Opportunistic Biases

Their Origins, Effects, and an Integrated Solution

Jamie DeCoster — *University of Virginia*
Erin A. Sparks — *Purdue University*
Jordan C. Sparks — *University of Minnesota*
Glenn G. Sparks — *Purdue University*
Cheri W. Sparks — *Indiana First Steps, Lafayette, Indiana*

Researchers commonly explore their data in multiple ways before deciding which analyses they will include in the final versions of their papers. While this improves the chances of researchers finding publishable results, it introduces an "opportunistic bias," such that the reported relations are stronger or otherwise more supportive of the researcher's theories than they would be without the exploratory process. The magnitudes of opportunistic biases can often be stronger than those of the effects being investigated, leading to invalid conclusions and a lack of clarity in research results. Authors typically do not report their exploratory procedures, so opportunistic biases are very difficult to detect just by reading the final version of a research report. In this article, we explain how a number of accepted research practices can lead to opportunistic biases, discuss the prevalence of these practices in psychology, consider the different effects that opportunistic biases have on psychological science, evaluate the strategies that methodologists have proposed to prevent or correct for the effects of these biases, and introduce an integrated solution to reduce the prevalence and influence of opportunistic biases. The recent prominence of articles discussing questionable research practices both in scientific journals and in the public media underscores the importance of understanding how opportunistic biases are created and how we might undo their effects.

Keywords: opportunistic, bias, replicability, methodology, statistics, exploratory and confirmatory analyses

He uses statistics as a drunken man uses lampposts . . . for support rather than illumination.

—Andrew Lang (1844–1912)

Analyzing data from real experiments almost always involves complexities that are not covered in the tidy examples contained in statistics textbooks. In an idealized textbook problem, the investigator has conducted a study with the goal of answering a single, well-defined research question. The primary challenge for students is to determine how to analyze the provided data in a way that will answer the question. In the real world, however, the typical study collects information related to a wide variety of research questions. The primary challenge for practicing analysts is to determine which of the relations found in the data can be presented together as part of a coherent narrative that will be accepted by an academic journal. As a result, practicing analysts typically conduct a much larger number of statistical tests and examine a much larger range of hypotheses than would a student solving a textbook problem. Conducting multiple tests gives analysts additional opportunities to find effects in their data, but it also makes it more likely they will observe significant results even when there are no real effects present. This creates an "opportunistic bias" in the results, so that published studies report larger estimates, smaller confidence intervals, and lower p values relative to unbiased versions of the same tests.

Opportunistic biases occur whenever researchers examine multiple analyses before deciding exactly which ones to present as part of a report. The selection process makes it more likely for the researcher to find significant results and larger effect sizes. For example, a researcher who examines the effect of a treatment on a single outcome is much less likely to find a significant result than a researcher who examines the effect of the same treatment on 10 different outcomes. Similarly, when trying to demonstrate that a treatment has large effects, the researcher examining 10 outcomes is much more likely to observe an effect of the desired magnitude than the researcher examining a single outcome. While this provides investigators with more opportunities to obtain publishable findings, these procedures will bias any estimated relations away from their true values. The ultimate result is that findings affected by opportunistic biases will typically be less likely to replicate, and will produce smaller effects when they do replicate.

Jamie DeCoster, Center for Advanced Study of Teaching and Learning, University of Virginia; Erin A. Sparks, Department of Psychological Sciences, Purdue University; Jordan C. Sparks, Department of Psychology, University of Minnesota; Glenn G. Sparks, The Brian Lamb School of Communication, Purdue University; Cheri W. Sparks, Indiana First Steps, Lafayette, Indiana.

We thank Marcello Gallucci, Jason Gitler, Chan Ha, John Greene, Saul Miller, Jim Tyler, and Steve Wilson for comments made on an earlier version of this article.

Correspondence concerning this article should be addressed to Jamie DeCoster, Center for Advanced Study of Teaching and Learning, University of Virginia, P.O. Box 800784, Charlottesville, VA 22908-0784. E-mail: jamied@virginia.edu

Vol. 70, No. 6, 499–514 http://dx.doi.org/10.1037/a0039191

Jamie DeCoster

The effects of opportunistic biases can be thought of as a specific example of "regression toward the mean," a phenomenon originally observed by Galton (1886). Any time that researchers choose a sample based on whether or not the individual subjects exceed a criterion score, statistics calculated from this sample will be biased estimates of the true population values. When those who exceeded the criterion initially are measured a second time, their scores will typically regress toward the mean (i.e., will be less extreme) because the bias is removed. Where does this bias come from? All measurements will necessarily be influenced by random factors, but we can expect that the average of a group of such measurements will typically provide an unbiased estimate of the true sample mean because the random effects will tend to cancel each other out (because we can expect them to be positive about half the time and negative about half the time). However, subjects chosen because they have particularly large or small values are more likely to have random factors aligning in a way to enhance their extreme nature. In this case, the random factors no longer cancel each other out (because random effects making the estimate more extreme will be more common than those making it less extreme), so that the mean of the observed measurements will typically be more extreme than the mean of the true values.

Most researchers are familiar with regression artifacts when selecting participants, but the same logic can also be applied to the selection of statistical tests for a research article. Like other estimates, the effects measured in research studies are influenced by random factors, so choosing effects because they are particularly large or because their tests have particularly low p values will typically identify effects that have random influences aligning in such a way to make them larger. The random influences will not typically align the same way in future studies, so the size of the effect in the initial study will typically overestimate the effect size that would be found in any replications. This overestimation is the opportunistic bias, and is equal to the difference between the observed effect and the true effect in the population.

Previous researchers have discussed procedures that create opportunistic biases using a number of different terms, including "snooping," "fishing," "hunting," and "data dredging" (Selvin & Stewart, 1966); "probability pyramiding" (Neher, 1967); "HARKing" (Kerr, 1998); and "p-hacking" (Simmons, Nelson, & Simonsohn, 2011). However, we prefer the term opportunistic bias to these other options for several reasons. First, unlike most of the prior terms, opportunistic bias is not specifically tied to the null hypothesis significance testing (NHST) paradigm. Given our desire to consider the effects of opportunistic biases across different analytic paradigms, such as Bayesian analysis and the use of confidence intervals, we believe that a paradigm-independent term is superior. Second, opportunistic bias reflects both the nature of the procedures we want to discuss (opportunistically looking for large or significant effects) as well as the effects of those procedures (biasing the results in the direction of the sought effects). Third, whereas most of the prior terms refer to probabilities, opportunistic bias can be used to refer to biased estimates, effect sizes, and confidence intervals. Finally, the prior terms have primarily been each applied to a specific source or a specific effect of opportunistic bias, and we wanted to have a more general term to discuss the entire collection of effects.

Procedures That Introduce Opportunistic Biases

There are many things that researchers can do when conducting studies and analyzing data that will lead to opportunistic biases. Some of these are performed so often that they are considered to be common practice, so that many researchers are unaware that the techniques are problematic. We would therefore like to describe some of the common practices that introduce opportunistic biases so that readers have a concrete understanding of the breadth of procedures that could bias results. We are not suggesting that researchers are intentionally using these techniques to amplify their results, nor even that these behaviors are necessarily unethical. In fact, we suspect that many researchers consider the procedures we discuss below to be a normal part of the theory-generation and data-exploration phase of research. However, these procedures all provide researchers with the opportunity to review multiple analyses before deciding which will be the focus of a paper or presentation, which introduces an opportunistic bias.

Measure a Large Collection of Variables and Only Report Desirable Results

Measuring a host of variables related to the topic of interest allows researchers to select their reported results from a large number of different tests. The variables can represent

Erin A. Sparks

different indicators of the same construct, the same measures collected at multiple time points, or measures of entirely different constructs. Using multiple variables can quickly escalate the number of tests that may be performed. Given a set of P predictor variables and a different set of O outcome variables, analysts have a total of $P \times O$ bivariate relations from which to select the results for their report. As an example, a researcher examining whether four demographic variables (age, sex, race, and income) were related to the Big 5 personality traits (Costa & McCrae, 1992) would have the opportunity to consider 20 different tests. If, instead, the analyst considers the relations found among a single set of T variables (without distinguishing predictors from outcomes), there are a total of $T(T - 1)/2$ unique bivariate relations to be explored.[1] An example of this would be a researcher examining whether the 11 clinical subscales of the Personality Assessment Inventory (Morey, 1991) were related to each other. This analysis would involve exploring 55 unique correlations. When researchers perform a large number of analyses and then choose to focus only on those representing the largest effects, their conclusions are subject to opportunistic biases.

Examine Different Ways of Transforming Variables

Researchers will sometimes change the way an existing variable is included in the statistical model if analyses of the variable in its original form do not produce the desired results. Those working with continuous variables sometimes perform inverse, logarithmic, square root, and other transformations on either the predictor or outcome variables. They can also examine the performance of the variables in their original form as well as after artificial dichotomization. Those working with categorical variables sometimes combine a set of groups into a single category after observing that the groups have similar means on important outcomes. Those working with scale composites sometimes redefine which items are included in the composites or how much weight is given to each item when the original composite performs poorly. Should researchers use these procedures to analyze their data in multiple ways before deciding how they will define their variables, taking advantage of the additional opportunities to obtain desirable findings will introduce a bias into their results.

Examine the Same Hypothesis Using Different Analyses

There is rarely just a single way to analyze a given research question. It is typically up to the researcher to determine which approach to take and to provide a justification for that approach. For example, if a researcher wants to compare two groups on a variable that is somewhat skewed, they could justify using a t test by citing the fact that it is robust to violations of normality, or they could justify using a Mann–Whitney U because it does not assume normality. Sometimes researchers take advantage of this flexibility to explore different analytic methods to find the one providing results most consistent with their hypotheses.

In addition to varying the analytic methods, researchers can also vary the statistical models they examine. When trying to explain an outcome variable, researchers can add and subtract different variables from their list of predictors to provide them with more opportunities to find desirable results. Given a set of P predictors, a total of $2^P - 1$ different models can be created just including main effects. This means that seven different models can be created from three predictors, 31 different models can be created from five predictors, 255 different models can be created from eight predictors, and 32,767 different models can be created from 15 predictors. There are also a total of $P(P - 1)/2$ two-way and $P(P - 3)/2 + 1$ three-way interaction effects that can be added to the P main effects, so researchers interested in examining moderating variables have an even greater number of models to explore.

A related issue, discussed by MacCallum, Roznowski, and Necowitz (1992), is the examination of multiple related models in covariance structure modeling. When performing path analysis or structural equation modeling, researchers receive statistics indicating whether or not the proposed model fits the observed data. Because conclusions cannot be easily drawn from models with poor fit, the common practice is to perform a "specification search" to find a model with good fit by adding paths to the model that improve its ability to explain the observed data (Leamer, 1978). However, MacCallum et al. (1992) indicates that tests of the final models chosen in this way capitalize on

[1] This differs from the prior equation because we have to exclude correlations of a variable with itself, and we must recognize that the correlation of variable X_1 with variable X_2 is the same as the correlation of variable X_2 with X_1.

Jordan C. Sparks

chance, so these models do not extrapolate to other samples as well as models that have not been modified using a specification search.

Examine the Same Hypothesis in Different Subgroups of Participants

If the desired results are not found in the data set as a whole, researchers can restrict the sample in different ways until they find a subsample providing the results they want. Each level of each grouping variable provides an additional opportunity for the analyst to examine the hypotheses of interest. In addition, analysts can vary the variables and the models examined in each subgroup, multiplying the total number of tests explored by the number of subgroups that are examined. For example, a researcher exploring whether there are gender or race effects on a particular relation could examine the relation within males, within females, within Blacks, within Whites, within Black males, within Black females, within White males, and within White females, giving them many opportunities to find the desired effect. If the researcher also considers multiple predictors, they can separately examine each of the possible models in each of the eight subgroups, providing even more opportunities to find the desired effect. Split enough ways, eventually the effect of interest will be significant somewhere, although findings uncovered using this procedure are likely to be spurious and unlikely to replicate. Although the use of an omnibus interaction test can reduce the inflation of Type I errors associated with exploring the effects within the different levels of a grouping variable, opportunistic biases will be introduced if researchers explore and discard the effects of multiple grouping variables, or if the subgroups are examined individually outside of the context of an interaction effect.

In a related way, researchers often have flexibility with regard to how they choose to handle outliers, which could be strategically used to help them obtain desirable effects. By definition, outliers are observations whose patterns of values substantially differ from others in a data set. Simple examples might have extremely high or extremely low values, but it is also possible to have multivariate outliers whose patterns are inconsistent with the correlations among the variables. It has been well established that outliers can drastically change the characteristics of an estimated regression line (Cohen, Cohen, West, & Aiken, 2003), and so it is important that researchers handle them intelligently. The most appropriate way to handle an outlier depends on why the observation is unusual (Barnett & Lewis, 1994). Sometimes outliers represent errors or observations that are not from the intended population, in which case they should be corrected or removed. Other times, outliers are valid but unusual cases, the handling of which requires judgment on the part of the researcher. The way that researchers choose to handle outliers can potentially introduce opportunistic biases into their results. A minor example would be when researchers first examine the results in the full data set, and then only choose to look for outliers when the initial results are not promising. A more serious example would be when researchers allow the knowledge of how the inclusion of a particular case would affect their results to influence whether or not they decide to keep that case in their sample.

Conduct Studies Examining the Same Hypothesis Using Different Methods

Researchers testing a relation can conduct multiple studies, each assessing or manipulating the variables in a slightly different way. Modifications can be made following each version of the study until a methodology producing large, significant effects is discovered. Assuming that the prior versions used theoretically valid implementations of the variables, each of the methodologies provides an equally valid estimate of the relation, so selectively reporting the one with the largest effect likely overestimates the strength of the relation. In addition, choosing to stop changing the experimental method only when a strong finding is observed makes the final method susceptible to biases arising from chance.

Scrutinize Undesirable Findings More Closely Than Desirable Findings

Occasionally researchers will observe a finding that runs counter to their expectations and hypotheses. In this case, they will often "double-check" the analytic procedure for the unexpected finding to see if it was the result of a statistical error, failed assumptions, or the presence of outliers. While the logic behind this check is reasonable, looking for errors when there are undesirable findings more often than when there are desirable findings systematically biases researchers' results in a way favoring the researchers' expectations. This process would prevent erroneous results that are contrary to the researcher's expectations, but would allow erroneous results that are consistent with

Glenn G. Sparks

the researcher's expectations. Ideally, researchers should double-check all of their results to prevent the introduction of this bias.

Collect Data Until Desirable Results Are Found

Researchers working within the NHST framework will often collect additional data when their tests provide low *p* values that are not quite significant. They will often justify this by saying that the purpose of the additional data collection is to "clarify" the marginally significant finding so that it is moved either past the threshold for significance or far enough away from the threshold that the researcher believes that collecting even more data will not likely produce a significant result. The problem with this procedure is that researchers will typically only use it when the initial result is above the threshold and not when it is below. The unbalanced nature of this correction will systematically bias researchers' results in a way favoring their expectations.

Prevalence of Opportunistic Biases

The scientific importance of our discussion depends on the degree to which opportunistic biases have influenced the published literature. We mentioned a number of different ways that researchers *can* bias their results, but so far have not discussed how often they *do* bias their results. The nature of opportunistic biases is such that they cannot be detected simply by reading research reports, as the selection of which effects to examine must necessarily occur before writing up a study. John, Loewenstein, and Prelec (2012) therefore conducted a survey of 2,155 psychologists to determine the prevalence rates of a number of behaviors that can lead to opportunistic biases and how these rates might vary across subgroups of researchers.

In their survey, John et al. (2012) asked participants to anonymously report whether they had personally engaged in 10 different questionable research practices, the proportion of their colleagues that they believed had engaged in the practices, and the percentage of their colleagues that they believed would admit to engaging in the practices. Obtaining information on the likelihood that their colleagues were to engage and report engaging in these behaviors allowed the authors to use Bayesian techniques (based on Prelec, 2004) to estimate the true prevalence of questionable research practices. Eight of these practices would be categorized as procedures that would create opportunistic biases, whereas the other two asked about instances of deliberate deception (intentionally misreporting findings and falsifying data). The raw prevalence rates for practices that create opportunistic biases ranged from 15.6% (for stopping data collection early once the desired result was obtained) to 63.4% (for failing to report all of the outcome variables collected as part of the study), with the Bayesian-corrected rates being substantially higher, reaching 100% in several cases. These authors provide compelling evidence that procedures that create opportunistic biases are commonly performed, suggesting that we can expect that opportunistic biases have had a substantial influence on the results published in our journals. Given that the majority of the estimated prevalence rates were greater than 50%, it appears that procedures creating opportunistic biases are part of the socially accepted norms for researcher behavior in psychology.

The Effects of Opportunistic Biases

John et al.'s (2012) survey indicated that procedures leading to opportunistic biases are commonly employed throughout the different areas of psychology. This naturally leads to the question of how the prevalence of these biases has affected researchers and their work. In this section, we would like to consider how opportunistic biases affect the interpretation of statistics calculated within the classic NHST paradigm, how they also affect statistical analysis using alternative paradigms like Bayesian analysis, how they affect the ability of researchers to develop generalizable theories, and how they affect the perception of scientific research.

Effects in NHST

The most direct and obvious effect of opportunistic biases is that the *p* values presented for hypothesis tests can no longer be interpreted as they should be (i.e., the probability of obtaining the results at least as extreme as those observed in the study due to chance alone). The widespread use of opportunistic biases suggests that, for a given study, the actual probability of obtaining the reported results if the null hypothesis was true is substantially higher than the reported *p* value. Exactly how much higher depends on how many different analyses were considered and then discarded without being reported in the final paper. Each extra test will increase the likelihood that the observed

Cheri W. Sparks

significant results are actually just due to chance alone. Specifically, if a researcher performs T independent tests using a confidence level of α, the probability of obtaining at least one spuriously significant result is equal to $1 - (1 - \alpha)^T$, which can be much higher than the original α (Abdi, 2007). According to this equation, conducting two tests leads to an error rate of .0975, conducting four tests leads to an error rate of .1855, conducting 10 tests leads to an error rate of .4013, and conducting 50 tests leads to an error rate of .9321. Given that researchers commonly explore the effects of multiple predictors on multiple outcomes in their studies, the probability that a reported "significant" finding is actually due to chance alone can often be quite high.

We are by no means the first to consider the problems that opportunistic biases create for the interpretation of statistical results within NHST. One of the most thorough treatments of this issue was presented by Ioannidis (2005), who calculated the probability of reported significant effects actually being true (rather than resulting from Type I errors) under a variety of different circumstances. Monte Carlo simulations demonstrated that significant results were less likely to be true when researchers probed a greater number of hypotheses, when the study had a smaller sample, when the effects being investigated were smaller, when the study was in a field that used a larger variety of methods, when there were financial interests at stake, and when the study was on a "hot" topic being investigated in a large number of different laboratories. The author then used the simulation results to estimate the likelihood that the results from different types of studies actually reflected true effects. These estimates ranged from .85 for adequately powered, unbiased clinical trials and meta-analyses of clinical trials to .001 for exploratory studies lacking prior theory that investigated a large number of different effects. Overall, this analysis suggests that significant results represent true effects between 20% and 40% of the time. Dishearteningly, Ioannidis (2005) concluded that "most claimed research findings are false" (p. 696) and that the estimates provided by studies "may often be simply measures of the prevailing bias" (p. 700).

Within psychology, Simmons et al. (2011) similarly discussed how "researcher degrees of freedom," the different choices that researchers make when deciding what to include in a research report, can substantially increase the likelihood of false-positive results within the NHST framework. They argue that the large number of ways that researchers may choose to report their studies allows almost any hypothesis to be reported as significant using $\alpha = .05$, no matter what the truth happens to be. They suggest that these effects are worst when researchers conduct studies with small sample sizes, because this allows them to conduct more studies on each topic, providing more opportunities to capitalize on opportunistic biases.

Effects in Other Statistical Paradigms

Although most of the criticism against opportunistic biases has focused on how they influence p values, their statistical effects are not limited to studies conducted within a NHST framework. Any analytic system where researchers can run multiple analyses and then choose what they want to report will have its results influenced by opportunistic biases. As examples, consider the use of confidence intervals and Bayesian inference, two commonly offered alternatives to NHST. Confidence intervals are calculated from many of the same statistics as hypothesis tests, but present the results in a way that focuses attention on the estimates calculated in the data and provides readers with a range of likely values for population parameters (Cohen et al., 2003). Similarly, Bayesian inference allows researchers to combine their prior knowledge with data observed in a study to estimate the "posterior probability distribution" of a parameter, which again focuses readers on the estimates and provides a range of likely values for population parameters (Congdon, 2006). These methods have been considered to be less susceptible to opportunistic biases because they do not have a specific pass/fail criterion like the significance level in NHST, reducing the need for researchers to bias their findings just to make an article acceptable for publication (Cohen, 1994).

Even within these frameworks, however, larger estimates will typically be seen as more important and interesting, which will provide a motivation for authors to selectively report variables and use methods that produce the strongest results. In addition, researchers conducting studies to test their own theories can selectively choose to report whichever results provide the best support for their ideas. If they want to claim a relation is present, they can search for methods and variables producing larger effects; if they want to discredit a relation, they can search for methods and variables producing smaller effects. Using these alternative frameworks will therefore change the motivation guiding the biases (i.e., having larger/smaller rela-

tions instead of lower *p* values), but will unfortunately not prevent people from biasing their results or make the influence of these biases any easier to detect.

This highlights an important point: Opportunistic biases will not only influence the conclusions that we draw from our analyses, but will also influence the size of the reported effects. If researchers systematically choose to report findings showing larger effects and avoid reporting findings showing smaller effects, we can expect that the literature as a whole will overestimate the magnitude of the relation being investigated. Researchers conducting subsequent studies in the area must then expect that their own effect sizes will regress toward the mean (Galton, 1886) and therefore have smaller magnitudes than those found in the literature. Power analyses and sample size calculations using effect sizes drawn from published articles will therefore underestimate the number of required participants for new studies, potentially causing researchers to waste resources on studies that are unlikely to produce significant results. Additionally, the presence of bias in the effect sizes means that meta-analyses will also be systematically overestimating the magnitude and importance of effects in their literatures (Rosenthal, 1979).

Effects on Developing Theories

Numerous problems arise when working in a field where the published results are commonly affected by opportunistic biases. Perhaps most obvious is that investigators will often waste their time trying to confirm or extend invalid findings. Given the possibility that the fault could be with the new study and not the original result, investigators may need to conduct an entire series of studies using different methods before they come to the realization that the conclusions of an initial article were incorrect. In addition, the prevalence of invalid results lowers the confidence that investigators can have in published findings. Many of the premier psychology journals rarely publish articles that do not replicate the original study multiple times, partially to counter the possibility that the original finding was just a chance result. Finally, the fact that we are less confident in each published study means that scientific progress in the field will be hindered, as we cannot generalize as strongly from each published article. If we could be more confident in the findings, we would be able to advance more quickly, with less effort devoted to checking and replicating earlier discoveries.

Effects on the Perception of Research

The prevalence of opportunistic biases and invalid findings has negatively affected the way that the public regards research. Doubt is often cast on the validity of research findings because the investigators are seen to be motivated by political, economic, or social agendas. This is at least partly because the investigators are given a great deal of latitude in deciding exactly what aspects of their research they will report and how they will present them, allowing the investigators to tailor their conclusions. Individual scientists biasing their results for short-term gains have accumulated long-term costs by damaging the reputation of scientific investigation as a whole. While all scientific disciplines are subject to negative consequences as a result of opportunistic biases, social scientists may be particularly likely to publish positive results stemming from "conscious and unconscious biases," as compared to the biological and physical sciences (Fanelli, 2010).

Opportunistic biases cannot only influence the way that nonscientists view research, but may also have detrimental effects on the attitudes of practicing scientists. Although it is certainly true that not all successful scholars bias their research, academic success and prestige can be influenced by the ability to strategically bias research findings as much as actual scientific ability. Researchers who are more careful about avoiding bias may become less competitive academically, which can demoralize and drive away the very people who are most likely to make unbiased contributions to the literature. In addition, when researchers must try multiple analyses on their data before obtaining publishable results, their own opinions about their theories will likely suffer. The commonly perceived need to bias research findings prior to publication can cause scholars to become jaded and pessimistic about their own and others' work.

Accommodations and Adjustments for Opportunistic Biases

Although the procedures that create opportunistic biases are widely used, scientists have not all ignored the problems that these biases create. Methodologists have proposed a number of solutions to either reduce the prevalence of procedures that create opportunistic biases or else protect the field from the erroneous conclusions that can result from their use. In the sections below, we review 12 suggestions that we believe successfully reduce the prevalence or impact of opportunistic bias, and two that have been proposed that we believe do not.

Replication

The most straightforward way to determine whether a particular finding is the result of opportunistic biases is to determine whether the results reappear in replications of the original studies. The truest test of whether a finding is valid is to see if it consistently appears where it is expected. Many statisticians and methodologists have emphasized the importance of replication to ensure that observed findings are not the result of random error (e.g., Shaver & Norton, 1980; Cohen, 1994; Schmidt, 2009), and some have specifically discussed its importance for protecting the field from experimenter biases (e.g., Valentine et al., 2011). Replicated studies have many of their features defined ahead of time by the original study, reducing the flexibility researchers have to perform multiple analyses, and correspondingly reducing their ability to take advantage of opportunistic biases.

In an effort to highlight the importance of successful replication, the Center for Open Science has sponsored a large-scale study designed to "estimate the reproducibility of published psychological science" (Reproducibility Proj-

ect & the Psychology Group, 2014). The goals of the project include identifying the rate that effects in the psychology literature can be replicated and identifying obstacles that prevent effective replications from taking place. This center has also sponsored the Many Labs Replication Project, an effort to conduct replications of 13 important effects in psychological science with over 6,000 participants to determine how replicability is impacted by sample and setting (Many Labs Replication Group, 2014). Recently, *Perspectives on Psychological Science* announced that it will begin publishing a new type of article, the Registered Replication Report. This change was implemented to encourage researchers to conduct high quality, multilab replications of important findings (Association for Psychological Science, 2014).

Distinguish Confirmatory Analyses From Exploratory Analyses

Several researchers have called for a greater emphasis on the distinction between exploratory and confirmatory analyses (e.g., Simmons et al., 2011; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). Exploratory analyses are helpful for generating ideas about future research directions and determining the feasibility of various designs, while confirmatory analyses are designed to systematically test specific, theory-driven hypotheses. The classic interpretation of a *p* value is based on the assumption that the test was used to investigate a confirmatory hypothesis. When *p* values are calculated from exploratory studies, the appropriate interpretation of the *p* value is less clear, as it will be influenced by opportunistic biases. However, it is clear that a confirmatory test provides more support for a hypothesis than an exploratory test with the same *p* value.

Although Simmons et al. (2011) and Wagenmakers et al. (2012) have provided specific guidelines to distinguish confirmatory from exploratory research, the field has not yet converged on a single definition of confirmatory research, nor has there been a consensus on how confirmatory studies should be treated compared to exploratory studies. Although it is clearly important that researchers be able to identify whether particular studies are confirmatory or exploratory, it is just as important that the field develop guidelines to help researchers appropriately draw conclusions when reviewing a literature containing a mix of these studies with varying results.

Arguing for the distinction between confirmatory and exploratory research is not a call to end exploratory research. Instead, it is a suggestion that the exploratory or confirmatory nature of a research finding be identified more clearly. Many journals do regularly publish exploratory findings, especially when the findings are novel or provide unique insights into a phenomenon. Unfortunately, even in these cases, editors and reviewers will still commonly request that authors present the research as if it derived from a confirmatory hypothesis. This advice, echoed in manuals for academic practice such as *The Compleat Academic* (Bem, 2003), is typically given to help authors make the presentation of their studies simpler and to make the findings easier to digest. However, it misrepresents the actual process by which a finding was obtained, and thereby misrepresents the confidence readers should place in the finding. As an alternative, we would suggest that researchers be asked to accurately describe their research as confirmatory or exploratory, and that the publication system provide outlets for both types of research. In such a revised system, journals could differ in their emphasis on exploratory and confirmatory research, imposing different standards for publishing exploratory and confirmatory findings.

Grassroots Efforts by Authors to Increase the Transparency of Their Research

Some researchers have argued that opportunistic biases should be combated by authors voluntarily choosing to improve the transparency of their own research. These efforts could range from coordinated efforts to voluntarily share data and study materials to increased reporting of methodological decisions and analyses. The Center for Open Science has recently developed the Open Science Framework (OSF) to facilitate this process via open collaboration (http://openscienceframework.org, Spies & Nosek, 2014). The OSF is a Web-based infrastructure in which scientists can publically distribute a variety of research materials. The OSF was created based on the idea that open collaboration helps to address problems in the research process more efficiently and comprehensively. It also provides a venue for researchers to report their confirmatory hypotheses and to share details about their study materials, data, and analyses.

Simmons, Nelson, and Simonsohn (2012) suggest that in the absence of journal-imposed disclosure requirements, authors can increase transparency by proactively labeling their research with a simple 21-word statement in the Method section: "We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study" (p. 1). On November 17, 2012, Etienne LeBel and colleagues (2013) drew upon Simmons, Nelson, and Simonsohn's (2011, 2012) work and launched PsychDisclosure.org, an open science initiative that gives authors a way to voluntarily disclose issues related to excluded subjects, unreported conditions and measures, and sample size determination. Almost 50% of authors randomly contacted by this organization to provide a voluntary disclosure of such issues responded (LeBel et al., 2013).

Develop and Enforce Better Reporting Standards

Some researchers have suggested that problems with the replicability of research could be solved by having more rigorous and thorough reporting standards for journal articles. The most recent edition of the *Publication Manual of the American Psychological Association* (APA, 2010) now explicitly suggests that authors adhere to the APA reporting guidelines, and a comprehensive book has been released to guide researchers in this adherence (Cooper, 2010). The guidelines include several recommendations that duplicate the solutions we mention above, including reporting the

intended sample size and the actual sample size, describing all the variables included in the study, identifying which analyses were "prespecified" and which were exploratory, and discussing the potential impact of "ancillary analyses for statistical error rates" (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008, p. 843).

Have Journals Require Increased Disclosure

Several researchers, such as Simmons et al. (2011) and Wagenmakers et al. (2012), have argued for disclosure guidelines that will allow readers to accurately assess the degree to which studies may be influenced by opportunistic biases. Building upon these suggestions, some journals have begun to require more comprehensive reporting about methodological and analytic procedures, in some cases requiring authors to provide data files or study materials. Within psychology, the newly formed APA journal *Archives of Scientific Psychology* has instituted requirements for openness that include having researchers submit a log of all performed analyses (Cooper & Vandenbos, 2013). *Psychological Sciences* has modified their manuscript submission procedure so that authors must now confirm that all excluded observations and reasons for exclusion have been reported, that all independent variables or manipulations have been reported, that all dependent variables or measures that were analyzed for the target research question have been reported, and that rules for determining sample size and stopping data collection were reported (Eich, 2014). *Psychological Sciences* has also implemented a "badge" system developed by Pierce et al. (2013). Manuscripts accepted for publication on or after January 1, 2014, can earn open data, open materials, and preregistered badges that will be documented on published versions of the manuscripts to highlight articles that have taken these additional steps to ensure the integrity and openness of their research. Other journals, such as the *Journal for Experimental Social Psychology* (2014), have begun to follow suit by revising their submission guidelines in attempt to encourage procedures that would reduce the presence and impacts of opportunistic biases.

Reduce the Bias Against the Null Hypothesis

Some of those concerned about how NHST encourages opportunistic biases have suggested that the main problem lies in people's misuse of the framework, not the framework itself. Specifically, a number of researchers have argued that journals should be more willing to publish null results. Different authors have argued for specific ways in which both researchers and editors can contribute to creating an environment in which a bias against the null hypothesis will be reduced. For example, Cohen (1962) suggested that researchers should conduct studies with greater power to insure that null results would be taken more seriously by editors. Hays (1963) suggested that simply including estimates of the magnitude of association would potentially make null findings more valuable. More recently, van Assen, van Aert, Nuijten, and Wicherts (2014) conducted a simulation study of selective publishing. They argued that meta-analytic results are more precise with a "publish everything" approach and further argued that selective publishing is less preferable due to the cost-benefit and time considerations involved. Most notably, Greenwald (1975) provided numerous suggestions for both researchers and editors to change attitudes toward the null hypothesis. In particular, he suggested that editors need to be willing to publish null results if the researchers take steps to make the null easier to accept gracefully, such as by selecting their sample size in advance based on a desired standard error for their estimate, presenting compelling evidence that all manipulations and measures are valid, presenting posterior probabilities of the null hypothesis, and reporting all results from the study. Laws (2013) echoed Greenwald's suggestions for multiple changes, arguing that reviewers, psychologists, and editors have "conspired to deny the existence of negative results" (p. 7).

Submit Research Designs for Publication Before Results Are Known

Some researchers have suggested that journals should evaluate submissions prior to the collection of any data. Under this system, research proposals would be accepted or rejected based on the soundness of the theory, methods, measures, and proposed analyses, much like when submitting a grant proposal or defending a dissertation prospectus. Once the initial plan is accepted, the journal would agree to publish the article no matter what results were found. In this system, researchers would not be implicitly motivated to engage in the strategies to ensure their results met certain criteria because any results would be publishable. There would also be no possibility of selectively presenting results because researchers would need to adhere to the originally proposed analysis plan.

Walster and Cleary (1970) argued for the acceptance of articles prior to data collection as a way to reduce bias against the null hypothesis, whereas Kupfersmid (1988) argues for this method as a solution to the file drawer problem. Walster and Cleary (1970) point out that while many problems plaguing the field have certainly been raised within the context of NHST, fewer people have focused on editorial policies as the source of the problems. They argue that if editors required researchers to submit articles for review before collecting data, no research effort would be wasted and biasing techniques that emerge during data collection and analysis would be eliminated.

Increase the Sample Size and Power of Studies

Simmons et al. (2011) noted that studies with fewer participants provide more "researcher degrees of freedom" because they allow investigators to conduct more studies given a fixed total number of participants, providing additional opportunities to take advantage of opportunistic biases. Schimmack (2012) also makes the point that researchers conducting multiple small studies may fail to report studies with nonsignificant findings, or may simply look for common findings among the set of studies and then present those as the predicted hypotheses, rather than allowing the

follow-up studies to act as true replications. Button et al. (2013) echo these concerns, noting that small studies tend to be accompanied by publication bias, selective reporting of outcomes, and worse design quality. These researchers all suggest that researchers should justify their choice of sample size in their articles, showing that the study as it was designed would have been powerful enough to detect the expected effects. This would help ensure that the studies have a reasonable chance to obtain significant results for the predicted hypotheses, which should reduce the need to use opportunistic biases to alter the results.

Use Sequential Analyses Instead of Informal Stopping Rules

In the section on procedures that introduce opportunistic biases, we mentioned that researchers inappropriately increase their likelihood of obtaining significant results if they extend data collection when they obtain marginally significant results. However, there are procedures that allow researchers to systematically examine their results throughout the course of data collection, possibly stopping the study if the results are particularly clear, without biasing their results. These procedures are called "sequential analyses." Compared to standard, fixed-sample analyses, sequential analyses typically lead to a substantial reduction in the number of observations required to reach a conclusion, making them preferable when data collection is costly or there is a need to increase the speed with which hypotheses are being tested. Some of these procedures allow researchers to perform standard hypothesis tests but adjust the sampling procedure and αs used by the tests (e.g., Frick, 1998; Botella, Ximénez, Revuelta, & Suero, 2006), whereas others involve tests specially designed for sequential analysis (e.g., Wald, 1947; Pocock, 1977).

Use Statistical Analysis to Detect Biased Results

Meta-analysts have long used "funnel plots" (Light & Pillemer, 1984) to help determine whether the reported distribution of effect sizes suffers from publication biases. More recently, researchers have proposed updated analytic methods to assess whether there is an excess of significant findings in an article or in the field as a whole due to any reason (such as publication bias, data fabrication, and selective reporting). Ioannidis and Trikalinos (2007) provide an "exploratory bias test" that determines whether the number of significant findings in a literature is statistically different from what we would expect if the results were unbiased, given the power of the studies in the literature. Applying their method, they observed evidence of bias in six of eight large-scale meta-analyses of clinical trials taken from the Cochrane Library. This is particularly noteworthy, given that the Cochrane Library specifically requires that their reviews meet stringent methodological standards so that they may be confidently used for medical decision making. Simonsohn, Nelson, and Simmons (2014) have instead suggested the use of a "*p* curve," which is designed to analyze the distribution of *p* values present in a particular group of findings to assess whether the set of findings has evidential value. Such a test could be applied to a group of findings from different authors or a particular set of studies from one author.

Although statistical methods of evaluating bias within a literature can be a valuable means to understand the trust that should be placed in a set of results, we have some concerns that these methods might be used to attack the work of individual researchers. It is important that any personal investigations of bias or fraud be based on multiple studies and a large amount of data to help ensure that the observed patterns are not coincidental. There is also a danger of those doing the investigations specifically targeting researchers with whom they have personal conflicts or whose theories conflict with their own. Although any suspected cases of academic misconduct should certainly be investigated, if those coming under suspicion are a nonrepresentative subset of scientists, then the investigations themselves could introduce a bias into the literature because invalid studies are more likely to persist among groups and individuals that have not been targeted for investigation.

Educate Researchers About Opportunistic Biases

One strategy for preventing opportunistic biases is to create a comprehensive education program that could make researchers more aware of problematic techniques and solutions. This program could include methodological curricula, chapters in methods and statistics texts, workshops at conferences, statements designed to highlight the problem in publication manuals, and editorial statements in journals. Most recently, the Society of Personality and Social Psychology e-mailed its members to announce the formation of a major task force to ". . . examine ethical conduct within the field, including what can be done to uncover misconduct, how the field can be more confident about the veracity of collected data, how training within the field can enhance ethical behavior, and how we can generally promote social and personality psychology as a credible scientific endeavor" (Devine, 2012). The need for the task force arose from recent cases of falsified data, but the scope of their charge includes the more general issue of opportunistic biases insofar as it seeks to explicate how the field can be more confident in the results reported in published studies.

Have Opinion Leaders Exert Authority

Many writers have issued a call for those who wield authority in deciding what articles are published and what grants are funded to adopt clear guidelines to reduce opportunistic bias in psychological research (e.g., Fidler, 2002; Hubbard & Ryan, 2000; Robinson & Wainer, 2002). Those in such positions have the greatest power to change perceptions of what practices are acceptable to the field and directly reduce the prevalence of opportunistic biases in the literature. For example, prominent researchers can encourage their colleagues and students to distinguish confirmatory and exploratory research, emphasize the importance of this distinction when reviewing journal articles, and provide examples of how to do this in their own work. Simi-

larly, should APA require that graduate methodology courses include a section discussing the importance and proper procedures for conducting replications, we can expect to see a larger presence of such articles in the future.

Stop Using NHST

As we have noted above, several authors have suggested that opportunistic biases are naturally rooted in the NHST framework. One might therefore try to reduce the prevalence of these biases by moving the field to alternative analytic methods. Authors such as Dienes (2011); Wetzels et al. (2011), and Kruschke (2011) suggest Bayesian inference is more coherent than NHST, Bayes factors can actually provide evidence in favor of the null hypothesis, and Bayesian inference is much richer than NHST. In keeping with this perspective, Johnson (2013) suggests that whenever possible, the Bayes factor should be reported instead of the corresponding *p* value. He further suggests that within the NHST framework, statistical significance should only be associated with *p* values less than .005. Some, such as Goodman (1993), have argued that significance testing should be abandoned in favor of techniques such as using mathematical likelihood to express evidential strength. Others (e.g., Hubbard & Armstrong, 1997; Natrella, 1960) have suggested the use of confidence intervals around point estimates to assess results for individual studies. Cumming (2014) recently argued for the use of a "new statistics" that focuses on practices based on effect sizes, confidence intervals, and meta-analysis.

Although we agree that the reliance of NHST on a single decision point (i.e., the critical value α) does encourage researchers to take advantage of opportunistic biases, specific decision points can be found in other analytic systems as well. Any time that the acceptability of a finding is based even partially on a specific criterion, whether that be a *p* value, a posterior probability, or an effect size, researchers have an obvious motive to preferentially report analyses that meet that criterion. In addition, there will always be a motivation for researchers to report the results most consistent with their own theories, even in the absence of a statistical criterion for acceptance. Consistent with this perspective, Simonsohn (2014) recently conducted a simulation study demonstrating that Bayesian approaches "are as invalidated by selective reporting as *p* values are" (p. 1). We do not agree that changing the nature of the reported statistics will guarantee a reduction in these biases and therefore cannot endorse a movement away from NHST as a solution for opportunistic biases, although we would encourage researchers to consider alternative analytic paradigms to overcome some of the important limitations of NHST.

Statistically Control Type I Error Rates

Researchers have long recognized that the probability of committing a Type I error increases as a function of the number of tests conducted in an article (e.g., Betz & Gabriel, 1978), and it is relatively common that reviewers would request that authors use statistics that preserve the studywide error rate (e.g., Abdi, 2007) if a submitted article reported a large number of tests. Given the existing success of these methods in preventing increases in Type I error rates within a study, it is reasonable to consider such statistical corrections as a solution to opportunistic biases more generally.

While we certainly advocate that researchers employ appropriate statistical methods to control the studywide error rate when performing multiple tests, we do not believe that these methods will be of much use in preventing opportunistic biases. The issue is that most of the exploratory analyses that are performed are never reported in the final research article, so these corrections only rarely take them into account. There is an additional difficulty in that the appropriate application of these adjustments requires knowledge about the correlations among the tests, which is often not available. Although we might ask researchers to correct for tests that are not included in their articles, there is no way to easily determine exactly what prior tests are relevant and should be included in these corrections. The only possibility is to leave the decision up to the judgment of the author, which still allows researchers to bias their results by making motivated decisions about how many of their prior analyses are relevant enough to be included in the formulas for correction. In addition, there are a number of opportunistic biases, such as halting data collection once a significant statistic is obtained, that cannot be addressed using a statistical correction for multiple tests. We therefore would not endorse statistically controlling for Type I errors as a solution for opportunistic biases, although we would endorse the use of these corrections to prevent the inflation of the overall error rate within a particular study.

Characterizing and Categorizing the Solutions

Common Features of the Solutions

The proposed accommodations and adjustments for opportunistic biases vary greatly in their purposes and methods. Although we believe that it is important to use a multifaceted approach to combat opportunistic biases, the value that a proposal might have to an individual or organization would depend on the specific needs of that entity. To facilitate evaluations of and comparisons among the different proposals, we present Table 1, which codes each proposal on six important dimensions that are described next.

Minimum level of implementation. This dimension represents whether the solution is one that could be applied by individuals in their own work ("Individual"), whether it would be incorporated as part of journal submission or reviewing guidelines ("Journal") or whether it would need to be endorsed by a larger professional organization in order to be successful ("Organization"). Note that the code given in the table is the minimum level of implementation. It is possible, for example, that a solution coded as Individual be part of the submission requirements of a journal or that a solution coded as Journal be endorsed by a professional organization.

Table 1
Characteristics of the Different Solutions for Opportunistic Biases

Solution	Minimum level of implementation	Scope	Requires retraining	Extra labor if finding is valid	Precise or vague	Implementation requires systemic change
Replication	Individual	Broad	No	Additional study	Precise	No
Distinguish confirmatory analyses from exploratory analyses	Individual	Specific	No	None	Precise	Yes
Grassroots efforts by authors to increase the transparency of their research	Individual	Broad	Yes	Posting of materials	Vague	Yes
Develop and enforce better reporting standards	Organization	Broad	Yes	None	Precise	Yes
Have journals require increased disclosure	Journal	Broad	Yes	None	Precise	Yes
Reduce the bias against the null hypothesis	Journal	Specific	No	None	Precise	Yes
Submit research designs for publication before results are known	Journal	Broad	No	Additional submission	Precise	Yes
Increase the sample size and power of studies	Individual	Broad	No	Additional subjects	Precise	Yes
Use sequential analyses instead of informal stopping rules	Individual	Specific	Yes	Additional analysis	Precise	No
Use statistical analysis to detect biased results	Individual	Broad	Yes	None	Precise	No
Educate researchers about opportunistic biases	Organization	Broad	Yes	None	Vague	Yes
Have opinion leaders exert authority	Organization	Broad	No	None	Vague	No

Scope. This dimension represents whether the solution is designed to counter the effects of one or more specific sources of opportunistic bias ("Specific") or whether it would be effective at countering the influence of opportunistic bias no matter what source it came from ("Broad").

Requires retraining. This dimension represents whether researchers would ("Yes") or would not ("No") need to learn new statistical or methodological techniques in order to implement the solution.

Extra labor if finding is valid. This dimension represents whether the solution would or would not require authors of studies containing true effects to perform extra labor in order to get the studies published, and what extra labor would be involved. All of these solutions should be inhibiting studies that do not contain true effects, and so we do not consider extra labor in those cases to truly be a cost.

Precise or vague. This dimension represents whether the solution already has a specific implementation ("Precise") or whether the solution represents a general recommendation where the details would need to be determined by those attempting to employ the solution ("Vague").

Implementation requires systemic change. This dimension represents whether the implementation of the solution would ("Yes") or would not ("No") require that the field as a whole change the way it conducts, evaluates, publishes, or otherwise handles research findings.

In these terms, a solution will be easier to implement if it can be implemented at the individual level, it does not require retraining, it does not require extra labor for valid studies, and does not require systemic change. A solution will have a greater impact on the effects of opportunistic biases if it has a broad scope than if it has a specific scope. There are a number of other idiosyncratic factors that must be considered when evaluating these solutions (many of which we discussed in the descriptions of the individual solutions above), so we would not suggest that these guidelines be used to inhibit the development of specific solutions. We echo one of Pashler and Wagenmakers' (2012) conclusions about the problem of replicability, that ". . . it would be a mistake to try to rely upon any single solution to such a complex problem" (p. 529). However, we do believe that considering these dimensions can be useful for researchers trying to decide what solutions they have the need and the ability to implement. We also believe that they should be considered by statisticians and methodologists attempting to design solutions for opportunistic biases so that their solutions can influence the greatest number of researchers and help protect against the greatest number of biases.

An Integrated Solution to Reduce Opportunistic Biases

As we have discussed, opportunistic biases arise from a number of different sources and have multiple effects on the psychological literature. We therefore propose a multifaceted approach to remediate their effects, illustrated in Figure 1, integrating the different solutions mentioned above within a broader model suggesting how researchers can conduct and present research that is less influenced by opportunistic biases. The model focuses on those solutions we believe would be successful, and is organized around what investigators can do at each stage of the research process to produce research that is minimally biased. The model is multifaceted, suggesting that there are many approaches that can and should be used to reduce opportunistic biases.

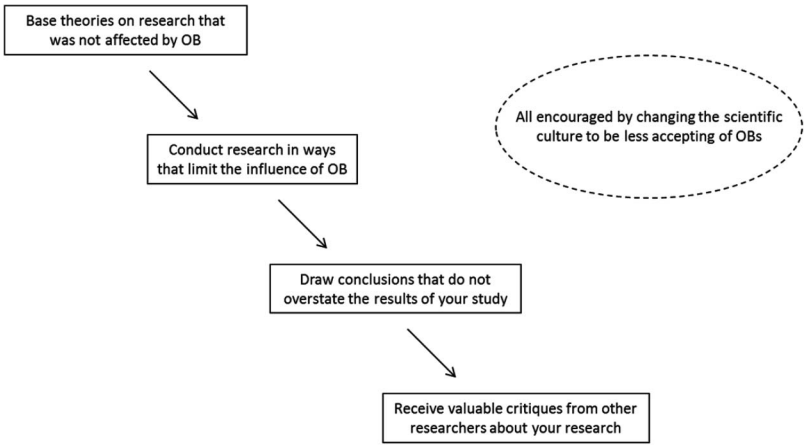
At the initial stage when investigators are developing their research ideas, they should do their best to base their studies on past research that has not been influenced by opportunistic biases. Theories based on prior studies that are free of bias are more likely to be correct, and methods taken from unbiased studies are more likely to affect outcomes through the theoretically proposed mechanisms. This combination produces more successful studies whose findings are easier to explain. Unfortunately, the information that would be required to identify studies that may have used biasing procedures is typically unavailable for older studies. However, we can use the statistical analyses that researchers have developed to detect biased results to determine the degree to which these were likely employed in a literature. The findings from topic areas where the collection of results do not conform to what we'd expect from an unbiased distribution would need to be treated as suspect. When researchers have questions about whether a particular finding is valid, they can conduct a direct replication of the primary result.

Once investigators have determined their research questions, they can reduce the possibility of opportunistic biases by using appropriate analytic methods. Studies using methods that prevent the influence of opportunistic biases should only produce a minimal number of false results (determined by the significance level when using NHST), so researchers can be more confident that their findings truly reflect relations among the constructs they are investigating. In general, researchers should let their research questions determine their analyses, rather than perform a wide sweep of analyses whose significant findings are used to identify viable research questions. Methodologists have also provided several specific suggestions to assist researchers during this phase, such as publicizing analytic plans before the results are known, ensuring that studies are adequately powered, and using sequential analyses instead of informal stopping rules.

After conducting a study, the influence of opportunistic biases can be reduced by ensuring that the discussion does not overstate the implications of the results and that it identifies any potential sources of bias that may remain. No study is perfect, and truthfully admitting any reservations helps others to accurately evaluate its findings. It can also help preserve a researcher's reputation should future investigations reveal that the initial interpretation was incorrect. Of great importance at this stage is to appropriately distinguish confirmatory from exploratory analyses, and to avoid slanting the results in favor of significant results or in favor of a particular theoretical perspective.

Finally, it is important to remember that even though the revision of a specific article will stop after it has been published, the scientific process itself is iterative and is strengthened by feedback between authors and readers. It is therefore important that investigators write their articles in a way that will maximize the ability of readers to provide accurate and useful criticisms about the study. This can be

Figure 1
An Integrated Solution for Conducting Research Free From Opportunistic Biases



accomplished by making study details publicly available so the potential influence of opportunistic biases is made clear, such as by improving the overall transparency of the research process and by enforcing better reporting standards for journal articles. When readers have the most accurate information about a study and the authors are open to providing details and sharing materials, the suggestions that readers make will be better informed, more accurate, and will provide the most insight into the phenomenon being studied.

To encourage researchers to adopt research practices that will reduce the influence of opportunistic biases, efforts must be made to change the scientific culture within psychology so that procedures leading to opportunistic biases are viewed as unacceptable. Although there are a number of researchers discussing the importance of these issues, the origins, prevalence, and effects of opportunistic biases are not regularly taught to students in psychology graduate programs, nor are the solutions part of everyday practices for most psychologists. Changing the way that people evaluate research practices can be done in two major ways. First, those who want to promote using methods that reduce the effects of opportunistic biases can make efforts to educate their students and colleagues about their benefits. This may take the form of educational workshops at conferences, lectures for methods courses, or online tutorials. Second, those who are in positions of authority, such as journal editors or officers in scientific organizations, can make broader calls for change and can mandate the use of new methods for research submitted to academic journals or conferences. Other researchers might also lobby for change by writing to those in positions of authority to convince them both of the importance of adopting new methods and standards and of the presence of popular support for such changes.

Conclusions

A number of different issues related to opportunistic biases have become increasingly important to psychology. The last 5 years have seen increases in papers concerned with questionable research methods, the replicability of findings, and the benefits of open science. There has also been an unfortunate increase in papers being retracted for fraud (Jha, 2012). The prevalence and effects of biased research is more germane now than at almost any other time in the history of the field. We therefore believe that it is important that all psychologists become aware of the issues surrounding opportunistic biases and the proposed solutions to them that are currently under debate. The goal of our paper was to organize and synthesize the discussions related to opportunistic biases to help guide psychologists through the upcoming changes that the field will make regarding the accepted conventions for conducting, presenting, and evaluating research.

The final solution to opportunistic biases is to change the culture of psychological science so that the investigators who receive the greatest rewards are those who use the most rigorous methods. Changing a culture is incredibly difficult because it is based upon and simultaneously reinforces a large collection of individual behaviors (Knott, Muers, & Aldridge, 2008). There will necessarily be resistance from factions that benefit from the current culture, and those attempting to make changes on their own will likely face professional costs for failing to take advantage of biases in their research. They may even suffer social costs for opposing the established norms. However, we believe that there must eventually be a culture shift so that the field opposes research affected by opportunistic biases, partly because of the current momentum so evident for this type of change, partly because of the difficulty of defending the current norms allowing biased research against statistical arguments opposing those norms, and partly because of the clear benefits for science that will result from the change. As more researchers oppose biased research, become invested in performing open science, and come to value having accurate, replicable findings, the behaviors supporting the current norms will be less prevalent, allowing new standards opposing the use of procedures that create opportunistic biases to be established.

As computing power increases and it becomes simpler for researchers to perform exploratory analyses, the potential influence of opportunistic biases grows ever larger. The increasing role of data mining and other automated data-analytic techniques will reshape the social sciences, possibly in regrettable ways, unless researchers decide to provide direction for these changes. With scientists such as Ioannidis (2005) concluding that most published findings are actually false, it is unsurprising that many nonscientists dismiss the results of scientific research, believing that investigators commonly bias their findings so that they only represent the views that they endorse. The tradition of discussing the results from exploratory analyses as if they were confirmatory tests must be curtailed, as this misrepresents the actual confidence that we may have in the results. As the recent publicity of biased results in multiple fields (Begley & Ellis, 2012) suggests, if we are not able to reduce opportunistic biases in the practice of science, we risk losing continued scientific progress along with all of its practical benefits. We believe that this negative perception of scientific research can be changed if researchers pay more attention to the prevention of opportunistic biases. Scientists must be seen to reaffirm their dedication to uncovering true insights, instead of being seen as biasing their studies to support their theories or self-interests. If we do not make a change, the cynical perspective that scientific research only reflects the political, economic, or social beliefs of its practitioners will undermine any recommendations that we might make based on our work. Cultural change may take some time, but each individual can play a vital role in bringing about that change in the way they conduct their own research, how they evaluate articles as reviewers, how they advocate for change among their peers, and how they act in their positions of authority to insist on the importance of avoiding opportunistic biases. In the end, instead of being seen like Lang's "drunken man" grasping for support, researchers instead can be seen

as using the light of scientific discovery to reveal the truths about human nature.

REFERENCES

Abdi, H. (2007). Bonferroni and Šidák corrections for multiple comparisons. In N. J. Salkind (Ed.), *Encyclopedia of measurement and statistics* (pp. 103–107). Thousand Oaks, CA: Sage.

American Psychological Association (APA). (2010). *Publication manual of the American Psychological Association, sixth edition*. Washington, DC: Author.

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist, 63,* 839–851. http://dx.doi.org/10.1037/0003-066X.63.9.839

Association for Psychological Science. (2014). *Registered replication reports*. Retrieved March 2014 from http://www.psychologicalscience.org/index.php/replication

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (3rd ed.). Chichester, England: Wiley.

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature, 483,* 531–533. http://dx.doi.org/10.1038/483531a

Bem, D. J. (2003). Writing the empirical journal article. In J. M. Darley, M. P. Zanna, & H. L. Roediger III, (Eds.), *The compleat academic* (pp. 171–201). Washington, DC: American Psychological Association.

Betz, M. A., & Gabriel, K. R. (1978). Type IV errors and analysis of simple effects. *Journal of Educational Statistics, 3,* 121–143. http://dx.doi.org/10.2307/1164881

Botella, J., Ximénez, C., Revuelta, J., & Suero, M. (2006). Optimization of sample size in controlled experiments: The CLAST rule. *Behavior Research Methods, 38,* 65–76. http://dx.doi.org/10.3758/BF03192751

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14,* 365–376. http://dx.doi.org/10.1038/nrn3475

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology, 65,* 145–153. http://dx.doi.org/10.1037/h0045186

Cohen, J. (1994). The Earth is round ($p <$. 05). *American Psychologist, 49,* 997–1003. http://dx.doi.org/10.1037/0003-066X.49.12.997

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Congdon, P. (2006). *Bayesian statistical modeling*. Chichester, UK: Wiley. http://dx.doi.org/10.1002/9780470035948

Cooper, H. (2010). *Reporting research in psychology: How to meet journal article reporting standards*. Washington, DC: American Psychological Association.

Cooper, H., & Vandenbos, G. R. (2013). Archives of scientific psychology: A new journal for a new era. *Archives of Scientific Psychology, 1,* 1–6. http://dx.doi.org/10.1037/arc0000001

Costa, P. T., & McCrae, R. R. (1992). *The NEO-PIR professional manual*. Odessa, FL: Psychological Assessment Resources.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25,* 7–29. http://dx.doi.org/10.1177/0956797613504966

Devine, P. G. (2012, August 19). Responsible conduct in research. E-mail to SPSP members.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science, 6,* 274–290. http://dx.doi.org/10.1177/1745691611406920

Eich, E. (2014). Business not as usual. *Psychological Science, 25,* 3–6. http://dx.doi.org/10.1177/0956797613512465

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

Fidler, F. (2002). The fifth ed. of the *APA Publication Manual*: Why its statistics recommendations are so controversial. *Educational and Psychological Measurement, 62,* 749–770.

Frick, R. W. (1998). A better stopping rule for conventional statistical tests. *Behavior Research Methods, Instruments, & Computers, 30,* 690–697. http://dx.doi.org/10.3758/BF03209488

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland, 15,* 246–263. http://dx.doi.org/10.2307/2841583

Goodman, S. N. (1993). p Values, hypothesis tests, and likelihood: Implications for epidemiology of a neglected historical debate. *American Journal of Epidemiology, 137,* 485–496.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82,* 1–20. http://dx.doi.org/10.1037/h0076157

Hays, W. L. (1963). *Statistics for psychologists*. New York, NY: Holt, Rinehart & Winston.

Hubbard, R., & Armstrong, J. S. (1997). Publication bias against null results. *Psychological Reports, 80,* 337–338. http://dx.doi.org/10.2466/pr0.1997.80.1.337

Hubbard, R., & Ryan, P. A. (2000). The historical growth of statistical significance testing in psychology—and its future prospects. *Educational and Psychological Measurement, 60,* 661–681.

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

Ioannidis, J. P. A., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials, 4,* 245–253. http://dx.doi.org/10.1177/1740774507079441

Jha, A. (2012). Tenfold increase in scientific research papers retracted for fraud. *The Guardian*. Retrieved October 2012 from http://www.guardian.co.uk/science/2012/oct/01/tenfold-increase-science-paper-retracted-fraud

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. http://dx.doi.org/10.1177/0956797611430953

Johnson, V. E. (2013). Revised standards for statistical evidence. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 110,* 19313–19317. http://dx.doi.org/10.1073/pnas.1313476110

Journal for Experimental Social Psychology. (2014). *JESP editorial guidelines*. Retrieved May 2014 from http://www.journals.elsevier.com/journal-of-experimental-social-psychology/news/jesp-editorial-guidelines

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2,* 196–217. http://dx.doi.org/10.1207/s15327957pspr0203_4

Knott, D., Muers, S., & Aldridge, S. (2008). *Achieving culture change: A policy framework*. London, UK: Prime Minister's Strategy Unit. Retrieved October 2012 from http://webarchive.nationalarchives.gov.uk/20100125070726/http://cabinetoffice.gov.uk/media/cabinetoffice/strategy/assets/achieving_culture_change.pdf

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science, 6,* 299–312. http://dx.doi.org/10.1177/1745691611406925

Kupfersmid, J. (1988). Improving what is published: A model in search of an ed. *American Psychologist, 43,* 635–642. http://dx.doi.org/10.1037/0003-066X.43.8.635

Laws, K. R. (2013). Negativland: A home for all findings in psychology. *BMC Psychology, 1,* 2. http://dx.doi.org/10.1186/2050-7283-1-2

Leamer, E. E. (1978). *Specification searches: Ad hoc inference with non-experimental data*. New York, NY: Wiley.

LeBel, E. P., Borsboom, D., Giner-Sorolla, R., Hasselman, F., Peters, K. R., Ratliff, K. A., & Smith, C. T. (2013). PsychDisclosure.org: Grassroots support for reforming reporting standards in psychology. *Perspectives on Psychological Science, 8,* 424–432. http://dx.doi.org/10.1177/1745691613491437

Light, R., & Pillemer, D. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.

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. http://dx.doi.org/10.1037/0033-2909.111.3.490

Many Labs Replication Group. (2014). *Investigating variation in replicability: The "Many Labs" Replication Project*. Retrieved March 2014 from https://osf.io/wx7ck/

Morey, L. C. (1991). *Personality Assessment Inventory professional manual*. Odessa, FL: Psychological Assessment Resources.

Natrella, M. G. (1960). The relation between confidence intervals and tests of significance. *The American Statistician, 14,* 20–22, 38.

Neher, A. (1967). Probability pyramiding, research error and the need for independent replication. *The Psychological Record, 17,* 257–262.

Pashler, H., & Wagenmakers, E. J. (Eds.) (2012). Introduction to the special section on replicability in psychological science: A crisis of confidence. *Perspectives on Psychological Science, 7,* 528–530. http://dx.doi.org/10.1177/1745691612465253

Pierce, J., Nosek, B. A., Miguez, S., Holcombe, A. O., Spies, J. R., de-Wit, L., & Lewis, M. (2013). *Standards (and badges) for encouraging open science behaviors*. Retrieved July 2013 from http://opensciencefamework.org/project/TVyXZ/

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika, 64,* 191–199. http://dx.doi.org/10.1093/biomet/64.2.191

Prelec, D. (2004). A Bayesian truth serum for subjective data. *Science, 306,* 462–466. http://dx.doi.org/10.1126/science.1102081

Reproducibility Project & the Psychology Group. (2014). *Reproducibility Project: Psychology*. Retrieved March 2014 from http://openscienceframework.org/project/EZcUj/wiki/home

Robinson, D. H., & Wainer, H. (2002). On the past and future of null hypothesis significance testing. *The Journal of Wildlife Management, 66,* 263–271. http://dx.doi.org/10.2307/3803158

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86,* 638–641. http://dx.doi.org/10.1037/0033-2909.86.3.638

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods, 17,* 551–566. http://dx.doi.org/10.1037/a0029487

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13,* 90–100. http://dx.doi.org/10.1037/a0015108

Selvin, H. C., & Stewart, A. (1966). Data-dredging procedures in survey analysis. *The American Statistician, 20,* 20–23.

Shaver, J. P., & Norton, R. S. (1980). Randomness and replication in ten years of the *American Educational Research Journal*. *Educational Researcher, 9,* 9–15.

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. http://dx.doi.org/10.1177/0956797611417632

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). *A 21 word solution*. http://dx.doi.org/10.2139/ssrn.2160588

Simonsohn, U. (2012). It does not follow: Evaluating the one-off publication bias critiques by Francis (2012a, 2012b, 2012c, 2012d, 2012e, in press). *Perspectives on Psychological Science, 7,* 597–599. http://dx.doi.org/10.1177/1745691612463399

Simonsohn, U. (2014). *Posterior-hacking: Selective reporting invalidates Bayesian results also*. http://dx.doi.org/10.2139/ssrn.2374040

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General, 143,* 534–547. http://dx.doi.org/10.1037/a0033242

Spies, J., & Nosek, B. (2014). *Open science framework*. Retrieved March 2014 from http://openscienceframework.org/

Valentine, J. C., Biglan, A., Boruch, R. F., Castro, F. G., Collins, L. M., Flay, B. R., . . . Schinke, S. P. (2011). Replication in prevention science. *Prevention Science, 12,* 103–117. http://dx.doi.org/10.1007/s11121-011-0217-6

van Assen, M. A., van Aert, R. C., Nuijten, M. B., & Wicherts, J. M. (2014). Why publishing everything is more effective than selective publishing of statistically significant results. *PLoS ONE, 9,* e84896. http://dx.doi.org/10.1371/journal.pone.0084896

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7,* 632–638. http://dx.doi.org/10.1177/1745691612463078

Wald, A. (1947). *Sequential analysis*. New York, NY: Wiley.

Walster, G. W., & Cleary, T. A. (1970). A proposal for a new editorial policy in the social sciences. *The American Statistician, 24,* 5–10.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E. J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 *t* tests. *Perspectives on Psychological Science, 6,* 291–298. http://dx.doi.org/10.1177/1745691611406923