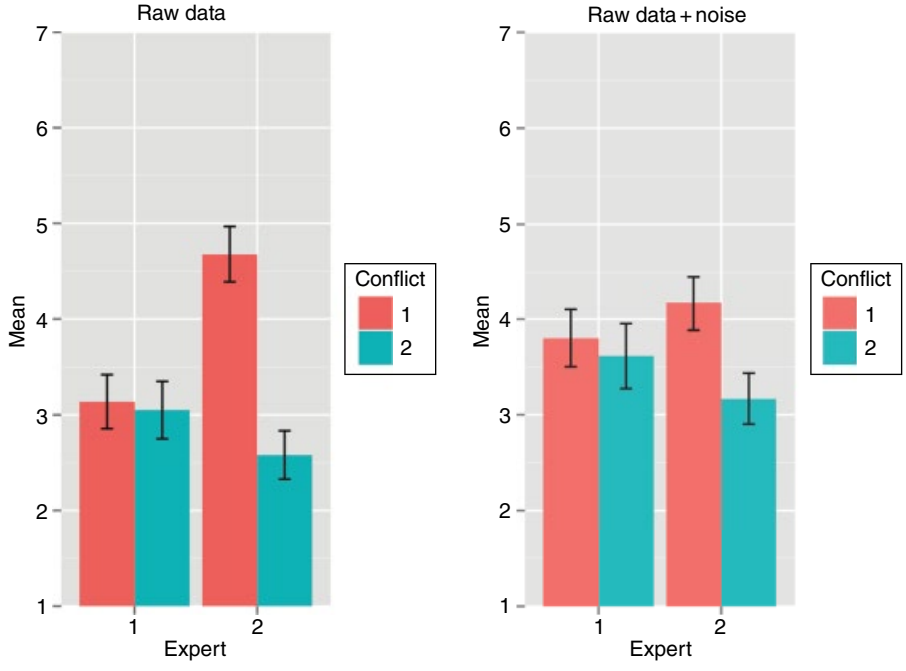

Figure 15.4 Raw means from Simulation 1 (left panel) and raw means perturbed by noise (right panel).

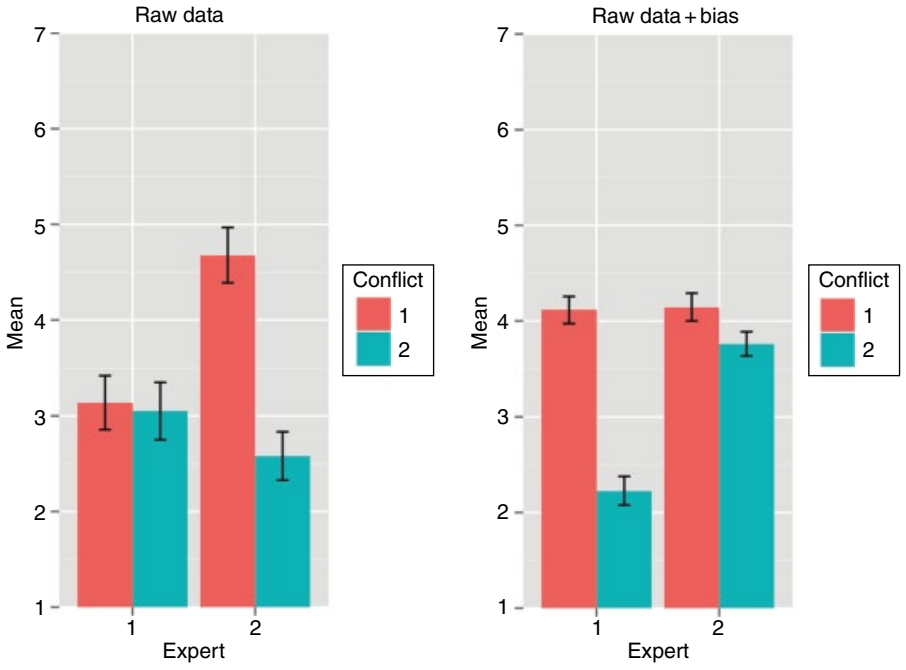

Figure 15.5 Raw means from Simulation 1 (left panel) and raw means perturbed by cell-specific bias (right panel).

strengthen the apparent effects. Adding noise is likely to be more effective as a blind when the measurement scale is less tightly bound than our narrow Likert-type scale (e.g., kilograms or miles or dollars).

Blinding method 2: Add cell bias. For our second blinding method, we perturb each of the 200 data points by averaging them together with the appropriate one of four cell-specific bias terms, each of which was sampled from a normal distribution with the same grand mean and SD as the full raw data distribution. As seen in Table 15.1, this produced significant main effects and a significant interaction – discouraging the temptation to *p*-hack. However, as seen in Figure 15.5, the qualitative pattern of means is quite different (e.g., the first cell mean is increased and the third cell mean is decreased by the blinding), so there is little reason to selectively edit the data.

Blinding method 3: Add noise + bias. Our third method simply combines the first two; we take the same vector of random numbers as method 1 and the same vector of bias terms from method 2, and average each of the 200 data points with their corresponding noise and bias terms (Figure 15.6).

Blinding method 4: Row scrambling. In our fourth method, we leave the raw outcome scores intact, but we "re-randomize" (or "post-randomize") the assignment to condition, so that the newly assigned cells no longer correspond to the true experimental condition for any given subject except by chance (in this case, a one-in-four chance). As seen in Table 15.1, as one might expect, row scrambling is strongly regressive, all but eliminating any hint of systematic effects in the data. This is the

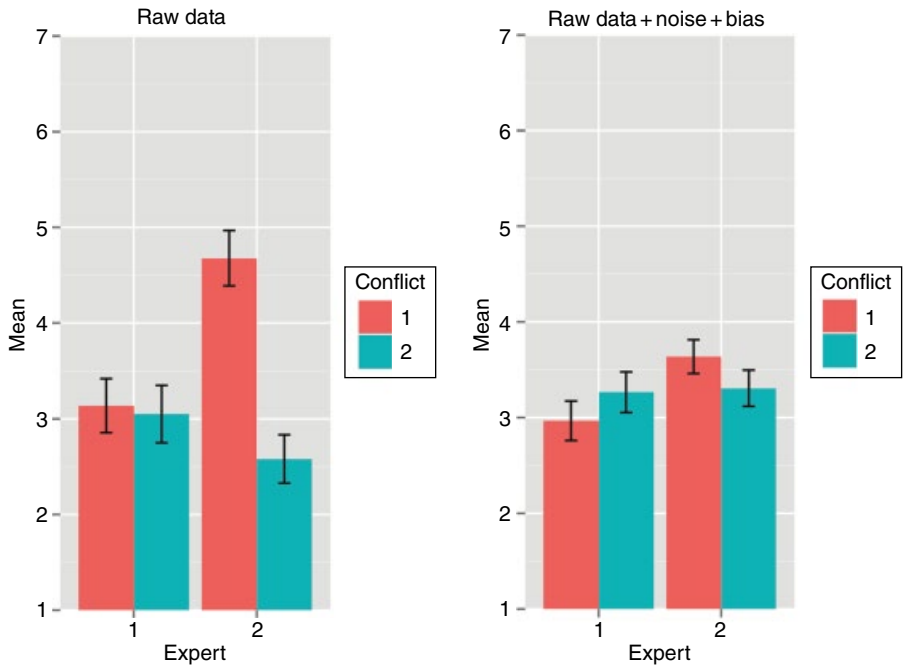

Figure 15.6 Raw means from Simulation 1 (left panel) and raw means perturbed by both noise and cell-specific bias (right panel).

most "blinding" of our methods – it obscures the qualitative pattern of effects while at the same time driving the F statistics so close to zero that all but the most egregious *p*-hacking is unlikely to be effective (Figure 15.7).

Such extreme blinding serves our "prevention-focused" motive of discouraging research bias. But it works so extremely that seeing the blinded data is little different from seeing no data at all, which seems little different in practice from simply pre-registering one's hypotheses and data analysis plans.

But earlier we argued that data analysis serves the "promotion-focused" goals of encouraging creative thinking about one's study and the possible mechanisms at play in respondent cognition and behavior. Is there a way to stimulate such thinking while at the same time discouraging researcher bias? Our fifth method attempts to fit the bill.

Blinding method 5: Cell scrambling. This is the method we report in the main text. Rather than scrambling individual data points, our fifth method keeps each cell's data together, but it scrambles the identities of the four cells of the design. For our four-cell design, there are 4! = 24 possible orderings of the cells. Rather than sampling one such ordering, imagine that the investigator is given *a set* of, say, six of them. (Our intuition here is that there may be a cognitive sweet spot between providing only a single permutation and providing all 24; six seems like just enough to encourage hard thinking about the data.) Note that the true raw data have a chance (in this case, one-in-four) of appearing in the ensemble of cell-scrambled

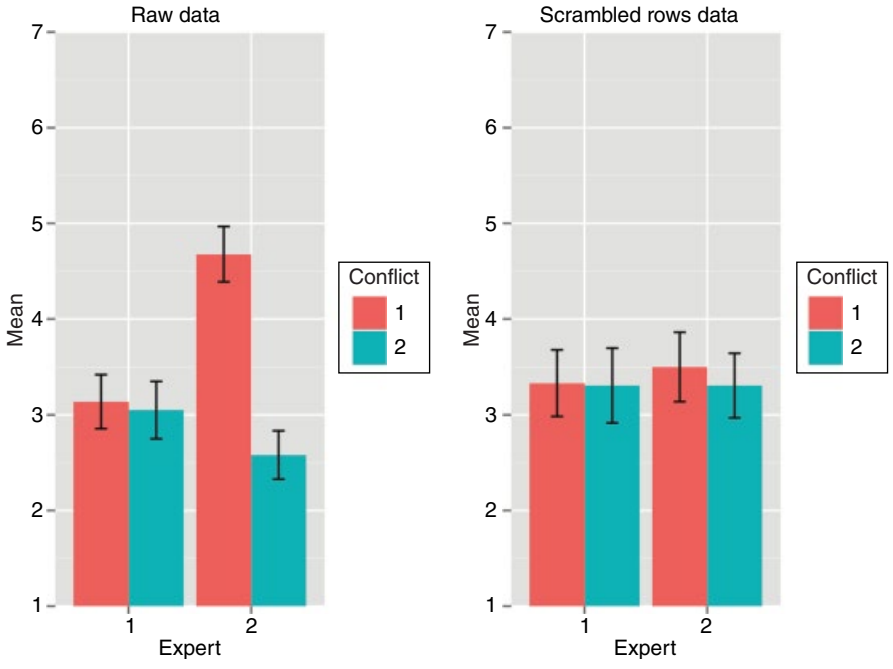

Figure 15.7 Raw means from Simulation 1 (left panel) and blinded means perturbed by "row scrambling" (right panel).

datasets. By coincidence, in this run, the very first scrambled set is in fact the true data set, though of course the investigator should not know that until the blind is lifted.

As seen in Table 15.1, cell scrambling influences the F statistics, but it does so in a different manner than the other methods. Note that all six cell-scrambled sets have the same three F statistics as the original raw data, so the investigator will know whether there are significant effects (and how many). But if only some of them are significant, the investigator will not know which ones have and have not crossed the $p < 0.05$ threshold.

A 2 × 2 Factorial with weaker effects and low power

In our second simulation, we tested the same blinding algorithms, but this time we reduced the cell 3 mean from 4.5 to 4, and we reduced the cell sizes from 50 to 25 – which, unfortunately, is closer to typical practice in psychology. As seen in Table 15.2, the raw data show no significant effects, though two of the three are very close to the $p < 0.05$ threshold (and prime candidates for p-hacking).

In this kind of situation, the regressive methods (adding noise and row scrambling) have little effect because we are so near the floor already. Cell scrambling retains the two marginal effects, but the investigator no longer knows which ones are

Table 15.2 Simulation 2: F Statistics.*

	Expert		*Conflict*		*Expert × Conflict*	
Raw data	0		3		3.4	
Raw + noise	0		3.5		1.8	
Raw + cell bias	23.4	***	59	***	45.5	***
Raw + noise + cell bias	7.1	**	5	*	46.1	***
Row scrambling	0.3		0		0.3	
Cell scrambling						
Set 1	0		3.4		3	
Set 2	3.4		0		3	
Set 3	0		3.4		3	
Set 4	3.4		3.4		0	
Set 5	3.4		0		3	
Set 6	3		0		3.4	

* Each effect has 1 df, and the error term has 196 df.

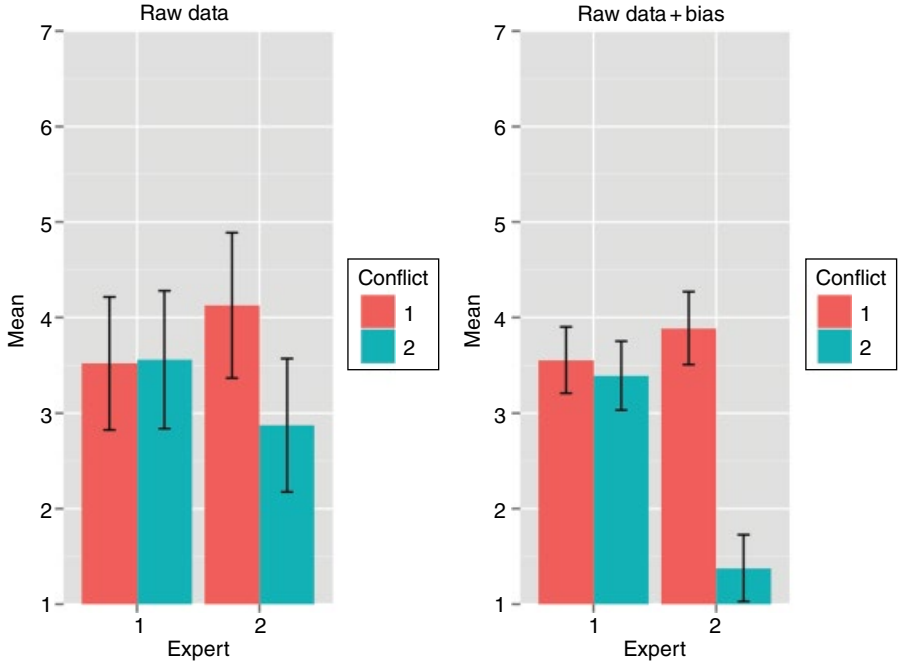

Figure 15.8 Raw means from Simulation 2 (left panel) and blinded means perturbed by cell-specific bias (right panel).

near threshold. As such, cell scrambling will not fully discourage *p*-hacking – though it will make it more difficult. But methods 2 and 3 – which perturb the data with cell-specific bias terms – serve to push all three effects well into the significant range. In this case, the investigator is now so beyond the significance threshold that there is little temptation to *p*-hack (Figure 15.8).

Endnotes

1 Yong (2012) provides a good overview. In-depth treatments appear in the symposia in *Perspectives on Psychological Science* on "Replicability in psychological science: A crisis of confidence?" (November 2012), "Advancing science" (July 2013), and "Advancing our methods and practices" (May 2014).

2 Over the long run, we also see evidence of a different problem: the most recent estimates tend to fall well outside many of the previous confidence intervals – a clear sign of judgmental overconfidence (see Henrion & Fischhoff, 1986).

3 This is similar to the notion of disconfirmation bias discussed earlier; there is a continuum anchored by "principled skepticism" on one end (where considerable prior evidence or well-tested theory argue against accepting a finding) and "motivated skepticism" on the other (where one simply does not like a finding).

4 See Einhorn (1972). The classic demonstration, using hypothetical data, is Armstrong (1967). Empirical examples are documented in Fabrigar, Wegener, MacCallum, and Strahan (1999) and MacCallum, Roznowski, and Necowitz (1999).

References

Anderson, C. A., & Anderson, K. B. (1996). Violent crime rate studies in philosophical context: A destructive testing approach to heat and southern culture of violence effects. *Journal of Personality and Social Psychology*, *70*, 740–756.

Armstrong, J. S. (1967). Derivation of theory by means of factor analysis or Tom Swift and his electric factor analysis machine. *American Statistician*, *21*, 17–21.

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science*, *9*, 333–342.

Bruner, J., & Potter, M. (1964). Inference in visual recognition. *Science*, *144*, 424–425.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd edn). Hillsdale, NJ: Erlbaum.

Conley, A. et al. (The Supernova Cosmology Project) (2006). Measurement of Ω_m, Ω_Λ from a blind analysis of Type Ia supernovae with CMAGIC: Using color information to verify the acceleration of the universe. *The Astrophysical Journal*, *644*, 1–20.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition*, *31*, 187–276.

Ditto, P. H., & Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology*, *63*, 568–584.

Duarte, J. L., Crawford, J. T., Stern, C., Jussim, L., & Tetlock, P. E. (2015). Political diversity will improve social psychological science. *Brain and Behavioral Sciences*, *38*, e130.

Edwards, K., & Smith, E. E. (1996). A disconfirmation bias in the evaluation of arguments. *Journal of Personality and Social Psychology*, *71*, 5–24.

Einhorn, H. J. (1972). Alchemy in the behavioral sciences. *Public Opinion Quarterly*, *8*, 367–378.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*, 272–299.

Faia, M. A. (2000). Three can keep a secret if two are dead (Lavigne, 1996): Weak ties as infiltration routes. *Quality & Quantity*, *34*, 193–216.

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PLoS ONE*, *5*, e10068, 1–10.

Fanelli, D., & Ioannidis, J. P. A. (2013). US studies may overestimate effect sizes in softer research. *PNAS*, *110*, 15031–15036.

Fang, F. C., Steen, R. G., & Casadevall, A. (2012). Misconduct accounts for the majority of retracted scientific publications. *PNAS*, *109*, 17028–17033.

Feynman, R. (1985). Cargo cult science. 1974 speech, reprinted in *Surely you're joking, Mr. Feynman!* New York, NY: W. W. Norton.

Fiedler, K., Kutzner, F., & Krueger, J. I. (2012). The long way from α-error control to validity proper: Problems with a short-sighted false-positive debate. *Perspectives on Psychological Science*, *7*, 661–669.

Fung, B. C. M., Wang, K., Chen, R., & Yu, P. S. (2010). Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys*, *42*, 1–53.

Furr, R. M. (2002). Psychology and astrophysics: Overcoming physics envy. *SPSP Dialogue*, *17*, 17–18.

Goldin, C., & Rouse, C. (2000). Orchestrating impartiality: The impact of "blind" auditions on female musicians. *American Economic Review*, *90*, 715–741.

Gross, N., & Fosse, E. (2012). Why are professors liberal? *Theory and Society*, *41*, 127–168.

Hedges, L. V. (1987). How hard is hard science, how soft is soft science? The empirical cumulativeness of research. *American Psychologist*, *42*, 443–455.

Henrion, M., & Fischhoff, B. (1986). Assessing uncertainty in physical constants. *American Journal of Physics*, *54*, 791–798.

Higgins, E. T. (1998). Promotion and prevention: Regulatory focus as a motivational principle. *Advances in Experimental Social Psychology*, *30*, 1–46.

Humphreys, L. G., Ilgen, D., McGrath, D., & Montanelli, R. (1969). Capitalization on chance in rotation of factors. *Psychological Measurement*, *29*, 259–271.

Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science*, *7*, 645–654.

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

Jacobs, R. C., & Campbell, D. T. (1961). The perpetuation of an arbitrary tradition through several generations of a laboratory microculture. *Journal of Abnormal and Social Psychology*, *62*, 649–658.

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

Jost, J. T., Rudman, L. A., Blair, I. V., Carney, D. R., Dasgupta, N., Glaser, J., & Hardin, C. D. (2009). An invitation to Tetlock and Mitchell to conduct empirical research on implicit bias with friends, "adversaries," or whomever they please. *Research in Organizational Behavior*, *29*, 73–75.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, *64*, 515–526.

Kashima, Y. (2014). How can you capture cultural dynamics? *Frontiers in Psychology*, *5*(955), 1–16.

Kashy, D. A., Donnellan, M. B., Ackerman, R. A., & Russell, D. W. (2009). Reporting and interpreting research in PSPB: Practices, principles, and pragmatics. *Personality and Social Psychology Bulletin*, *35*, 1131–1142.

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

Kerr, N., MacCoun, R. J., & Kramer, G. (1996). Bias in judgment: Comparing individuals and groups. *Psychological Review*, *103*, 687–719.

Klayman, J., & Ha, Y. W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*, 211–228.

Klein, J. R., & Roodman, A. (2005). Blind analysis in nuclear and particle physics. *Annual Review of Nuclear and Particle Physics*, *55*, 141–163.

Klein, R. A. et al. (2014). Investigating variation in replicability: A "Many Labs" replication project. *Social Psychology*, *45*, 142–152.

Kruglanski, A. W., & Webster, D. M. (1996). Motivated closing of the mind: "Seizing" and "freezing." *Psychological Review*, *103*, 263–283.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*, 480–498.

Lewin, K. (1931). The conflict between Aristotelian and Galileian modes of thought in contemporary psychology. *Journal of General Psychology*, *5*, 141–177.

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Consider the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology*, *47*, 1231–1243.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*, 2098–2109.

Lyons, L. (2008). Open statistical issues in particle physics. *The Annals of Applied Statistics*, *2*, 887–915.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*, 490–504.

MacCoun, R. (1998). Biases in the interpretation and use of research results. *Annual Review of Psychology*, *49*, 259–287.

MacCoun, R. J., & Paletz, S. (2009). Citizens' perceptions of ideological bias in research on public policy controversies. *Political Psychology*, *30*, 43–65.

MacCoun, R. J., & Perlmutter, S. (2015). Hide results to seek the truth. *Nature*, *526*, 187–189.

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy Research*, *1*, 161–175.

Marcus, G., & Davis, E. (2014, April 6). Eight (no, nine!) problems with Big Data. *New York Times*, A23.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*, 103–115.

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., Glennerster, R., Green, D. P., Humphreys, M., Imbens, G., Laitin, D., Madon, T., Nelson, L., Nosek, B. A., Peterson, M., Sedlmayr, R., Simmons, J. P., Simonsohn, U., & Van der Laan, M. (2014). Promoting transparency in social science research. *Science*, *343*, 30–31.

Navarro, D. J., & Perfors, A. F. (2011). Hypothesis generation, sparse categories, and the positive test strategy. *Psychological Review*, *118*, 120–134.Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*, 175–220.

Nosek, B. A. (2014, March). Improving my lab, my science, with the Open Science Framework. *APS Observer*, *27*, 12–15.

Olive, K. A., et al. (2014). Particle data group. *Chinese Physics C*, *38*, 090001.

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*, 528–530.

Payne, J. L. (1974). Fishing expedition probability: The statistics of post hoc hypothesizing. *Polity*, *7*, 130–138.

Platt, J. R. (1964). Strong inference. *Science*, *146*, 347–353.

Rabin, M., & Schrag, J. L. (1999). First impressions matter: A model of confirmatory bias. *Quarterly Journal of Economics*, *144*, 37–80.

Redding, R. E. (2012). Likes attract: The sociopolitical groupthink of (social) psychologists. *Perspectives on Psychological Science*, *217*, 512–515.

Sagarin, B. J., Ambler, J. K., & Lee, E. M. (2014). An ethical approach to peeking at data. *Perspectives on Psychological Science*, *9*, 293–304.

Saks, M. J., & Koehler, J. J. (2005). The coming paradigm shift in forensic identification science. *Science*, *309*, 892–895.

Schulz, K. F., & Grimes, D. A. (2002). Blinding in randomized trials: Hiding who got what. *The Lancet*, *359*, 696–700.

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.

Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological Science*, *24*, 1875–1888.

Snyder, M. (1984). When belief creates reality. *Advances in Experimental Social Psychology*, *18*, 247–303.

Stolberg, M. (2008). Inventing the randomized double-blind trial: The Nuremberg salt test of 1835. *Journal of the Royal Society of Medicine*, *99*, 642–643.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, *10*, 557–570.

Taleb, N. N. (2001). *Fooled by randomness: The hidden role of chance in life and in the markets*. New York, NY: Random House.

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist*, *24*, 83–91.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science*, *6*, 100–116.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, *4*, 274–290.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, *14*, 779–804.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*, 129–137.

Watson, J. D. (1968). *The double helix*. New York, NY: Atheneum.

Yong, E. (2012). Bad copy. *Nature*, *485*, 298–300.