Journal of Experimental Social Psychology 66 (2016) 116–133

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

Interpretations and methods: Towards a more effectively self-correcting social psychology☆

Lee Jussim [a,*], Jarret T. Crawford [b], Stephanie M. Anglin [a], Sean T. Stevens [a], Jose L. Duarte [c]

[a] Rutgers University, United States
[b] The College of New Jersey, United States
[c] Arizona State University, United States

ARTICLE INFO

Article history:
Received 16 September 2014
Revised 5 October 2015
Accepted 11 October 2015
Available online 24 March 2016

Keywords:
Research methods
Statistics
Scientific integrity
Best practices
Interpretation

ABSTRACT

We consider how valid conclusions often lay hidden within research reports, masked by plausible but unjustified conclusions reached in those reports. We employ several well-known and cross-cutting examples from the psychological literature to illustrate how, independent (or in the absence) of replicability difficulties or questionable research practices leading to false positives, motivated reasoning and confirmation biases can lead to drawing unjustified conclusions. In describing these examples, we review strategies and methods by which researchers can identify such practices in their own and others' research reports. These strategies and methods can unmask hidden phenomena that may conflict with researchers' preferred narratives, in order to ultimately produce more sound and valid scientific conclusions. We conclude with general recommendations for how social psychologists can limit the influence of interpretive biases in their own and others' research, and thereby elevate the scientific status and validity of social psychology.

"Getting it right" is the sine qua non of science (Funder et al., 2014). Science can tolerate individual mistakes and flawed theories, but only if it has reliable mechanisms for efficient self-correction. Unfortunately, science is not always self-correcting (Ioannidis, 2012). Indeed, a series of threats to the integrity of scientific research has recently come to the fore across the sciences, including questionable research practices, failures to replicate, publication biases, and political biases (Begley & Ellis, 2012; Duarte et al., 2015; Ioannidis, 2005; Simmons, Nelson, & Simonsohn, 2011). In response to these issues, individuals and organizations have begun addressing how to improve scientific practices through reforms targeting transparency, statistics, and data collection methods.

The term "methods" typically refers to ways of collecting data (construction of measures and research design); the term sometimes also includes statistics. More generally, however, "method" refers to how scientists go about conducting science. Our view is that every step of "how one goes about reaching scientific conclusions" is "method." In this paper, we consider how valid conclusions often lay hidden within research reports, masked by plausible but unjustified conclusions reached in those reports. These conclusions do not necessarily involve the use of questionable research practices. Invalid conclusions may be reached based, not on failing to report dropped conditions, failed studies, or nonsignificant analyses, but on selective *interpretations* of data that highlight researchers' preferred conclusions while masking more valid ones. In this paper, we consider ways to identify, unmask, and correct invalid conclusions that mask valid ones.

1. Masked interpretations, phenomena, and alternative explanations

We characterize situations in which the data justify a different conclusion than reached in a published report as situations in which that different conclusion is "masked." Masked phenomena may constitute *alternative explanations* for a pattern of results, reasons to believe the published interpretations are true but exaggerated, or reasons to believe the published interpretation is simply incorrect. These conclusions are typically masked because the original report does not even consider or acknowledge them, and because the data that are presented usually create the superficial appearance of support for the presented conclusions. We next discuss two simple and well-known examples of masked phenomena to illustrate how we use the concept.

☆ Work on this paper has been supported by grants from the Hewlett Foundation (2014-1735), the Fetzer Foundation, and the Center for Advanced Study in the Behavioral Sciences (CASBS), Stanford. The initial draft of this paper was completed while Lee Jussim was a CASBS Scholar. We thank Ifat Maoz for providing the notion of "Wow Effect" developed in this paper.

* Corresponding author.
E-mail addresses: jussim@rutgers.edu (L. Jussim), crawford@tcnj.edu (J.T. Crawford), stephaniemanglin@gmail.com (S.M. Anglin), seantstevens@gmail.com (S.T. Stevens), jlduarte@asu.edu (J.L. Duarte).

http://dx.doi.org/10.1016/j.jesp.2015.10.003
0022-1031/

1.1. Simpson's paradox

Simpson's paradox refers to the fact that a valid statistical conclusion for an entire sample may be invalid for all subsamples (Simpson, 1951). As such, it is the classic example of a masked phenomenon. In the 1970s, UC Berkeley was sued for gender bias in graduate admissions because about 44% of men, but only 35% of women were admitted (see Bickel, Hammel, & O'Connell, 1975 for the evidence). This difference is close to that identified by Greenwald, Banaji, and Nosek (2015) as meeting legal standards for the possibility of discrimination, and similar disparities have been interpreted as suggesting discrimination (e.g., Ledgerwood, Haines, & Ratliff, 2015; Shen, 2013).

In the particular case of Berkeley, however, it turned out that women were as or more likely to be admitted to the departments to which they applied as were men (Bickel et al., 1975). How is this even possible? It is possible because *women disproportionately applied to the departments with lower admissions rates, not because, within departments, women were less likely to be admitted*. Berkeley had 85 departments; details regarding the six largest departments are available on Wikipedia under "Simpson's paradox". Interested readers can also consult Bickel et al. (1975) for more details.

Table 1 presents a hypothetical example. If one examined only the overall admission rate, one would find what appears to be massive evidence of gender bias. Only 290/1000 women are admitted, whereas 710/1000 men are admitted. However, women are admitted at higher levels in both the competitive (22% vs. 10%) and easy (90% vs. 78%) departments. There is evidence here that women apply disproportionately to the more difficult department, but there is no evidence that either department discriminates against women. Thus, that women were being disproportionately accepted into each program was masked behind the aggregate data. Of course, explaining why women disproportionately applied to the more difficult program was beyond the scope of these analyses, leaving open the possibility that there was bias against women *somewhere else* in the social processes culminating in graduate applications. The data do not address the existence of bias against women *writ large*; they only refute the claim that departmental admissions committees discriminated against women by selecting proportionately more men than women.

1.2. Experimenter (lack of) blindness to conditions

Phenomena may often be masked because researchers failed to include procedures that could reveal them. The simplest example is experimenter blindness to conditions. Many reports of experimental studies that involve experimenters interacting with live participants (as opposed to, e.g., studies conducted completely online) do not explicitly declare that experimenters were blind to condition. Indeed, we randomly selected 20 papers reporting at least one experiment published in *Journal of Personality and Social Psychology* in 2007, and coded: 1. Whether they involved live interactions between experimenters and participants; and 2. Whether the methods section described experimenters as blind to condition. Of the 66 experiments reported in these 20 papers, 63 of them involved a participant–experimenter interaction. Of these, only 15 explicitly declared that experimenters were blind to condition. This raises the possibility that experimenter effects (Rosenthal & Fode, 1963), rather than the authors' stated hypothesis, explains all or some of the results of these studies.

It is possible that experimenters were blind in some of these studies, even though the published reports failed to state so. Regardless, if no statement of blindness appears in the published report, we cannot assume that experimenters were blind. If experimenters were not blind an experimenter effect account may explain all or some of the obtained findings. These studies rarely, if ever, even acknowledged this potential problem — thus experimenter effects remain masked, an alternative explanation hiding in plain sight "underneath" the text of the publish reports. This analysis is not purely hypothetical. In a rare case of researchers correcting their own research, Lane et al. (2015) reported failures to replicate their earlier findings (Mikolajczak et al., 2010, same team). They noted that experimenters had not previously been blind to condition, which may have caused a phantom effect.

Simpson's paradox is a good example of a masked phenomenon, not because we have any reason to believe that social psychology is riddled with data misinterpreted due to researchers missing evidence of Simpson's paradox, but because it is a clear example of a more general potential problem: researchers' data may be clean (obtained without any questionable practices) and analyses performed statistically appropriately, and their conclusion may still be wrong. The problem of experimenter blindness to condition is a good example for a different reason. Researchers have known about this problem since the early 1960s. Nonetheless, our results raise the general point that just because some methodological procedures for minimizing masked phenomena may be well-known does not mean they are in widespread use. If they are in widespread use but just not being reported, then explicitly articulating this aspect of method should be encouraged, or even required, by journal editors and reviewers, so that consumers of those reports will know that experimenter effects do not explain the obtained results. Lacking such an explicit statement, we are left with the possibility that something very different than what the authors have claimed explains the results.

The rest of this paper focuses on three issues: 1. Identifying social psychological theoretical bases for predicting that researchers would not always adopt the procedures needed to unmask hidden phenomena; 2. Reviewing substantive examples from highly influential work in social psychology in which alternative phenomena went unmasked for years; and 3. Identifying practices researchers can adopt to reduce their vulnerability to allowing their analyses and interpretations to leave better interpretations and explanations masked.

2. Sources of the failure to expose masked phenomena

Exposing masked phenomena requires four ingredients, all of which are necessary, and none of which are sufficient:

1. Awareness of the possibility of masked phenomena.
2. The motivation to expose them.
3. The expertise necessary to expose them.
4. The data necessary to test for them.

A failure in any one can lead to a failure to expose a masked phenomenon. In the Berkeley case, failure to expose the masked bias in favor of women could plausibly have resulted from three of these four sources. Perhaps the plaintiffs were unaware of Simpson's paradox. Or, perhaps

Table 1
Simpson's paradox, hypothetical example.

	Men accepted	Men rejected	Women accepted	Women rejected
Competitive admissions department	10	90	200	700
Easy admissions department	700	200	90	10

Overall, proportionately fewer women than men are admitted (290/1000 versus 710/1000), but a higher proportion of women are admitted to both the easy department (90% vs. 78%) and the competitive department (22% vs. 10%). Higher admission rates for women, within each department, are revealed here, though they are hidden by the overall higher admission rate for men (71% vs. 29%).

the aggregate bias in favor of men was so "obviously" sexism to the plaintiffs that their motivation to be sure was short-circuited by overconfidence. Or, perhaps they simply did not have the data broken down by department readily available.

These four ingredients are not necessarily independent, and have not been presented in any chronological ordering. One might presume that basic scientific training would lead social psychologists to possess all four ingredients. Perhaps that is true, but we are aware of no evidence that bears on this question. More important, however, are the existence of countervailing social and psychological forces that can lead to practices that undermine researchers' likelihood of uncovering masked phenomena, which are discussed next.

2.1. Confirmation bias and motivated reasoning

2.1.1. Confirmation bias and motivated reasoning among laypeople

Motivated reasoning refers to biased information processing that is driven by goals unrelated to accurate belief formation (Kahan, 2011; Kunda, 1990). A specific type of motivated reasoning, confirmation bias, occurs when people seek out and evaluate information in ways that confirm their pre-existing views while downplaying, ignoring, or discrediting information of equal or greater quality that opposes their views (Nickerson, 1998; also referred to as myside bias, see Stanovich, West, & Toplak, 2013). People intensely scrutinize counter-attitudinal evidence while easily accepting information supporting their views (e.g., Ditto & Lopez, 1992; Lord, Ross, & Lepper, 1979). Although these processes are affectively driven (e.g., Jacks & Devine, 2000; Munro & Ditto, 1997; Zuwerink & Devine, 1996), people generate convincing arguments to justify their automatic evaluations, producing an illusion of objectivity (Haidt, 2001; Nickerson, 1998).

2.1.2. Confirmation bias and motivated reasoning among scientists

Scientists are not immune to confirmation biases and motivated reasoning (Ioannidis, 2012; Lilienfeld, 2010). Values influence each phase of the research process, including how people interpret research findings (Duarte et al., 2015). Reviewers' theoretical (Epstein, 2004; Greenwald, Pratkanis, Leippe, & Baumgardner, 1986; Mahoney, 1977) and ideological (Abramowitz, Gomes, & Abramowitz, 1975) views can influence their evaluation of research reports, leading them to judge studies that oppose their beliefs more critically than studies supporting their views. Consequently, they are then less likely to recommend publication of studies with undesired findings or funding for studies based on undesirable theories or hypotheses. Confirmation bias is sometimes defensible from a decision theory or Bayesian perspective (see MacCoun, 1998, for a review). Of course, just because confirmation bias *can* be justified does not mean most of its occurrences are, in fact, defensible (e.g., one of MacCoun's, 1998, conditions under which such biases can be defensible is when researchers specifically articulate reasons for holding a certain conclusion to a higher standard than some alternative conclusion — an articulation that, in our experience, is rarely found in the literature, and which appears in *none* of the examples presented in the remainder of this review).

2.2. The power of the story and the academic incentive structure

There are powerful incentives for psychologists to present a strong, compelling story when describing their research, and such practices have been encouraged (Bem, 2002; Jordan & Zanna, 2007). Most of us are motivated to get the science right, but we are also motivated to get the studies published and our grants funded. We want our colleagues to find our research sufficiently interesting and important to support publishing it, and then to cite it, preferably a lot. We want jobs, promotions, and tenure. We want popular media to publicize our research and to disseminate our findings beyond the confines of our lab. We might even hope to tell a story so compelling we can produce a bestselling popular book and receive lucrative consulting and speaking engagements, or have our findings influence policy decisions.

In brief, powerful incentives exist that motivate us to achieve — or, at least, appear to achieve — a "Wow Effect" (Jussim & Maoz, 2014). A "Wow Effect" is some novel result that comes to be seen as having far-reaching theoretical, methodological, or practical implications. It is the type of work likely to be emulated, massively cited, and highly funded.

But how can our stories be sufficiently compelling and persuasive to draw attention when the average effect in our field is $r = .20$ (an estimate which itself is probably inflated by the existence of publication biases and other distortions — e.g., Bakker, van Dijk, & Wicherts, 2012; Simmons, Nelson, & Simonsohn, 2012)? How can we create beautiful, coherent stories from data that is almost always messy, only partially supports our claims, is difficult to replicate (even when we are right — Krosnick, 2015), even more difficult to scale up into real world interventions and policies (Sampson, Winship, & Knight, 2013), and typically subject to many different alternative explanations?

Compelling, persuasive narratives are amply rewarded by promotions, grants, named chairs, etc., but the relationship of "compellingness of narrative" to validity (effect size, replicability, generalizability, etc.) is currently unknown. This raises the possibility that for some unknown and possibly substantial portion of the time, we are rewarding research practices that produce Wow Effects that are false, distorted, or exaggerated. We next demonstrate, with examples drawn from actual scholarship, how mundane explanations for the same data remain hidden in the depths of the theorizing, methodology, statistics, and conclusions of some major areas of psychological science.

3. The New Look in Perception, confirmation bias, blind spots, and masked veridicality

The New Look in Perception of the 1940s is a classic case of confirmation bias and masked phenomena. The dominant behaviorist perspective of the period banished fears, needs, and expectations from scientific study, dismissing such internal states as unscientific. The New Look researchers then came and, en masse, set out to demonstrate ways in which such internal states could influence and distort perception (see Allport, 1955 for a review). The main claims of the New Look could be captured by two concepts: Perceptual vigilance and perceptual defense. Perceptual vigilance referred to the tendency for people to be hypersensitive to perceiving stimuli that met their needs or were consistent with their values, beliefs, or personalities. Perceptual defense referred to the tendency for people to avoid perceiving stimuli that was uncomfortable or threatening.

The New Lookers generated an impressive body of literature seeming to demonstrate influences of bodily needs, reward and punishment, personal values, personality, and motivations to avoid taboos on perception (see reviews by Allport, 1955; Bruner, 1957; Jussim, 2012a,b). But for most studies, nonperceptual alternatives were not ruled out. Hungry people were sometimes more likely to associate food with various (nonfood) pictures (e.g., Levine, Chein, & Murphy, 1942), but, of course "associating" food with a picture is not the same thing as perceiving food. People were slower to report recognizing taboo words than nontaboo words (McGinnies, 1949), but it was never clear whether this was *difficulty perceiving the words* (consistent with a perceptual defense explanation) or whether they were simply more reluctant to verbalize such words to an experimenter (F. H. Allport also identified several other veridical perception explanations for these results). As Allport (1955) amply demonstrated, underneath *every* New Look claim to have produced evidence of fears or desires influencing perception there was a mundane but viable hidden explanation that could not be ruled out.

The New Look can be viewed as researcher confirmation bias writ large. Researchers clearly *wanted* to find what was then the "Wow Effect" of motivations distorting perception, and so leaped to *interpretations* of such effects with insufficient skepticism, at least as manifested in their published scholarship. Very little of the New Look literature

acknowledged even the existence of the potential ways in which their empirical results could reflect veridical perception rather than motivational factors, a blind spot writ large. The *New Look* was a *confirmatory search* for evidence interpretable as perceptual defense or vigilance — leaving other, often more viable, interpretations masked by virtue of being neither tested nor mentioned in the scientific articles reporting new empirical studies.

One might view the New Look as an example of science self-correcting because, in recognition of these issues, by the 1960s, the claims had been largely dropped from social psychology (although social psychology's long emphasis on error and bias, and work on priming social behaviors and automaticity can be viewed as having roots in the New Look). That is undoubtedly true to some degree, and it is why we have only a brief discussion of it here. Nonetheless, *the scientific processes* that led to about 20 years of unjustified conclusions are a good example of how confirmation biases (seeking information interpretable as evidence of motivational influences on perception, but not seeking to disconfirm such influences) led to many unjustified or inadequately justified conclusions.

4. Examples of masked phenomena and how they could have been uncovered

4.1. How the remaining examples included in this review were selected

Exposing masked phenomena requires expertise specific to the topic and context being investigated. The present authors' expertise is primarily in the areas of political psychology, intergroup relations, and social cognition/social perception; as social psychologists, we are all familiar with many of the most famous and influential conclusions in social psychology (e.g., "the power of the situation"). As such, those are the domains in which most of the remainder of this review focuses.

For example, exposing masked phenomena in some of our examples required knowledge of the number of plays in a football game, the El Greco fallacy, or a sophisticated understanding of analysis of covariance. We suspect that vanishingly few readers of this article have the requisite knowledge on all three topics upon first read (although obtaining that knowledge is not particularly difficult). Lacking very specific expertise relevant to any particular research area, it will often be *impossible* to expose masked explanations.

Thus, examples here are illustrations of how suboptimal practices lead to unjustified *interpretations* of data. They reflect a mix of classic and current research primarily in social psychology, with some at the intersection of social and cognitive psychology. Our view is that true sciences self-correct, and in that spirit, we think that the common but unjustified conclusions we highlight here warrant correction, especially since those unjustified conclusions often still appear in the contemporary literature.

Our goal is not to perform a systematic assessment of the frequency or prevalence of these problems in social psychology. This is not a meta-analysis of problems of interpretations or the presence of masked phenomena, and there are no known methods for performing one. Consequently, we have no information about effect sizes, distributions, or moderators of such problems.

For each example of an unjustified conclusion presented below, we also present one or more examples of papers on the same topic reaching justified conclusions. We present *illustrative case studies* of how social psychology has gone wrong in order to avoid not just those specific errors going forward, but other similar errors.

4.2. They saw (nearly) the same game: Hastorf and Cantril (1954)

4.2.1. The study

This early study is a classic because it demonstrated subjectivity and bias in social perception, themes that were to become a mainstay of modern social psychology. In 1951 Dartmouth and Princeton played a hotly contested, aggressive football game. A Princeton player received a broken nose; a Dartmouth player broke his leg. Accusations flew in both directions: Dartmouth loyalists accused Princeton of playing a dirty game; Princeton loyalists accused Dartmouth of playing a dirty game. Hastorf and Cantril (1954) showed a film of the game to 48 Dartmouth students and 49 Princeton students, and had them rate the total number of infractions by each team. Dartmouth students saw both the Dartmouth and Princeton teams as committing slightly over four (on average) infractions. The Princeton students also saw the Princeton team as committing slightly over four infractions, but they saw the Dartmouth team as committing nearly ten infractions.

4.2.2. The conclusions

Because the Dartmouth and Princeton students diverged in the number of infractions they claimed were committed by Dartmouth, Hastorf and Cantril (1954) concluded that Princeton and Dartmouth students seemed to be actually seeing different games. Hastorf and Cantril's (1954, p. 133) own extraordinary interpretations of their study were as follows:

> "There is no such 'thing' as a 'game' existing 'out there' in its own right which people merely 'observe" and "The 'thing' simply is not the same for different people…."

Not surprisingly, the study has long been cited as a demonstration of how strongly motivations and beliefs color social perception (e.g., Ross, Lepper, & Ward, 2010; Schneider, Hastorf, & Ellsworth, 1979; Sedikedes & Skowronski, 1991). As Ross et al. (2010, p. 23) put it: "The early classic study by Hastorf and Cantril (1954) … reflected a radical view of the 'constructive' nature of perception that anticipated later discussions of naïve realism."

4.2.3. The masked phenomenon: they saw (mostly) the same game

Unfortunately, the study's results do not support Hastorf & Cantril's (1954) own extreme interpretations or any radical form of constructivism at all. This is quite easy to see from: 1. Their data; and 2. A minimal understanding of football. There was no difference in the infractions perceived by Dartmouth and Princeton students regarding the Princeton team. Thus, for half the game, there was no evidence that the students saw a different game; put positively, the evidence indicated that, for the Princeton half, *they saw the same game*.

What about the other half of the game? Perceptions of the Dartmouth team did show about a six perceived infraction difference between the Princeton and Dartmouth students. This is indeed bias, and it was statistically significant. However, it is also useful to consider *how much* of a bias this was. Most college football games have about 100 plays, or more. If one conservatively estimates that this particular game only had 60 plays (a low estimate biases conclusions in *favor* of bias), then a bias of six means that 54 judgments, or *90%, were unbiased*. This point is not presented in the paper (Hastorf & Cantril, 1954). Instead, it is masked by the data on perceived infractions, which are excerpted for their table out of a crucial context: the rest of the game.

Half the judgments (regarding the Princeton team) were completely unbiased; half the judgments (regarding the Dartmouth team) were at least 90% unbiased. Thus, *at least 95% of the time, judgments were unbiased* based on the measure of bias employed in the study itself, but when applied to the whole game rather than just perceived infractions (Hastorf & Cantril, 1954). For half a century, this study has been extolled as evidence for the power of subjectivity and bias. Overwhelmingly, however, the Princeton and Dartmouth students saw the same game.

4.2.4. Objections

One might object to the preceding analysis as missing the point — the study did indeed demonstrate group-serving biases in social perception (Hastorf & Cantril, 1954). Furthermore, small effects can be important. Winning and losing can sometimes hinge on a single important play, so it is valuable to know whether people view such plays differently. Lastly, who is to say what constitutes the "same" or a "different"

game? No standards exist for determining what is the "same" or "different."

We agree with many of these objections, though we see none of them as undermining our point that, overwhelmingly, the evidence shows that the Dartmouth and Princeton students saw the same game. Existence of bias is not evidence that bias swamps accuracy. There is nothing wrong with citing this study as evidence that "bias exists," though there is something quite wrong with concluding that a study showing that people saw the same game at least 95% of the time is evidence of radical constructivism.

It is also true that small effects can be important and the outcome of a game can hinge on a single play. This, however, does not justify reaching a conclusion that "there is no such thing as a game." That the *outcome* of the game might hinge on a single play does not mean that, *in general*, people saw "different" games. Judgments of "importance" are highly subjective; researchers have every right to conclude that this study (Hastorf & Cantril, 1954) discovered an "important" bias. They did not, however, discover a situation in which the subjectivity of perceptions exceeded, equaled, or even approximated the level of the objectivity of perception.

Lastly, lack of standards for what constitutes the "same" or "different" games is indeed a problem — but it is a problem with the original study, not with our analysis. In the original study, no a priori standards for what would constitute viewing the game as the same or different were articulated. By not articulating such standards, any result, even one showing that their participants viewed the game as the same at least 95% of the time, could be interpreted as meaning that they saw "different" games. Without an a priori standard for "sameness" or "differentness" modern readers are left to come up with their own.

In such situations, one option would be to use traditional standards in our field, thereby protecting researchers from the risk of confirmation bias manifested as setting up post hoc standards to advance their preferred narrative. For example, social psychologists routinely use multiple measures of a construct and combine them to form a scale. Although there are no hard and fast rules for doing so, social psychologists would typically feel justified combining two measures that correlated with one another $r = .90$ (which would produce a Cronbach's alpha of .95), which is well above conventional standards. If we applied the Hastorf and Cantril (1954) standard (that disagreeing 5% of the time means "different") to social psychological research, researchers would have to conclude that two variables were measuring "different" things if they "merely" correlated $r = .90$ with one another, because, as per a binomial effect size display, they disagree 5% of the time. This strikes us as unjustified, but it does nicely convey why, using any reasonable conventional standard common in social psychology (one could also use Cohen's (1988) standards for small, medium, and large effect sizes) leads to the conclusion that, in fact, the Dartmouth and Princeton students overwhelmingly saw the same game (Hastorf & Cantril, 1954).

4.2.5. Doing better: unmasking accuracy in studies of bias

Sometimes, research is designed only to test for biases in social perception, and is neither intended nor capable of assessing levels of accuracy, agreement, or unbiased responding. There is nothing inherently problematic with such research. It only becomes problematic when research that has only attempted to study bias, and has no information about accuracy or unbiased responding, is cited as a basis for reaching conclusions about the relative power of bias over accuracy or unbiased responding.

Over the last 30 years, more and more studies have been obtaining and reporting data capable of assessing levels of both unbiased (or accurate) and biased responding (see reviews by Jussim, 1991, 2012a,b; West & Kenny, 2011). As a result of longstanding controversies over whether stereotypes are mostly accurate or inaccurate (e.g., Allport, 1954/1979; McCauley, Stitt, & Segal, 1980) assessing both accuracy and inaccuracy is now routine in this area (see Jussim, Crawford & Rubinstein, 2015, for a review). For example, a study of Canadian ethnic stereotypes (Ashton & Esses, 1999) found that although about 30% more people exaggerated the real differences in educational achievement than underestimated them, people's beliefs about the groups' achievement generally correlated quite highly with board of education records ($r = .69$). If this study had only investigated bias (exaggeration), and not reported correlation accuracy, accuracy would have been masked by the reported results of bias, and it would likely have reached a distorted conclusion about degree of distortion vs. accuracy in those ethnic stereotypes.

4.3. But for stereotype threat, African-Americans and Whites would have equal standardized test scores (Steele & Aronson, 1995)

This was once a common interpretation of the early classic (Steele & Aronson, 1995) research on stereotype threat. However, continued group differences in achievement, even under the nonthreatening conditions, were masked by the presentation of evidence seeming to suggest achievement equality (see Sackett, Hardison, & Cullen, 2004, for a review). To understand how requires understanding sources of this widespread misinterpretation: A presentation of the original results that masked how achievement changed, and a plethora of technically correct but misleading interpretations that continues to the present.

4.3.1. The studies

Four experiments were reported, of which three examined the effects of racial stereotype threats on the test performance of African-American and White college students (Steele & Aronson, 1995). In addition, participants reported their SAT scores, which were included as a covariate when testing effects of threat vs. nonthreat on subsequent performance. Across the three studies, results consistently showed that, in the nonthreatening conditions, African-American and White covariate adjusted test score means were about the same, whereas, under threatening conditions, the typical racial difference in test performance emerged. These results were widely interpreted as evidence that, when threat was removed (Steele & Aronson, 1995), racial achievement test score differences were erased (e.g., American Psychological Association, 2006; see Sackett et al., 2004 for a review).

4.3.2. The misleading presentation

Although the text clearly states that an *analysis of covariance* was performed and that *adjusted means* were reported, their Figure 2 (re-created in our Fig. 1), which depicted the *covariate adjusted means*, has the X-axis labeled "Mean test performance Study 2" (p. 802). Absent a close reading of the text, and a sophisticated understanding of adjusted means in analysis of covariance, misinterpreting this figure is easy. Because the *adjusted means* for African-American and White students were nearly identical in the no threat (nondiagnostic test) condition, it is easy to come away with the *false* impression that these analyses showed that removing stereotype threat eliminated racial differences. They did not. When pre-existing participant group differences are equal across conditions, equal *adjusted* means in ANCOVA occur because pre-existing differences are unaffected by the manipulation, not because the means are equal. Equal *adjusted* means are not equivalent to *equal means*.

This unjustified interpretation was once widely advanced. In referring to the original study (Steele & Aronson, 1995), Aronson, Lustina, Good, Keough, Steele and Brown (1999, p. 30) claimed that African-American students performed "… about as well as Whites when the same test was presented as a nonevaluative problem solving task", and Wolfe and Spencer (1996, p. 180) declared that, "One simple adjustment to the situation (changing the description of the test) eliminated the performance differences between Whites and African-Americans."

After Sackett et al. (2004) pointed all this out, Steele and Aronson (2004, p. 48) acknowledged that the "gap was not eliminated" interpretation was indeed correct: "… in fact, without this [covariate] adjustment, they would be shown to perform still worse than Whites…"

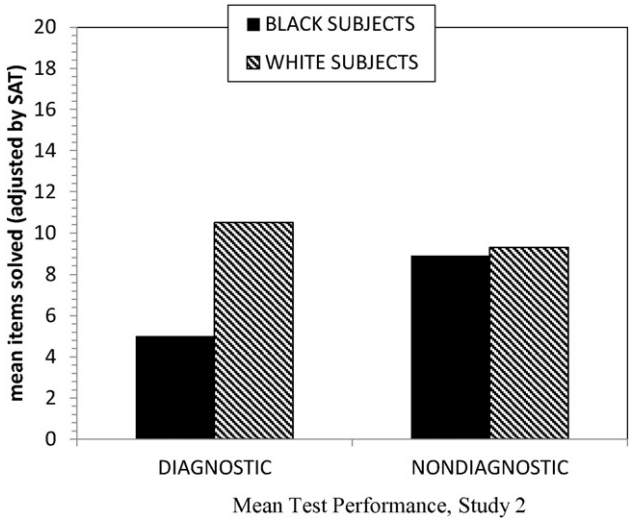

Fig. 1. Mean test performance, Study 2. Based on Figure 2 from Steele and Aronson (1995), page 802. The Figure 2 *caption* statement is technically incorrect (they are covariate adjusted means, not "mean test performance" scores, thereby rendering the figure deeply misleading). The nearly equal covariate adjusted means in the nondiagnostic condition *do not mean* that Blacks and Whites had equal scores. Instead, if random assignment succeeded at producing groups with no a priori differences in SAT scores, they mean that the *pre-existing differences* (of about 40 points) were maintained in the nondiagnostic condition. Stereotype threat increased achievement test differences; removing it did not reduce the mean differences between African-Americans and Whites.

and explained that an ANCOVA was conducted in order to reduce error variance. Although this is a valid use of ANCOVA, it does not justify the claim that removing threat eliminated the racial difference in test scores.[1]

4.3.3. Misleading presentations 2.0

One view of the exchange between Sackett et al. (2004) and Steele and Aronson (2004) is that this is science functioning well, self-correcting as errors are pointed out. Unfortunately, however, subsequent characterizations of Steele and Aronson's (1995) results did not change as much as this exchange suggests they should have. Eliminating racial achievement differences is clearly more of a Wow Effect than is exacerbating them (threatening conditions clearly exacerbated the pre-existing racial differences in the Steele & Aronson, 1995, studies). If researchers are motivated to promote the phenomena they study as "Wow Effects" one might predict that the misleading characterization of the original study (Steele & Aronson, 1995) would continue to appear even after Sackett et al.'s (2004) critique.

That is exactly what has happened, with one minor twist. To illustrate, Schmader, Johns, and Forbes (2008, p. 336) claimed that the original study (Steele & Aronson, 1995) showed that: "… African-American college students performed worse than their White peers on standardized test questions when this task was described to them as being diagnostic of their verbal ability but that their performance was equivalent to that of their White peers when the same questions were simply framed as an exercise in problem solving (and after accounting for prior SAT scores)." Similarly, Walton, Spencer, and Erman (2013, p. 5) wrote: "In a classic series of studies, Black students performed worse than White students on a GRE test described as evaluative of verbal ability, an arena in which Blacks are negatively stereotyped. But when the same test was described as nonevaluative — rendering the stereotype irrelevant — Blacks performed as well as Whites (controlling for SAT scores; Steele & Aronson, 1995)."

These statements are technically true, highly convoluted and not unique to these papers (see e.g., APA, 2006; Appel & Kronberger, 2012; Walton & Spencer, 2009). The language needs to be convoluted, because for the statements to be technically true, the declaration that African-American and White scores are "equivalent" in nonthreatening conditions needs to be walked back by adding the parenthetical regarding "controlling for prior SAT scores." The actual result — pre-existing differences continued even under no threat conditions — is never explicitly stated and remains hidden in these descriptions of Steele and Aronson (1995).

4.3.4. The misleading nature of declaring two groups equal controlling for prior differences

In general, declaring two groups "equal controlling for prior differences on the same variable" is meaningless. Fig. 2a shows the mean temperatures in Nome, Alaska, and Tampa, Florida, on 20 days scattered throughout the year 2014. Tampa was much warmer than Nome (means = 82.75 and 43.1°, respectively, $F(1, 38) = 79.68$, $p < .0001$). However, through ANCOVA, we can make this huge difference disappear by controlling for previous temperatures. We did so by selecting the day before each of the 20 days in the first analysis, and then using them as a covariate. Fig. 2b displays those results. The covariate-adjusted means now show no difference (means = 64.5 and 61.35 for Tampa and Nome, respectively, $F(1,37) = .49$, $p = .488$). Or, put differently, the following statement is just as technically true as the

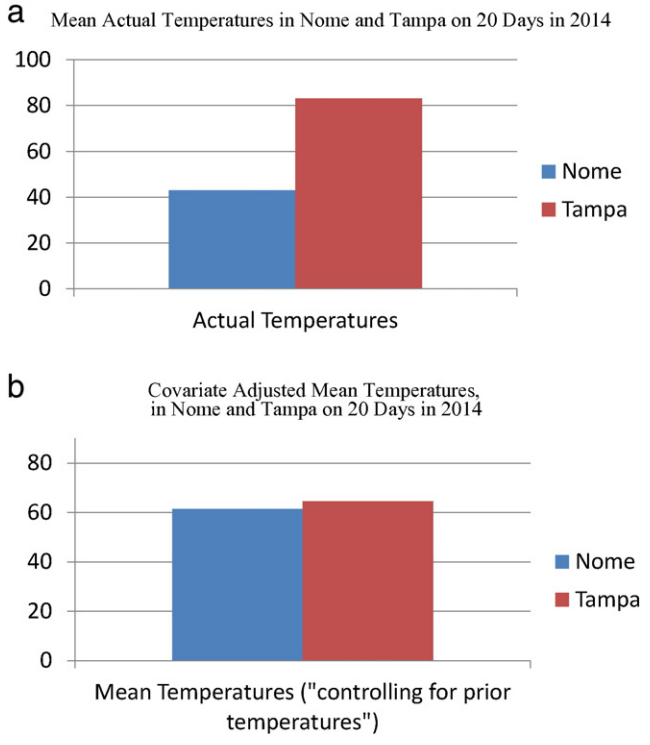

Fig. 2. a) Mean actual temperatures in Nome and Tampa on 20 days in 2014. b) Covariate adjusted mean temperatures, in Nome and Tampa on 20 days in 2014.

[1] In a true experiment in which participants are successfully randomly assigned to conditions, there should be little or no difference between the means of those conditions pre-manipulation. That is the entire point of random assignment, to equalize such pre-existing differences. In such a situation, equal *adjusted means* on the outcome (adjusting for the pre-existing differences) indicates that pre-existing differences between intact groups (say, different ethnic groups) were maintained, not that they were actually equal. There is a possible exception to this interpretation of ANCOVA adjusted means. Equal post-manipulation adjusted means (controlling for pre-manipulation means) might not equal the pre-manipulation difference if there was a failure of random assignment, and the *pre-manipulation* African-American/White means were already significantly different in the threat vs. the no threat, equal adjusted means would not necessarily entirely reflect the pre-existing differences. However, if random assignment failed, the studies could not be considered true experiments, thereby casting doubt on the ability to reach any causal interpretation of the results. The only resolution to this issue would be for Steele and Aronson (1995) to make their data publicly available so these issues could be explored. This strengthens our broader point calling for greater transparency.

statements quoted above regarding the original stereotype threat findings (Steele & Aronson, 1995): "The mean temperature in Nome is as high as the mean temperature in Tampa (controlling for prior differences)."

4.3.5. Objections

One potential objection to our analysis is that some reviews of stereotype threat now explicitly recognize that stereotype threat effects are quite modest and only explain, at most, a fraction of the racial achievement gap (e.g., Walton & Spencer, 2009). Another objection might be that even modest stereotype threat effects can be important. Both of these claims are valid, but they are irrelevant to our point — which is *not* that stereotype threat research is "invalid" or that the effects are unimportant. Indeed, we have not discussed *stereotype threat, in general*. Instead, our review has focused on the unjustified (pre-Sackett et al., 2004) and misleading (post-Sackett et al., 2004) claims routinely made specifically regarding the original findings (Steele & Aronson, 1995).

Another potential objection is that stereotype threat researchers have made lots of other, more valid claims, and routinely recognize, *in general*, that stereotype threat is only one of many contributions to racial and gender achievement gaps. This objection is also true (e.g., Schmader et al., 2008; Walton et al., 2013), and we return to this point in our next section. Regardless, valid claims about stereotype threat in general, or judgments of the phenomenon's importance, do not justify inaccurate or misleading representations of Steele and Aronson's (1995) findings.

4.3.6. How to do better: fully transparent data and valid interpretations

The pitfalls involved in appropriately interpreting analysis of covariance have been articulated elsewhere (e.g., Miller & Chapman, 2001; Yzerbyt, Muller, & Judd, 2004). These critiques make the following points: 1. Use of ANCOVA to compare naturally occurring groups can yield significant but spurious results, especially when the groups differ on the covariate; 2. Its use is often inappropriate if the covariate is not independent of the groups; and 3. If the covariate interacts with the grouping variables, or if such interaction is not tested, ANCOVA can produce misleading results. One easily understandable problem is that, if the groups differ substantially on the covariate, then adjusted means may be meaningless. Adjusted means are the predicted value (as in a regression equation) on the outcome for a particular group at the overall mean for the sample (grand mean, in an experimental design). If there is a large difference in scores on the covariate (as there was in Steele & Aronson, 1995), then "equal adjusted means" merely indicate that African-Americans in the upper end of the achievement distribution of African-Americans performed about the same as Whites in the lower end of the achievement distribution of Whites. If unadjusted means were presented, readers could then judge for themselves how meaningful and dramatic such findings are.

ANCOVA and related regression techniques are not inherently problematic and, for certain questions, used under the right conditions, and interpreted appropriately narrowly, can provide important insights. It is also clear, however, that such techniques can and do obscure much of the very phenomenon about which researchers aspire to reach conclusions (e.g., change in ethnic differences in achievement).

Therefore, the ideal stereotype threat report would provide all of the following information: Unadjusted means on all pre-manipulation measures; unadjusted means on all post-manipulation measures; covariate adjusted means; correlations among all measures, overall and within conditions; the standardized and unstandardized coefficients relating pre-manipulation achievement scores to post-manipulation achievement scores; and a test of whether that relationship differed by condition. Table 2 provides a set of simplified hypothetical examples (based solely on raw means) in which the pre- and post-threat manipulation means and standard deviations are made explicit, and communicates why this full set of information is critical for the appropriate interpretation of stereotype threat studies. One can rarely infer much, if anything, about the pattern of change produced by an experimental manipulation on the basis of adjusted means alone.

Unfortunately, research on the role of stereotype threat in racial and ethnic achievement almost never reports this information, and instead, typically reports adjusted means but not the unadjusted means or standard deviations for the outcomes and covariate (e.g., Aronson, Fried, & Good, 2002; Gonzales, Blanton, & Williams, 2002). One stereotype threat paper which presented results in a far more transparent manner than is typical (Kellow & Jones, 2008) simply reported the outcome means for African-American and White students when taking a test under evaluative (threatening) or non-evaluative (non-threatening) conditions (they did not use a covariate). Results were strikingly inconsistent with the interpretations usually reached on the basis of adjusted means studies, even though the racial difference in achievement was much larger under threat than under no threat. This occurred, not because African-American achievement rose to approximate that of Whites under nonevaluative conditions but because Whites' performance was much higher than all other groups in the evaluative condition. Stereotype threat raised Whites' achievement, whereas African-Americans performed similarly under threat and no threat conditions.

To be clear, we are not concluding that the Kellow and Jones' (2008) finding is generally true, or that it invalidates or alters the interpretation of other stereotype threat studies. We are simply holding this study up as an example of: 1. How easy it is to be transparent; and 2. How transparency can readily unmask alternative explanations for patterns of differences and nondifferences among adjusted means. Without such transparency, alternatives may be masked throughout the literature on the role of stereotype threat in the racial achievement gap. Identifying practices that can reduce achievement gaps is too important to be jeopardized by potentially faulty conclusions that could have been avoided with some basic transparency.

Last, presenting a clearly valid interpretation of Steele and Aronson (1995), without text implying that their nonthreat conditions eliminated the achievement gap, is not particularly difficult. Rather than the oft-used convoluted language necessary to render claims of racial equivalence in that study technically true, there is a simple, 12-word description of the study that is valid: "Steele and Aronson (1995) found that stereotype threat increased racial achievement differences." This description does not *deny* the existence of stereotype threat, and it does accurately describe the only unambiguous findings. Harackiewicz et al. (2014, p. 376), provide this model: "Numerous laboratory experiments have shown that minority group members (or women in math and science contexts) perform more poorly when told that a test is diagnostic of ability, or when stereotypes about their group are made salient, relative to nonevaluative, nondiagnostic, controls…" (which includes a long list of citations, including Steele & Aronson, 1995). This is simple, clear, and valid, and devoid of misleading claims about eliminating the gap "controlling for prior differences."

4.4. Curious cases of confusing correlations with means and distributions

4.4.1. Climate skeptics described as believing the Moon Landing was a hoax

"NASA Faked the Moon Landing — Therefore (Climate) Science is a Hoax" — is a title of a paper (Lewandowsky, Oberauer, & Gignac, 2013) that implies that people who doubt global warming believe conspiracy theories. The main hypothesis was that conspiracist ideation predicts skepticism of anthropogenic climate change. Evidence for these conclusions were data on 1145 respondents' beliefs in various conspiracies and their acceptance of science conclusions (HIV causes AIDs, burning fossil fuels increases atmospheric temperatures, etc.). These measures were subjected to latent variable modeling and did indeed indicate that "conspiracist ideation" negatively and significantly predicted acceptance of climate science.

Table 2
How transparency can clarify the meaning of stereotype threat results: simplified examples.
All hypothetical examples below present pre-manipulation SAT scores, and post-manipulation scores correct on a test conducted post-threat/no threat manipulations. The unadjusted post-manipulation means can be hidden beneath adjusted means and can reflect very different patterns of differences than do adjusted post-manipulation means. We do not present covariate adjusted means here because *no matter what the covariate adjusted means are,* the patterns shown here would not necessarily be revealed. These examples show the *necessity* of reporting unadjusted means to fully interpret stereotype threat effects.

Panel 2a: 1. Random assignment succeeded; 2. Ethnic differences were eliminated under no threat; and 3. Ethnic differences increased under threat. In these data, there is a 0.50 ethnic difference pre-manipulation; there is a 1.0SD difference post-manipulation in the threat conditions; and there is no difference in the no threat conditions.

	African-American no threat	White no threat	African-American threat	White threat
Pre-manipulation	450 (100)	500 (100)	450 (100)	500 (100)
Post-manipulation	15 (5)	15 (5)	10 (5)	15 (5)

Panel 2b: 1. Random assignment succeeded; 2. Ethnic differences were unchanged under no threat (0.50SD); and 3; Ethnic differences increased under threat from 0.50SD to 1.0SD (this is a stereotype threat effect in which threat increases ethnic differences, but no threat leaves the original ethnic differences unchanged)

	African-American no threat	White no threat	African-American threat	White threat
Pre-manipulation	450 (100)	500 (100)	450 (100)	500 (100)
Post-manipulation	12.5 (5)	15 (5)	10 (5)	15 (5)

Panel 2c: 1. Random assignment failed; and 2. Covariate adjusted means would obscure the fact that there is no post-manipulation difference in the size of the ethnic difference in the threat vs. no threat conditions, because the amount a mean of 350 is adjusted would not be the same as the amount that a mean of 450 was adjusted. Nonetheless, the no threat conditions reduced the initial ethnic difference from 1.50 SD to 1SD, whereas threat had increased the difference (there is a 0.50 SD ethnic difference pre-manipulation and a 1.0 SD difference post-manipulation.

	African-American no threat	White no threat	African-American threat	White threat
Pre-manipulation	350 (100)	500 (100)	450 (100)	500 (100)
Post-manipulation	10 (5)	15 (5)	10 (5)	15 (5)

Panel 2d: 1. Random assignment succeeded; 2. No threat reduced ethnic differences; and 3. Standardization renders amount of the reduction in ethnic differences ambiguous. In standardized units, no threat reduces a 1SD difference to a 0.5SD difference. However, the SD of on pre-manipulation scores of the sample is only 50, even though the SD of SAT scores in the population is 100, rendering interpretation of this result ambiguous. No threat cuts the standardized sample difference in half. However, the pre-manipulation sample difference, though a full standard deviation for the sample, is actually half a standard deviation for the population (in which 100, not 50, is the SD). The post-manipulation difference under no threat is half a standard deviation. The meaning is ambiguous, because it is not clear what the correct baseline comparison is: if to the sample, then removing threat reduced the achievement gap; if to the population, it did not because even post-manipulation, there is a 0.50 SD ethnic difference. A 0.50 SD difference in *population* SAT scores would be about 50 points -- which corresponds to the sample pre-manipulation difference.

	African-American no threat	White no threat	African-American threat	White threat
Pre-manipulation	450 (50)	500 (50)	450 (50)	500 (50)
Post-manipulation	12.50 (5)	15 (5)	10 (5)	15 (5)

4.4.2. 98% of climate skeptics did not believe the Moon Landing was a hoax

Latent variable modeling masked the invalidity of the titular implication that climate skeptics tend to believe in silly conspiracy theories. The invalidity of this conclusion cannot be found in the structural equation model results; it can, however, be found in the simple distribution of responses. In the sample of 1145, only ten participants endorsed the moon-landing hoax. Of the 134 who believed climate science was a hoax, only three endorsed the moon-landing hoax (on a four-point scale of Strongly Disagree, Disagree, Agree, and Strongly Agree, we are treating both "agree" and "strongly agree" responses as agreement). Thus, almost no one, including those who rejected climate science, believed the moon landing was a hoax.

The abstract reported that "Endorsement of free markets also predicted the rejection of other established scientific findings, such as the facts that HIV causes AIDS and that smoking causes lung cancer." However, only 16 participants in their sample of 1145 rejected the fact that HIV causes AIDS, and only 11 participants rejected the fact that smoking causes lung cancer. There were 176 free market endorsers in their sample. Nine of them rejected the HIV–AIDS link, and seven of them rejected the smoking–lung cancer link. Thus, 95% and 96% of free market endorsers agreed with those scientific facts.

The structural equation modeling performed was a sophisticated set of analyses (Lewandowsky et al., 2013). Interpretations of such analyses as evidence that climate skeptics believe in silly conspiracy theories conflate the sign of the correlational results with participants' actual placement on the items. Correlations resulted from covariance in levels of explicit agreement with reasonable positions (e.g., disbelieving the moon landing hoax and disbelieving that climate science is a hoax — see Table 3). It would be fair to characterize their results as indicating "the more strongly people disbelieved hoaxes, the more strongly they

Table 3
Almost no one believed the moon landing was faked.
Data from Lewandowsky et al. (2013).

		The moon landing was faked			
		Strongly disagree	Disagree	Agree	Strongly agree
Global warming is a hoax	Strongly disagree	892	39	2	2
	Disagree	53	20	1	2
	Agree	65	5	0	0
	Strongly agree	57	4	1	2

10/1145 believed the moon landing was faked. 134 believed global warming is a hoax; 3 of them believe the moon landing was a hoax. The correlation is nonzero almost entirely because there is covariance among the reasonable positions (disagreeing that the moon landing was faked and that global warming is hoax). There is no evidence here that people who believe global warming is a hoax were also more likely to believe the moon landing was faked. Also, these data are so skewed, and have so few response options, that it is not clear that the type of structural equation models used in the original report are appropriate.

believed in climate science," but too few people actually believed in hoaxes to warrant reaching *any* conclusions about them. Similar patterns occurred for the other conspiracy beliefs.

4.4.3. When racial prejudice IAT scores predict anti-White behavior

This problem of conflating correlations with levels of a construct is not unique to research on global warming beliefs. One early study reported that the IAT predicted anti-Black discrimination, because the IAT correlated about $r = .30$ with discrimination (McConnell & Leibold, 2001). However, a simple scatterplot of the data (Blanton et al., 2009) showed that there was almost no evidence of anti-Black discrimination. Instead, most participants treated the African-American target more positively than they treated the White target, and most of the remainder treated targets nearly equally. The correlation occurred because higher IAT scores corresponded to *egalitarian behavior*, and lower scores corresponded to anti-White behavior. It cannot be concluded that the IAT predicted anti-Black discrimination among data in which nearly all behavior was pro-Black or egalitarian.

4.4.4. Objections

Climate skepticism is inconsistent with a mountain of evidence indicating both that the Earth is warming and that human activity has contributed to it. Clearly, the study of why people maintain erroneous beliefs in light of such evidence is an interesting and important psychological topic. Racial prejudice is also an ongoing social problem, and understanding the role of unconscious prejudice in discrimination is also extremely important. The present paper does not attempt to adjudicate controversies about the ability of the IAT to predict discrimination more broadly (contrast, e.g., Greenwald, Poehlman, Uhlmann, & Banaji, 2009 with Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2013). However important it may be to study such phenomena, they do not they justify reaching invalid conclusions about the results of particular studies investigating beliefs about global warming or about prejudice and discrimination.

4.4.5. How to unmask data inconsistent with one's preferred story: simple analyses and transparency

There are many examples in which researchers perform correlational or structural equation analyses after also providing basic descriptive statistics on the variables included in their models (e.g., Caprara, Alesandri, & Eisenberg, 2012; Guimond et al., 2013). Such a practice provides insurance against misinterpreting correlational relationships as indicating the absolute levels of each variable, because the absolute levels are then readily apparent.

Faulty conclusions drawn from correlations and structural equation models masking means and distributions could be avoided with greater detail and transparency (we prefer them to be available in the main report, but, if necessary, at least in supplementary materials). If simple frequencies and descriptive statistics had been reported, the fact that almost no one actually believed the moon landing hoax (Lewandowsky et al., 2013) would have been far more apparent. It is clear that both sophisticated (e.g., SEM, HLM) and simple (e.g., correlation) analytic techniques can obscure fundamental patterns in the data that can and should substantially influence how those results are interpreted. We are not arguing against the use of sophisticated statistical techniques; rather, in general, authors should also provide basic descriptives, frequencies, correlations, standard deviations, ranges, and unadjusted cell means for anything presented in a research report. Scatterplots will also often be very revealing. And if authors do not provide them, reviewers and editors should request that they be provided.

4.5. The power of the situation

Social psychologists have long emphasized the power of the situation (see reviews by Ross & Nisbett, 1991; Ross et al., 2010). For example, one highly cited work (Ross & Nisbett, 1991) has section headings titled, "The Weakness of Individual Differences" (p. 2) and "The Power of Situations" (p. 3). Others, too, have reached similar conclusions: "The first century of experimental social psychology, then, has been devoted largely to demonstrating *the power of construal and the power of the social situation*…" (Jost & Kruglanski, 2002, p. 172, emphasis in original).

At first glance, this probably seems reasonable, because social psychologists have indeed discovered some extraordinarily powerful situations (see any introductory social psychology text for examples involving conformity, obedience, helping, and more) and testaments to the power of the situation can be found throughout social psychological scholarship. Furthermore, one of the field's most classic discoveries, the fundamental attribution error, provides abundant evidence that people often underestimate the power of situations (Ross, 1977).

There are, however, several problems with this perspective, if it is construed to mean that situations are far more powerful than individual differences. First, some of the supposedly most powerful situations actually reveal maximal individual differences (see Krueger & Funder, 2004). The exemplar for this conclusion is one of the most dramatic "power of the situation" phenomena in social psychology — obedience to authority. Milgram's (1974) studies found unexpectedly high levels of willingness to shock "learners" among the "teachers." When situations dominate behavior, most people act the same in that situation; when individual differences dominate, people act differently. For a dichotomous outcome (such as "going all the way up to 450 volts or not"), maximal individual differences occur when half the people do and half do not (meaning that knowing the situation does not allow you to predict behavior better than a coin flip).

Across many variations of the study in which the teacher had to flip the shock switch, willingness to go all the way to 450 V hovered around 50–60% — near the point of maximal individual variation (see also Krueger, 2009; Krueger & Funder, 2004 for more detailed expositions of this point, which are also then applied to research on conformity, roles, and helping in emergencies).

Second, the evidence that laypeople underestimate the power of situations is not quite as dramatic as implied by the "Wow Effect" version of this story. Even though people often do make correspondent inferences, the evidence that people actually systematically or generally underestimate the power of situations is weak and inconsistent (Gawronski, 2004; Malle, 2006). Furthermore, the related actor–observer difference has been shown to have an effect size of essentially zero (Malle, 2006).

As important as all these critiques may be, even more important is that scholarship emphasizing the power of situations rarely, if ever, has explicitly reported and compared effect sizes for situations versus persons. Thus, powerful dispositional influences potentially — and, as it turns out, actually — were masked by the bona fide evidence of powerful situations. Masking occurred because social psychologists emphasizing the power of situations held to a particular narrative, and they did not investigate or report the power of individual differences to predict behavior. Empirical investigations that *have* compared the power of situations to the power of individual differences in predicting behavior have found the effect sizes to be of similar magnitude (e.g., Fleeson, 2001; see reviews by Fleeson, 2004; Funder, 2008). Furthermore, whereas situations do influence people's momentary behaviors (i.e., people show considerable variability in behavior across situations), personality traits are better predictors of people's typical behavior over time (Fleeson, 2004).

4.5.1. Objections

One might object to our analysis on many grounds. First, we have not reviewed much of the literature demonstrating situational influences on people's behavior. Second, we have not provided a thorough review of the person–situation debate.

These objections are valid. There are many extensive reviews of those issues in the literature (e.g., Fleeson, 2004; Funder, 2008;

Krueger & Funder, 2004; Ross et al., 2010), and reviewing those literatures is not our purpose. Instead, we mean to point out that: 1) much modern scholarship *still* emphasizes the power of situations relative to persons; 2) the scholarship that does so has, as far as we can tell, *never* reported effect sizes for persons and situations; thereby 3) Permitting evidence of person effects that are as strong as situational effects to be masked by strong claims emphasizing powerful situations.

4.5.2. "Towards a balanced social psychology" (Krueger & Funder, 2004, article title)

Scholarship that acknowledges and reports effects for both persons and situations is generally more balanced, and more valid, than scholarship that does not. Indeed, it is only by virtue of a blind spot with respect to a large and growing literature that has demonstrated the cross-situational power of individual differences that it is possible to maintain the classic "situationist" perspective in social psychology that individual differences hardly matter.

Scientists should not be in the business of simply ignoring literature that they do not like because it contests their view. We are not attempting to adjudicate the continuing debate between those arguing that situational effects swamp person effects (compare, e.g., Bargh, 2007 with Krueger, 2009). Perhaps it is possible for the strong situationist perspective to be plausibly defended, even after acknowledging the now-large literature showing that person effects approximately equal situation effects in magnitude. Our only point is that, to reach *any* justified conclusions about the power of situations relative to persons, *one must actually review the evidence bearing on both issues*, and at least report effect sizes for each. Equal effect sizes do not automatically *dictate* that scientists are compelled to reach the conclusion that two effects are actually equal in the wider world — that conclusion hinges on many considerations beyond the particular studies that have been conducted. Nonetheless, our view is that *overlooking a large body of research that appears to directly conflict with one's conclusions* is a problematic practice whenever it occurs. And the solution is simple — cite it, grapple with it, and, if one is claiming one effect is stronger than another, report effect sizes for both.

4.6. Citation practices: the masking (and unmasking) of findings that do not fit the "story"

4.6.1. The saga of "stereotypes lead to their own confirmation"

This was the conclusion reached in a study showing that: 1) There was no stereotype bias in person perception in the absence of individuating information; and 2) There was a stereotype bias in the presence of individuating information (Darley & Gross, 1983). This conclusion has been so widely embraced by social psychologists that the paper has been cited over 1000 times according to Google Scholar.

A problem, however, appears in 1996, because in 1995, failed replications were published (Baron, Albright, & Malloy, 1995). This was quite striking for several reasons. First, they obtained the original stimulus materials, so this was an attempt to closely follow the original procedures. Second, not one, but two successive failed replication attempts (Ns = 81 and 80, respectively; N = 67 in Darley & Gross, 1983) were reported (Baron et al., 1995). Third, rather than simply producing null results, the findings were statistically significant in the *opposite direction* of the original. In short, the failed replications (Baron et al., 1995) found (twice, with a total of more than twice as many participants) that stereotypes biased person perception in the absence but not presence of individuating information.

One might expect most scientists publishing on these issues after 1995 to cite both papers in an attempt to grapple with the inconsistent results. Instead, since 1996, the original study has been cited 852 times, while the failed replications have been cited just 38 times (according to Google Scholar searches conducted on 9/11/15). This means that nearly all discussions of the original study since 1996 have simply overlooked the failed replications. We recognize that it is not possible for every researcher to be aware of every study that has ever been published in their field. However, 852 vs. 34 is not random variation in awareness. *JPSP* does have an impact factor about twice that of *PSPB*, which, perhaps, could explain about a 2:1 citation advantage to Darley and Gross (1983), but not a 22:1 advantage.

This citation pattern is common. Failed replications (whether exact or conceptual) often receive a fraction of the citations of the original narrative. Other sorts of correctives — meta-analyses that include a wider range of studies, failed "conceptual replications" showing that the original finding may be restricted to extremely limited conditions — have a similar fate. Table 4 displays some examples of papers where: 1) an initial paper had high impact; 2) follow-up research was performed using stronger methodological standards (e.g., in each case in Table 4, the follow-ups research had larger, and often much larger, samples); and 3) the follow-ups continued to be mostly overlooked and the originals extensively cited even after the follow-ups were published. If social psychology is to become a self-correcting science, these blind spots need to be uncovered. Or, as Gelman (2015) put it: "Don't privilege something that happens to have been published once and declare it true. If you do that, and you follow up by denying the uncertainty that is revealed by failed replications … well, then you're offering nothing more than complacent happy talk."

4.6.2. Objections

One objection to our analysis is that failed replications do not necessarily invalidate the original findings. We agree, and have not argued otherwise. However, simply *ignoring* the failed replications should not be an option either.

Another objection is that it is not the original authors' fault if research failing to replicate their results is not cited. We agree; failure to cite subsequent failed replications is not a problem with the original study, but a broader field problem. In each case, literally hundreds of papers either intentionally ignore the failed replications (e.g., because the failures conflict with the "narrative") or they simply marched on failing to acknowledge the doubts or qualifications raised by the failures. Hundreds of papers per topic imply blind spots among hundreds, possibly thousands of social psychologists. True sciences do not act as if data that conflicts with a preferred narrative simply do not exist.

Another objection is that there is so much literature out there that one cannot know about every failed replication that ever gets produced. This, too, is valid. It is, however, one thing to not know about *every* failed replication, and another to not know about *any*. If most social psychologists know about *most* failed replications in the areas that they are writing about (and in which, presumably, they are experts), and chose not to ignore them, one might get citation ratios of 1.1:1, or 1.5:1, but one would not get citation ratios of 10:1, or worse, for originals vs. failed replications.

Finally, another objection is that research can be cited for many reasons, so that there is no simple, straightforward interpretation of citation patterns. This, too, is undoubtedly true. However, a paper is most typically cited in support of some claim. If, however, that claim or its generalizability is called into question by failed replications, then at least some of those failed replications need to be also cited and discussed.

4.6.3. Beyond cherrypicking

These examples reflect a broad problem in the field rather than anything wrong with the original research reports. This pattern of ignoring correctives likely leads social psychology to overstate the extent to which evidence supports the original study's conclusions. For example, publications that cite *only* Darley and Gross (1983), often emphasize the power of stereotypes to bias person perception judgment (e.g., Brown, 2011; Greenwald & Pettigrew, 2014). There are very few publications that cite both Darley and Gross (1983) and Baron et al. (1995). Those that do (e.g., Jussim, 2012a,b; Regner, Huguet, & Monteil, 2002) do not

Table 4
Social psychological self-correction?

Publication	Narrative	Key aspects of methods	Citations	
			Total	Since 1996
Darley and Gross (1983)	Stereotypes lead to their own confirmation; stereotype bias in the presence but not absence of individuating information	People judge targets with vs. without relevant individuating information. Single experiment. N = 59–68, depending on analysis.	1054	853
Baron et al. (1995)	Failed replication of Darley & Gross, 1983. Positive results in the opposite direction: stereotype bias in the absence of individuating information; the presence of individuating information eliminated stereotype bias.	Close replication (and extension) of Darley & Gross, 1983. Two experiments. Total N = 161.	41	38
			Total	Since 2011
Jost, Glaser, Kruglanski, and Sulloway (2003)	Conservativism is a syndrome characterized by rigidity, dogmatism, prejudice, and fear	Meta-analysis of 88 studies, including two unpublished studies. No articulation of study selection criteria.	1920	1030
Van Hiel, Onraet, and De Pauw (2010)	Liberal/conservative psychological differences in cognitive style were modest to nonexistent.	Meta-analysis of 124 studies, including five unpublished studies. Clear articulation of study selection criteria.	67	60
			Total	Since 2013
Bargh, Chen, and Burrows (1996)	Automatic effects of stereotypes on behavior.	Two experiments. Total N = 60.	3717	900
Doyen, Klein, Pichon, and Cleeremans (2012)	Failed replication of Bargh et al. (1996). No effects of stereotypes on behavior except when experimenters were not blind to condition	Two close replication and extension experiments. Total N = 170.	212	194
			Total	Since 1984
Snyder and Swann (1978)	People seek to confirm their interpersonal expectations	Four experiments. Total N = 198. People had to choose among confirmatory or disconfirmatory leading questions (no option was provided for asking diagnostic questions)	915	841
Trope and Bassok (1983)	People rarely seek to confirm their interpersonal expectations. Instead, they seek diagnostic information.	Three experiments. Total N = 342. People could seek information varying in the extent to which it was diagnostic versus confirmatory.	131	126

Citation counts were obtained from Google Scholar between 9/22/15 and 10/3/15.

declare the results of the original studies "false" because of the failed replications. Rather, they are more circumspect, nuanced, and two-sided than papers promoting the conclusions reached in the original studies. For example, Regner et al. (2002, p. 254) wrote that: "The generalizability of the SES bias, therefore, remains unclear." Awareness and acknowledgement of this type of failed replication improves the quality of scientific claims by justifiably raising doubts about what can be concluded. Future studies might ultimately support the broad generalizability of Darley and Gross's (1983) findings, but, until such research is actually produced, it behooves researchers to grapple with the full literature, not just the studies conducive to their preferred arguments.

The incentives that reward the telling of compelling narratives in social psychological scholarship encourage cherrypicking. To some extent, the practice of cherrypicking presents a classic social dilemma: whereas it is in most individual scientists' self-interest to tell compelling stories (facilitated by cherrypicking), it is clearly not in the interest of the field of social psychology as it undermines the field's validity and credibility.

We anticipate several rewards for telling far less compelling narratives based on messy and contradictory data. First, we maintain our own scientific integrity. Second, we maintain the integrity of our field. Third, acknowledgement of conflicting results and messy data provides an opportunity for theoretical advance and new empirical research to resolve those conflicts, either by showing that one set of results are irreplicable, or by identifying conditions under which both sets of conflicting findings can be consistently obtained. Thus, the more traditional rewards may then become available to the researcher capable of resolving such conflicts.

Regardless, with respect to practices that can elevate the validity and credibility of social psychology, failed replications (especially if published), corrective reviews, and meta-analyses need to be acknowledged. Such failures and correctives are, themselves, not immune to criticism, and scientists may differ in the credibility they give to original studies vs. subsequent potential correctives. Thus, we are not arguing for a particular *outcome* — for researchers to give more weight to failures or meta-analyses providing evidence of weak effects or otherwise distorted literatures. We are merely arguing for a *process* that acknowledges and wrestles with data that does not comport with one's preferred narrative.

The problem of researchers simply not being aware of failed replications is a thorny problem. How can one cite Baron et al. (1995) if one does not know it exists? Obviously, one cannot, so the solution involves the answer to a different question: How can researchers raise their awareness of the existence of failed replications and other scholarship that indicates prior widely accepted conclusions may not be correct?

This difficulty has been compounded by the following historical pattern: 1. Until recently, it has been very difficult to publish failed replications, in part, because editors would often send the failure to the scientists authoring the original — who then has a vested interest in evaluating the failure negatively and may be motivated to block publication (e.g., Funder, 2012); so that 2. In order to publish, some nonreplicators would often frame their studies in ways designed to be so camouflaged or inoffensive that the fact that the study is a failed replication (either direct or conceptual) is not apparent from the types of information that readily comes up in a search (e.g., title or abstract), and is only apparent upon a close reading of the paper.

Fortunately, it has become somewhat easier to identify published failed replications over the last few years as recognition of the importance and acceptance of replication attempts have increased. In addition to widely publicized multi-site attempts to replicate many different studies (e.g., Open Science Collaboration, 2015), resources for listing

original studies followed by published replication attempts in general (e.g., curatescience.org and www.http://bps.stanford.edu) and in specific topic areas (e.g., Jussim, 2012a,b) have begun to become more readily accessible. It behooves researchers purporting to be experts on a topic to make strong efforts to scour the literature for failed exact and conceptual replications, and for meta-analyses and reviews reaching opposing conclusions.

5. Meet the new New Look; same as the old New Look?

Evidence that what we think or desire alters our sensory perceptions is *still* a Wow Effect even 60 years after the New Look. Recently, there has been a flood of research (over 160 papers since 1995; see Firestone & Scholl's, in press, review) on top-down effects on visual perception. The term "top-down" may mean many different things, but Firestone & Scholl (in press, manuscript page 12) make it clear that they are focusing only on the most dramatic meaning — "…the provocative claim that our beliefs, desires, emotions, actions, and even the languages we speak can directly influence what we see." Firestone and Scholl's (in press) review, however, concludes that the modern research fails to actually demonstrate such effects. Claims of top down effects, in their analysis, are only reached through a combination of confirmation biases and suboptimal practices (which they refer to as "pitfalls"). Two examples will have to suffice here (see Firestone & Scholl, in press for many more).

5.1. Top down effects on visual sensory perception: two examples

In one study (Banerjee, Chatterjee, & Sinha, 2012), participants first described one of their own past ethical or unethical behaviors. Those describing an unethical behavior rated the room as darker on a 7 point rating scale. The paper was titled, "Is it Light or Dark? Recalling Moral Behavior Changes Perceptions of Brightness" and the conclusion reached was that: "… metaphorical associations go beyond mere linguistic coupling to influence the actual perception of the physical world — the perception of light" (p. 408).

Other studies have found that, when asked to categorize briefly-presented stimuli as words or non-words, participants categorize moral words (e.g., justice, crime, guilty) more accurately than non-moral words (akin to perceptual vigilance; Gantman & Van Bavel, 2014), a phenomenon they termed the "moral pop out effect" to capture the claim that moral stimuli are especially salient in the visual field. This is a Wow Effect because it is often interpreted as evidence that moral stimuli are privileged in the mind and thus are more readily visually perceived: "The current research suggests that moral concerns shape our basic awareness of perceptually ambiguous stimuli" (Gantman & Van Bavel, 2014, p. 29).

5.2. No Wow Effects after all

A paper amusingly titled "Top-down effects where none should be found" (Firestone & Scholl, 2014) made use of the *El Greco fallacy* to identify and test why the initial paper (Banerjee et al., 2012) failed to unmask a non-perceptual explanation. El Greco was an artist whose human figures were strangely elongated. One explanation offered was that he suffered severe astigmatism, which visually stretched his visual environment, and he simply painted what he saw. Although this sounds reasonable, this analysis suffers a logical flaw. If El Greco perceived an elongated world and painted what he saw, *everything should have been equally stretched out, including his painting canvasses, so that, in the end, the distortions would have canceled out*. If he saw a 10 in. head as 15 in., he would have seen 10 in. of canvass as 15 in. long — so the head would have been painted subjectively (to him) 15 in., but objectively, 10 in. long on the canvass. This would produce no distortion.

In much the same manner, if rating the room as darker after thinking about unethical actions was truly a *perceptual* effect, then *everything* should be seen as darker, not just the room. Thus, the effect should disappear when, rather than using a darkness numerical rating scale, participants must choose a shade of darkness from a range of grayscale patches. The effect should disappear because, *everything* should look darker, including the grayscale patches, thereby leading to *equal* ratings of darkness across conditions. In contrast to the claim that visual *perception* was actually affected, participants *still* chose darker grayscale patches after considering unethical actions (Firestone & Scholl, 2014). This means that thinking about unethical actions led people to choose darker options, but that alteration of actual visual perception could not have caused that result (because alteration of actual visual perception would have caused *no difference* in ratings of darkness across conditions). In our terms, influences on something other than *visual perception* were masked by the experimental demonstrations, because, as they put it (Firestone & Scholl, in press), those promoting top-down explanations used overly confirmatory research strategies — they sought to find evidence of the phenomenon when their theory says it should appear, but *failed* to seek evidence of the absence of the phenomenon when their theory says it should *not* appear. Thus, the invalidity of the hypothesis that unethical actions influence visual perception was *masked* by its success at predicting that the room would be rated as darker and by the *lack* of a skeptical test of conditions under which the effect should disappear in the original studies.

With the respect to the "moral pop-out" studies (Gantman & Van Bavel, 2014), semantic relatedness is confounded with the manipulation, such that spreading activation in semantic memory (a bottom-up process) can explain apparent top-down effects on perception. Moral words may have been easier to recognize because of semantic priming; moral words were semantically related to each other, whereas non-moral words were not. Indeed, a paper titled, "Enhanced Visual Awareness For Morality And Pajamas?" showed that, when other non-moral categories (rather than unrelated assortments) of words are used (e.g., clothing words or transportation-related words), such trivial but categorically-related words also "pop out" more readily than do unrelated words (Firestone & Scholl, 2015). Thus "pop out" effects can be attributed to semantic priming rather than to morality having some sort of unique influence on sensory perception.

5.3. The quest for clear evidence of top-down effects on sensory perception

Firestone and Scholl's (in press) critique is not restricted to studies of the influence of morals on perception. Supposed top down influences of target race on perceptions of darkness of skin tone, influences of labels on perceptions of nonsense and ambiguous stimuli, the effect of task difficulty (e.g., wearing heavy backpacks) on perceived distance, and many more have all been shown to arise from processes other than perception (including, but not restricted to, experimental demand characteristics, attention, and features of stimuli confounded with judgments).

Demonstrating top-down effects on visual perception would be a true Wow Effect, one that would change psychological orthodoxy regarding organization of the mind. That research has, so far, failed to provide evidence of such effects does not mean claims of top-down influences are inherently false. Perhaps someday, such evidence will be produced. For now, though, as Firestone & Scholl (in press, page 6, manuscript version) put it: "… there is in fact *no* evidence for such top-down effects of cognition on visual perception…" (emphasis in original).

It is, perhaps, worth noting here that *none* of the unjustified conclusions that have emerged from this literature were seen to be a result of questionable research practices. Firestone and Scholl (in press) never disputed the data, and, in some cases they have themselves replicated the original studies' results. The *results* were generally sound. The *interpretations* of the results were not. Much as with the original New Look, research that more thoroughly seeks falsification and does a much better job ruling out non-perceptual alternative explanations may, in

the future, provide evidence for top-down effects of cognition on perception.

6. How to limit misinterpreting data and leaving valid phenomena masked

> "*It Ain't What You Don't Know That Gets You Into Trouble. It's What You Know For Sure That Just Ain't So.*"
>
> [(Mark Twain)]

The nature of scientific progress is that we will get many things wrong. A healthy science, however, will: 1. keep such errors to a minimum; and 2. quickly self-correct when errors have been made. Limiting its practices that camouflage true phenomena in the name of promoting Wow Effects and preferred narratives is one way to accomplish these goals.

The present paper has attempted to contribute to providing such correctives in two ways. First, we have highlighted how the widespread *interpretations* of many studies are not justified. Results long interpreted as testaments to the power of subjectivity and bias reflect far more evidence of objectivity and agreement; famous studies advancing egalitarian narratives actually provided evidence of continuing, not eliminated, achievement gaps; studies claiming to have identified predictors of beliefs in bizarre hoaxes or of bias against African-Americans actually found no evidence of hoaxes or bias to be predicted; the power of situations over persons are extolled without reporting actual effect sizes (which turn out to be similar); citation practices reflect widespread failure to acknowledge failed replications and other correctives; and supposedly dramatic evidence of emotions, motivations, and other top-down processes influencing sensory perception have, thus far, always been explained by non-perceptual phenomena. These errors occurred, not because of p-hacking or questionable research practices, but because of unjustified interpretations.

In addition to providing specific correctives to particular common but unjustified conclusions, this paper has also reviewed: 1. How and why such interpretations routinely go wrong; and 2. Practices in the published literature that model the way towards more valid interpretations. We conclude our paper with specific recommendations for improving researcher interpretations of evidence.

6.1. Archive of replication attempts and outcomes

Given that researchers' conclusions are often justifiably more nuanced and tentative when they are aware that a study has had difficulty replicating, it seems that, in the interest of facilitating self-correction, it is in social psychology's interest to make it easier for researchers to discover failed replications, especially of famous and influential studies. Thus, we have a very specific recommendation. The field could benefit from a well-organized public archive of studies and the success or failure of exact or close replications. Such attempts by individuals and small groups already exist (psychfiledrawer.com, curatescience.org, http://bps.stanford.edu/?page_id=2367), but they are haphazard and incomplete. Filedrawer.com archives unpublished studies, which is valuable, but, given the Darley and Gross (1983) vs. Baron et al. (1995) citation patterns, we would argue that the first step would be to archive *published* replications. To some extent, this is the purpose of reviews and meta-analyses, which undoubtedly perform a service for the areas being meta-analyzed. Some findings, however, became extremely influential despite few attempts at replication, and, in some cases, overlooked failed replications. Such an archive could reduce such scientific errors and facilitate self-correction. As the archive expands, more distant attempts at replication, "conceptual replications," can be added. Indeed, such archives would also greatly facilitate the conduct of meta-analyses by putting many such sources in a single place.

6.2. The journal of strong falsification

Recent research highlights many similarities among, rather than differences between, liberals and conservatives (see Brandt, Reyna, Chambers, Crawford, & Wetherell, 2014 for a review). This work has embraced a fundamental principle of philosophy of science that appears to be rarely practiced — attempting to falsify hypotheses (Popper, 1934; in this case, attempting to falsify longstanding claims that conservatives are fundamentally different than liberals — more biased, more prejudiced, etc.). It is generally very difficult to falsify psychological *theories* (e.g., Meehl, 1990). It is, however, usually far less difficult to falsify specific empirical *predictions* for a particular study, including a meta-analytic study. How would one seek falsification that social perception is so subjective that there is no objective reality? We doubt one could. But one could seek to falsify the claim that "there was far more bias than agreement in Princeton and Dartmouth students' perceptions of a controversial football game in 1951." One probably cannot falsify the general claim that "stereotype threat affects the performance of stigmatized students," but one can seek to falsify the claim that "African-American and White students had equal test scores in the non-threatening conditions of Steele and Aronson's (1995) studies." One probably cannot falsify that claim that "top down processes influence perception" but one can attempt to falsify the claim that thinking about unethical acts led people to actually perceive the world as darker in a particular study. The more that strong attempts at falsification themselves fail to actually falsify a prediction, the more confidence we can have in the validity of that prediction (see Gildersleeve, Haselton, & Fales, 2014, for a good example).

Journals have wide latitude to experiment with different approaches to upgrading the validity and robustness of the research that they publish. One possibility (suggested by a reviewer of an earlier version of this manuscript) would be for one or more journals to adopt constraints on overpromotion of "Wow Effects" by requiring authors to articulate a theoretical basis for opposite predictions and opposite interpretations for the same prediction, and then either test them or explain why they were not tested. Similarly, authors could be required to articulate likely boundary conditions and moderators and, again, either test them or explain why they did not. It is possible that many of the problematic interpretations highlighted here would have been avoided had such procedures been adopted. The sooner researchers recognize that some effect is not large or dramatic, or only holds under certain conditions, even if it is real, the sooner they will realize they need to run very high powered and nuanced studies to capture them. As such, social psychological research may become more valid, robust, and replicable.

This has never been tried, and it is always possible that such a policy would produce more unintended negative consequences than it would solve problems. The effectiveness of such a policy is an empirical question (although one that probably could not be answered for a very long time, if at all), in the same way that the effectiveness of *Basic and Applied Social Psychology*'s ban on significance testing is an empirical question. But the only way we will get answers to such questions is by putting them to empirical test.

6.3. (More reasons to) report effect sizes and confidence intervals

Many psychology journals are now emphasizing the importance of reporting effect sizes (e.g., rs, ds, or the original metric) and confidence intervals (CIs) to support replication, meta-analysis, and concerns about null hypothesis significance testing (Funder et al., 2014). We add another reason here: reporting effect sizes and CIs can help reduce the overclaiming that results from motivated reasoning and the desire to promote weak, mixed, or messy results as Wow Effects.

It becomes harder to promote testaments to the power of some phenomenon when one is also compelled to report small effect sizes with very large CIs. This is especially important for *major reviews* that appear in influential outlets. When one is making *relative* claims (e.g., the

power of situations versus person effects; the power of biased versus unbiased social perceptions), one should provide some *comparison* of the effect sizes. This is *especially* true of narrative and theoretical reviews. Such reporting is routine in meta-analysis, but is absent from many narrative reviews (a simple skim through *Handbook of Social Psychology* or *Annual Review* chapters will reveal vanishingly few reports of actual effect sizes, and even fewer confidence intervals).

Without the need to compare effect sizes, psychologists may (to paraphrase Churchill commenting on a very different context), "Occasionally stumble over the truth, but pick themselves up and hurry off as if nothing had happened." This becomes much more difficult when effect sizes are reported. Comparing effect sizes for bias versus agreement could have prevented social psychology from failing to recognize that the participants in Hastorf and Cantril's (1954) study overwhelmingly saw the same game. It might have prevented social psychologists from implying that the power of situations to predict behavior exceeds that of individual differences. Indeed, as meta-analyses have come in showing that stereotype threat effects explain, at most, a modest fraction of the achievement gap between African-American and White students, soaring claims that removing stereotype threat eliminates achievement gaps have generally disappeared, replaced by conclusions that stereotype threat is merely one piece of the achievement gap puzzle (e.g., Walton et al., 2013).

Effect sizes are no panacea, and are not a gold standard. They are a function of how various phenomena have been studied, operationalized, and measured, which will often always be somewhat arbitrary. Furthermore, because of continued controversies over what constitutes a "large" versus a "small" effect size (contrast, e.g., Greenwald et al. (2015) with Hyde's, 2005 very different characterizations of identical effect sizes), researchers still have considerable latitude in how they frame findings. Tiny effects that we want to promote can still be characterized as "important" whereas identically-sized effects that we dislike can still be dismissed as trivial (see Jussim, Crawford, Stevens, Anglin & Duarte, in press for examples). Nonetheless, requiring reporting of effect sizes and confidence intervals in any discussion of the "power" of some phenomena will likely increase the extent to which psychologists grapple with results that appear to run counter to their preferred narratives.

6.4. Transparency and detail in data presentations

Nearly all empirical articles should present overall means, medians, ranges, standard deviations and correlations. When researchers are interested in reaching conclusions about people at some *level of some variable*, and not merely relationships among variables, they also need to report how many participants are actually at various levels on the scale measuring that variable. This can be done with frequencies and scatterplots. For example, to make claims about people who believe the moon landing was a hoax, characteristics of people who range from pretty sure to certain that the moon landing was *not* a hoax are completely irrelevant. Instead, we need to know whether anyone in the sample actually believed the moon landing was a hoax.

Similarly, full ANOVA tables (complete with old fashioned sums of squares, mean square errors, and degrees of freedom) and full reports of *both* standardized and unstandardized results in regression and SEM models could reduce misleading conclusions that result from, in essence, masked data and results. Similarly, when reporting analyses of covariance, the raw means on the covariate and main dependent variable should always be reported, rather than just reporting the covariate-adjusted means. If we want to reach conclusions about how much some intervention reduced achievement gaps, we need to know: 1) What the pre-intervention gap was; and 2) What the post-intervention gap was. Justified conclusions about gap reductions are left masked by analyses reporting only covariate adjusted means. Even absent increased journal space, such information can often be easily provided as online supplementary materials. Such details will be invaluable components of the scientific record, making it far easier for future researchers to identify exactly what prior research found.

6.5. Publicly posting data

Posting data publicly should become the default for published articles. Occasionally, there might be participant confidentiality issues at stake or other unusual justifications for not posting. Therefore, we are not proposing that posting data be a *rule*, but certainly the *default*. The typical series of 2×2 experimental laboratory studies that appear in our major journals would not usually appear to be the type of data a researcher could exploit in many separate articles, and is not likely to have participant confidentiality issues. Indeed, authors publishing empirical research in APA journals have long been required to sign a document agreeing to make their data available to competent professionals who request it for up to five years post-publication (a requirement that is, apparently, often disregarded without consequence; see Wicherts, Bakker, & Molenaar, 2011).

6.6. Data blinding

To reduce confirmation bias in data analysis and interpretation, MacCoun and Perlmutter (2015) propose adopting the types of methods of blind analysis used in physics. Blind analysis refers to hiding some aspect of the data to the researchers; only after the researchers have completed their analyses do they lift the blind. There is no single technique for performing a blind analysis. The appropriate method depends on the nature of the study and measures. Ideally, researchers will select data blinding procedures that allow them to conduct all necessary analyses (including preliminary analyses such as data cleaning and exclusions).

When researchers have experimental data, they can scramble the cell means so they do not know where observed mean differences lie. For example, in a 2×2 design, there are 24 possible orders of cell means. Researchers could inspect a subset of these different permutations, considering possible explanations for each pattern. Cell scrambling utilizes the "consider-the-opposite strategy" (Lord, Lepper, & Preston, 1984), encouraging researchers to consider alternative conclusions to those they expect. However, a limitation to cell scrambling is that it still preserves the F statistic, informing researchers whether a significant finding is obtained (and thus may not eliminate p-hacking entirely).

Although data blinding is a useful tool for reducing confirmation bias in data analysis and interpretation, there are many questions researchers must consider before implementing data blinding (besides simply how to blind the data; MacCoun & Perlmutter, 2015). For example, who should apply the blinding technique? A member of the research team or a third party? When should the data blind be removed? Is it acceptable for researchers to perform analyses after the blind has been lifted? In addition, it is sometimes necessary to see the actual pattern of results in order to detect an error. Thus, although data blinding can reduce researcher confirmation biases, there is no simple recipe for optimal blinding. Furthermore, blinding is not particularly important for true a priori hypothesis testing, especially pre-registered hypotheses. It ismost relevant for exploratory data analyses, as a means to reduce cherrypicking, researcher degrees of freedom, and p-hacking.

6.7. Viewpoint diversity and identification of alternative explanations and hypotheses

When in doubt, we can seek out colleagues with very different views than our own. We do not have to agree with them or be persuaded by their arguments. But those who *disagree with us* will probably have very different blind spots than we have, and will usually be quite happy to point ours out. Of course, just because a colleague *claims* we

have missed something important does not mean we actually have. The point is to reveal masked findings, studies, and explanations that our own blind spots have led us to miss. Once unmasked, nothing prevents us from critically evaluating them, too — we still can conclude that they are not as important as our critics presume. But at least we will have an opportunity to address them, rather than marching on as if they did not exist.

Ideally, when alternative explanations exist for a phenomenon, researchers will develop methodologies that pit alternative hypotheses against one another. The point is not to demonstrate that one is "true" and the other "false." Indeed, some influential social psychological scholarship has advanced the position that most hypotheses are true under some conditions (Barret, 2015; Greenwald et al., 1986; McGuire, 1983). If one subscribes to this view, it is downright silly to try to "disprove" any theory. Even if one holds this view, our perspective is that it is *still* invaluable to pit alternative perspectives against one another *in particular research contexts*. If everything is true *under some conditions*, then any particular hypothesis is probably not true *under all conditions*. To find out which conditions Hypothesis X accounts for all or most of the data, and under which conditions the alternative, Hypothesis Y does, we need to test both (indeed, there are likely more than two alternative hypotheses that can be brought to bear on most situations studied by social psychologists).

This is why adversarial collaborations have considerable potential to advance the field. No matter how prone we are to confirmation bias, and how difficult it may be to be completely objective in our interpretations, we often have colleagues who are ready, willing, and able to tell us how wrong we are. To address these issues, then, we suggest social psychologists play to one of their strengths. The field has long embraced diversity, in part on the grounds that people from diverse backgrounds bring different experiences to bear on psychological problems. In short, diverse people have diverse ideas, thereby enriching the "marketplace" of ideas.

Such collaborations are probably quite difficult, because those on opposing "sides" of some debate — whether theoretical or political — often hold considerable hostility for one another (e.g., Brandt et al., 2014). Nonetheless, one of the few known solutions to confirmation bias is to adopt an alternative desired conclusion (Kunda, 1990). It may not be easy, but our prediction is that it will usually be worth it.

And what about failed attempts at such collaborations? One criticism often leveled at adversarial collaborations is that they often do not work, because the adversaries are sufficiently hostile to one another, or one another's views, that they cannot work together. This, we would argue, is a testament to just how powerful researcher confirmation biases can be. The lie is put to the ideal image of objective scientists reaching conclusions entirely on the basis of logic, method, statistics, and data by all such failures. Both sides may be equally culpable, or, perhaps, one side is biased and the other is not. Regardless, such failures are a strong signal that something other than the objective and dispassionate pursuit of truth is going on.

At the very least, however, acknowledging other explanations for their findings protects researchers from leaving valid phenomena masked. Clark and Hatfield (1989) offer a good example of how to present multiple possible interpretations of findings. In two separate studies, they had attractive male and female experimenters approach members of the opposite sex on campus and ask one of three questions: "Would you go out with me tonight?" "Would you come over to my apartment tonight?" "Would you go to bed with me tonight?" Both men and women were equally willing to accept the offer for a date, whereas only the males agreed to go to the female experimenter's apartment and go to bed with her. Clark and Hatfield (1989) described their findings as follows: "Men readily accepted a sexual invitation. Women were extremely reluctant to do so. We now know that this is so. We are not quite sure *why* this is so" (p. 51). They then proceeded to describe a sociocultural explanation (the sexual double standard), a sociobiological explanation (based on the principles of sexual selection and the theory of parental investment), and a practical explanation (the situation was riskier for women than men because women are less equipped to protect themselves against physical assault). Because their findings could not rule out one explanation over another, they presented all equally.

Psychology's aspirations have often led it to borrow practices from the natural sciences (such as widespread use of experimentation). In that spirit, a recent article in astronomy (Loeb, 2014) has made important points about diversity of ideas that may be relevant to social psychology. Loeb (2014) highlighted example after example where prestigious astronomers "believed" something to be true on the basis of little or no evidence, obstructed the ability of younger scientists and others with new ideas to make progress on that problem because the alternatives were perceived as outlandish. In each case, many years later, it was ultimately discovered that the "outlandish" claim turned out to be true. In our terminology, unjustified but confidently-held conclusions masked the evidence, and sometimes even the search for evidence, of more valid ones. He concludes (p. 617) his article with this: "Uniformity of opinions is sterile; the co-existence of multiple ideas cultivates competition and progress. Of course, it is difficult to know in advance which exploratory path will bear fruit, and the back yard of astronomy is full of novel ideas that were proven wrong. But to make the discovery process more efficient … funding agencies should dedicate a fixed fraction of their resources (say 10–20%) to risky explorations. This can be regarded as affirmative action to promote a diversity of ideas…." Psychology would do well to adopt similar practices.

7. Conclusion

This review is not intended to provide any information about the frequency or pervasiveness of practices leading to unjustified or distorted conclusions in social psychology. Nonetheless, others have reviewed ample evidence of similar problems in many areas that we have not reviewed, including research on: the size and importance of gender differences (e.g., Eagly, 1995, 2013); the role of researcher politics in political psychology (Tetlock, 1994); research on the (in)accuracy of stereotypes (Jussim, Crawford & Rubinstein, 2015); and the psychological characteristics of liberals and conservatives (Brandt et al., 2014; Duarte et al., 2015).

We generally oppose adding onerous bureaucratic requirements to the already-difficult research process. Nonetheless, some relatively simple institutional changes are already coming down the pike (recent adoption of transparency guidelines that are sweeping many scientific fields, including but not restricted to publicly posting data). Journals, editors and reviewers can require or at least strongly recommend that researchers include far more transparent reports of their data analyses than is currently common.

In addition, individual researchers who wish to improve the validity of their work can take steps to do so by limiting their propensity towards confirmation bias. The 21-word solution (Simmons et al., 2012) was a great start. One complementary method social psychologists can use to limit their potential for motivated reasoning and confirmation biases is to ask themselves a few pointed questions. In that spirit, we conclude our paper with this *Personal Use Checklist*, which we envision as something for personal use, not to be required by journals or other organizations. Furthermore, just because this version of the checklist works for us does not mean it will necessarily work for others. We encourage researchers to adapt this as they see appropriate, or develop their own (see Washburn, Morgan, & Skitka, 2015 for a comparable checklist for checking one's political biases). This is intended to assist well-meaning researchers to become more aware of their own potential for biases, in order to be more able to limit or eliminate them, *without mandates from authorities, editors, or organizations*.

A checklist for increasing confidence that our empirical research is relatively free of motivated biases:

1. What do I want to happen and why?* An honest and explicit self-assessment is a good first step towards recognizing our own tendencies towards bias, and is, therefore, a first step to building in checks and balances in our research to reduce them.
2. Am I shooting for a "Wow Effect!"? Am I painting a weak and inconsistent result as dramatic in order to tell a compelling story? Scientific ambition is not inherently problematic, and may be a powerful constructive force for scientific advancement. But we want our literature to have true, valid, Wow Effects, not ones that cannot be replicated or ones promoted as powerful and pervasive, which upon further reflection (or evidence-gathering) are, in fact, weak, fragile, and fleeting, or which can be easily called into question under critical scrutiny.
3. Do I have a long track record of research that systematically validates a particular political or social narrative or agenda? This is not about one's *intentions* but rather one's *results*. If one's results consistently validate a particular set of beliefs, values or ideology, one has failed this check, and suggests that attempts at falsification may be in order.
4. Am I receiving remuneration (e.g., speaking or consulting fees) for reaching a particular conclusion? Conflicts of interest, though they do not invalidate one's conclusions, plausibly place one at greater risk of dubious research and interpretation practices more generally.
5. Have I generated theoretical arguments for *competing and alternative hypotheses* and designed studies to incorporate and test them?* Honest tests of alternatives can go a long way to reducing personal bias.
6. Have I read some of the literature highlighting the invidious ways our motivated biases, morals, and politics can creep into our scientific scholarship (e.g., Duarte et al., 2015; MacCoun, 1998)? Doing so can alert one to ways in which our preferences might distort our science. After having done so, have I made a good faith attempt to eliminate such biases from my scholarship?
7. Have I sought feedback from colleagues with very different preferences and perspectives than mine or with track records of scholarship that often contest my preferred narratives?

Asterisks (*) indicate items that are from the checklist developed by Washburn et al. (2015).

It may not always be possible for researchers to meet all of these checks. However, as a starting heuristic, meeting six of the seven probably justifies confidence that the research has kept bias mostly in check. What to do if one cannot meet at least six (or, alternatively, one fails too many of one's own such questions?). Although that, too, is a matter of judgment, one possibility will be to *start over*. The first check may appear to be an open-ended question that one can neither pass nor fail. However, if one has strong preferences for how a study "should" come out, then one's ego may be invested in the outcome and one has failed this check. Checks 2 through 5 are easy enough to conduct, though *implementing* 5 after one has realized one has not met it, may require new research. Check 6 requires a little reading and probably can do double duty as a required assignment in advanced undergraduate and graduate courses on methodology, social cognition (confirmation bias among scientists!), and scientific practices. The hardest part about Check 7 is finding enough people so that the ones from whom one seeks feedback are not overburdened.

References

Abramowitz, S. I., Gomes, B., & Abramowitz, C. V. (1975). Publish or perish: Referee bias in manuscript review. *Journal of Applied Social Psychology*, *5*, 187–200.

Allport, F. H. (1955). *Theories of perception and the concept of structure.* NY: Wiley.

Allport, G. W. (1954/1979). *The nature of prejudice.* Reading, MA: Addison-Wesley.

American Psychological Association (2006). Stereotype threat widens achievement gap. Retrieved on 8/20/15 from http://www.apa.org/research/action/stereotype.aspx

Appel, M., & Kronberger, N. (2012). Stereotypes and the achievement gap: Stereotype threat prior to test taking. *Educational Psychology Review*, *24*, 609–635.

Aronson, J., Fried, C. B., & Good, C. (2002). Reducing the effects of stereotype threat on African American college students by shaping theories of intelligence. *Journal of Experimental Social Psychology*, *38*, 113–125.

Aronson, J., Lustina, M. J., Good, C., Keough, K., Steele, C. M., & Brown, J. (1999). When white men can't do math: Necessary and sufficient factors in stereotype threat. *Journal of Experimental Social Psychology*, *35*, 29–46.

Ashton, M. C., & Esses, V. M. (1999). Stereotype accuracy: Estimating the academic performance of ethnic groups. *Personality and Social Psychology Bulletin*, *25*, 225–236.

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

Banerjee, P., Chatterjee, P., & Sinha, J. (2012). Is it light or dark? Recalling moral behavior changes perception of brightness. *Psychological Science*, *23*, 407–409.

Bargh, J. A. (2007). Social psychological approaches to consciousness. In P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *The Cambridge handbook of consciousness* (pp. 555–569). New York, NY: Cambridge University Press.

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, *71*, 239–244.

Baron, R. M., Albright, L., & Malloy, T. E. (1995). Effects of behavioral and social class information on social judgment. *Personality and Social Psychology Bulletin*, *21*, 308–315.

Barret, L. F. (2015, Sept 1). Psychology is not in crisis. *New York Times* (Retrieved on 10/1/15 from: http://www.nytimes.com/2015/09/01/opinion/psychology-is-not-in-crisis.html?_r=0).

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

Bem, D. (2002). Writing the empirical journal article. In J. M. Darley, M. P. Zanna, & H. L. RoedigerIII (Eds.), *The complete academic: A career guide*. Washington, DC: American Psychological Association.

Bickel, P. J., Hammel, E. A., & O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, *187*, 396–404.

Blanton, H., Jaccard, J., Klick, J., Mellers, B., Mitchell, G., & Tetlock, P. E. (2009). Strong claims and weak evidence: Reassessing the predictive validity of the IAT. *Journal of Applied Psychology*, *94*, 567.

Brandt, M. J., Reyna, C., Chambers, J. R., Crawford, J. T., & Wetherell, G. (2014). The ideological-conflict hypothesis: Intolerance among both liberals and conservatives. *Current Directions in Psychological Science*, *23*, 27–34.

Brown, R. (2011). *Prejudice: Its social psychology.* John Wiley & Sons.

Bruner, J. (1957). On perceptual readiness. *Psychological Review*, *64*, 123–152.

Caprara, G. V., Alesandri, G., & Eisenberg, N. (2012). Prosociality: The contribution of traits, values, and self-efficacy beliefs. *Journal of Personality and Social Psychology*, *102*, 1289–1303.

Clark, R. D., & Hatfield, E. (1989). Gender differences in receptivity to sexual offers. *Journal of Psychology & Human Sexuality*, *2*, 39–55.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* New York: Academic Press.

Darley, J. M., & Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology*, *44*, 20–33.

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*, 569–584.

Doyen, S., Klein, O., Pichon, C., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS One*, *7*, e29081.

Duarte, J. L., Crawford, J. T., Stern, C., Haidt, J., Jussim, L., & Tetlock, P. E. (2015). Political diversity will improve social psychological science. *Behavioral and Brain Sciences*, *38*, 1–54.

Eagly, A. H. (1995). The science and politics of comparing women and men. *American Psychologist*, *50*, 145–158.

Eagly, A. H. (2013). *Sex differences in social behavior: A social-role interpretation.* Psychology Press.

Epstein, W. M. (2004). Informational response bias and the quality of the editorial processes among American social work journals. *Research on Social Work Practice*, *14*, 450–458.

Firestone, C., & Scholl, B. J. (2014). "Top-down" effects where none should be found: The El Greco fallacy in perception research. *Psychological Science*, *25*, 38–46.

Firestone, C., & Scholl, B. J. (2015). Enhanced visual awareness for morality and pajamas? Perception vs. memory in top-down effects. *Cognition*, *136*, 409–416.

Firestone, C., & Scholl, B. J. (2015). Cognition does not affect perception: Evaluating the evidence for 'top-down' effects. *The Behavioral & Brain Sciences*, 1–77. http://dx.doi.org/10.1017/S014052X15000965 (in press).

Fleeson, W. (2001). Towards a structure- and process-integrated view of personality: Traits as density distributions of states. *Journal of Personality and Social Psychology*, *37*, 1097–1126.

Fleeson, W. (2004). Moving personality beyond the person–situation debate. *Current Directions in Psychological Science*, *13*, 83–87.

Funder, D. C. (2008). Persons, situations, and person–situation interactions. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (pp. 568–580) (3rd edition). New York: Guilford Press.

Funder, D. C. (2012). The perilous plight of the (non)-replicator. Retrieved on 9/13/15 from: http://funderstorms.wordpress.com/2012/10/31/the-perilous-plight-of-the-non-replicator/

Funder, D. C., Levine, J. M., Mackie, D. M., Morf, C. C., Vazire, S., & West, S. (2014). Improving the dependability of research in personality and social psychology. *Personality and Social Psychology Review*, *18*, 3–12.

Gantman, A. P., & Van Bavel, J. J. (2014). The moral pop-out effect: Enhanced perceptual awareness of morally relevant stimuli. *Cognition*, *132*, 22–29.

Gawronski, B. (2004). Theory-based bias correction in dispositional inference: The fundamental attribution error is dead, long live the correspondence bias. *European Review of Social Psychology*, *15*, 183–217.

Gelman, A. (2015). To understand the replication crisis, imagine a world in which everything was published. Retrieved on 9/11/15 from: http://andrewgelman.com/2015/

09/02/to-understand-the-replication-crisis-imagine-a-world-in-which-everything-was-published/
Gildersleeve, K., Haselton, M. G., & Fales, M. R. (2014). Meta-analyses and p-curves support robust cycle shifts in women's mate preferences: Reply to Wood and Carden (2014) and Harris, Pashler, and Mickes (2014). *Psychological Bulletin*, *140*, 1272–1280. http://dx.doi.org/10.1037/a0037714.
Gonzales, P. M., Blanton, H., & Williams, K. J. (2002). The effect of stereotype threat and double-minority status on the test performance of Latino women. *Personality and Social Psychology Bulletin*, *28*, 659–670.
Greenwald, A. G., & Pettigrew, T. F. (2014). With malice toward none and charity for some: Ingroup favoritism enables discrimination. *American Psychologist*, *69*, 669–684.
Greenwald, A. G., Banaji, M. R., & Nosek, B. A. (2015). Statistically small effects of the Implicit Association Test can have societally large effects. *Journal of Personality and Social Psychology*, *108*, 553–561.
Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, *97*, 17.
Greenwald, A. G., Pratkanis, A. R., Leippe, M. R., & Baumgardner, M. H. (1986). Under what conditions does theory obstruct research progress? *Psychological Review*, *93*, 216–229.
Guimond, S., Crisp, R. J., Oliveria, P. D., Kamiejski, R., Kteily, N. S., Kuepper, B., ... Zick, A. (2013). Diversity policy, social dominance, and intergroup relations: Predicting prejudice in changing social and political contexts. *Journal of Personality and Social Psychology*, *104*, 941–958.
Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*, 814–834.
Harackiewicz, J. M., Canning, E. A., Tibbetts, Y., Giffen, C. J., Blair, S. S., Rouse, D. I., & Hyde, J. S. (2014). Closing the social class achievement gap for first-generation students in undergraduate biology. *Journal of Educational Psychology*, *106*, 375–389.
Hastorf, A. H., & Cantril, H. (1954). They saw a game: A case study. *Journal of Abnormal and Social Psychology*, *49*, 129–134.
Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist*, *60*, 581–592.
Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, e124.
Ioannidis, J. P. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science*, *7*, 645–654.
Jacks, J. Z., & Devine, P. G. (2000). Attitude importance, forewarning of message content, and resistance to persuasion. *Basic and Applied Social Psychology*, *22*, 19–29.
Jordan, C. H., & Zanna, M. P. (2007). Not all experiments are created equal: On conducting and reporting persuasive experiments. In R. J. Sternberg, D. Halpern, & H. L. RoedigerIII (Eds.), *Critical thinking in psychology* (pp. 160–176). New York: Cambridge University Press.
Jost, J. T., & Kruglanski, A. W. (2002). The estrangement of social constructionism and experimental social psychology: History of the rift and prospects for reconciliation. *Personality and Social Psychology Review*, *6*, 168–187.
Jost, J. T., Glaser, J., Kruglanski, A. W., & Sulloway, F. J. (2003). Political conservatism as motivated social cognition. *Psychological Bulletin*, *129*, 339–375.
Jussim, L. (1991). Social perception and social reality: A reflection-construction model. *Psychological Review*, *98*, 54–73.
Jussim, L. (2012a). *Social perception and social reality: Why accuracy dominates bias and self-fulfilling prophecy*. New York: Oxford University Press.
Jussim, L. (2012b). Unicorns of social psychology. Retrieved on 9/13/15 from: https://www.psychologytoday.com/blog/rabble-rouse/201208/unicorns-social-psychology
Jussim, L., & Maoz, I. (2014). Desperately seeking the "wow effect": Data interpretation and scientific story-telling as issues of research integrity. *Presentation at the October 2014 meeting of the Society for Experimental Social Psychology*.
Jussim, L., Crawford, J. T., & Rubinstein, R. S. (2015). Stereotype (in)accuracy in perceptions of groups and individuals. *Current Directions in Psychological Science*, *24*, 490–497.
Jussim, L., Crawford, J. T., Stevens, S. T., Anglin, S. M., & Duarte, J. L. (2016). Can high moral purposes undermine scientific integrity? To appear in In J. Forgas, L. Jussim, & P. van Lange (Eds.), *Sydney symposium on the social psychology of morality*. New York, NY: Taylor and Francis (in press).
Kahan, D. M. (2011). Neutral principles, motivated cognition, and some problems for constitutional law. *Harvard Law Review*, *125*, 1.
Kellow, J. T., & Jones, B. D. (2008). The effects of stereotypes on the achievement gap: Reexamining the academic performance of African American high school students. *Journal of Black Psychology*, *34*, 94–120.
Krosnick, J. A. (2015). Issues in scientific integrity. *Presentation at the annual conference of the Society for Personality and Social Psychology, Long Beach, CA*.
Krueger, J. I., & Funder, D. C. (2004). Towards a balanced social psychology: Causes, consequences, and cures for the problem-seeking approach to social behavior and cognition. *Behavioral and Brain Sciences*, *27*, 313–376.
Krueger, J. I. (2009). A componential model of situation effects, person effects, and situation-by-person interaction effects on social behavior. *Journal of Research in Personality*, *43*, 127–136.
Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*, 480–498.
Lane, A., Mikolajczak, M., Treinen, E., Samson, D., Corneille, O., de Timary, P., & Luminet, O. (2015). Failed replication of oxytocin effects on trust: The envelope T ask case. *PLoS One*, *10*(9), e0137000. http://dx.doi.org/10.1371/journal.pone.0137000.
Ledgerwood, A., Haines, E., & Ratliff, K. (2015). Not nutting up or shutting up: Notes on the demographic disconnect in our field's best practices conversation. Retrieved 9/22/2015 from http://sometimesimwrong.typepad.com/wrong/2015/03/guest-post-not-nutting-up-or-shutting-up.html
Levine, R., Chein, I., & Murphy, G. (1942). The relation of the intensity of a need to the amount of perceptual distortion: A preliminary report. *The Journal of Psychology*, *13*, 283–293.
Lewandowsky, S., Oberauer, K., & Gignac, G. E. (2013). NASA faked the moon landing—Therefore, (climate) science is a hoax an anatomy of the motivated rejection of science. *Psychological Science*, *24*, 622–633.
Lilienfeld, S. O. (2010). Can psychology become a science? *Personality and Individual Differences*, *49*, 281–288.
Loeb, A. (2014). Benefits of diversity. *Nature: Physics*, *10*, 616–617.
Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology*, *47*, 1231–1243.
Lord, C. G., Ross, L., & Lepper, M. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*, 2098–2109.
MacCoun, R. J. (1998). Biases in the interpretation and use of research results. *Annual Review of Psychology*, *49*, 259–287.
MacCoun, R. J., & Perlmutter, S. (2015). Blind analysis as a correction for confirmatory bias in physics and psychology. To appear in In S. O. Lilienfeld, & I. Waldman (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions*. John Wiley and Sons (in press).
Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research*, *1*, 161–175.
Malle, B. F. (2006). The actor–observer asymmetry in attribution: A (surprising) meta-analysis. *Psychological Bulletin*, *132*, 895–919.
McCauley, C., Stitt, C. L., & Segal, M. (1980). Stereotyping: From prejudice to prediction. *Psychological Bulletin*, *87*, 195–208.
McConnell, A. R., & Leibold, J. M. (2001). Relations among the Implicit Association Test, discriminatory behavior, and explicit measures of racial attitudes. *Journal of Experimental Social Psychology*, *37*, 435–442.
McGinnies, E. (1949). Emotionality and perceptual defense. *Psychological Review*, *56*, 244–251.
McGuire, W. J. (1983). A contextualist theory of knowledge: Its implications for innovation and reform in psychological research. *Advances in Experimental Social Psychology*, *16*, 1–47.
Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, *1*, 108–141.
Mikolajczak, M., Gross, J. J., Lane, A., Corneille, O., de Timary, P., & Luminet, O. (2010). Oxytocin makes people trusting, not gullible. *Psychological Science*, *21*(8), 1072–1074.
Milgram, S. (1974). *Obedience to authority: An experimental view*. New York, NY: Harper & Row.
Miller, G. A., & Chapman, J. P. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology*, *110*, 40–48.
Munro, G. D., & Ditto, P. H. (1997). Biased assimilation, attitude polarization, and affect in reactions to stereotype-relevant scientific information. *Personality and Social Psychology Bulletin*, *23*, 636–653.
Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*, 175–220.
Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716. http://dx.doi.org/10.1126/science.aac4716.
Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology*, *105*, 171–192.
Popper, K. (1934). *The logic of scientific discovery*. New York: Routledge.
Regner, I., Huguet, P., & Monteil, J. M. (2002). Effects of socioeconomic status (SES) information on cognitive ability inferences: When low-SES students make use of a self-threatening stereotype. *Social Psychology of Education*, *5*, 253–269.
Rosenthal, R., & Fode, K. L. (1963). The effect of experimenter bias on the performance of the albino rat. *Behavioral Science*, *83*, 183–189.
Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. *Advances in Experimental Social Psychology*, *10*, 173–220.
Ross, L., & Nisbett, R. E. (1991). *The person and the situation: Perspectives of social psychology*. New York: McGraw-Hill.
Ross, L., Lepper, M., & Ward, A. (2010). History of social psychology: Insights, challenges, and contributions to theory and application. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (pp. 3–50). New York: McGraw-Hill.
Sackett, P. R., Hardison, C. M., & Cullen, M. J. (2004). On interpreting stereotype threat as accounting for African American–White differences on cognitive tests. *American Psychologist*, *59*, 7–13.
Sampson, R. J., Winship, C., & Knight, C. (2013). Translating causal claims: Principles and strategies for policy-relevant criminology. *Criminology & Public Policy*, *12*, 587–616.
Schmader, T., Johns, M., & Forbes, C. (2008). An integrated process model of stereotype threat effects on performance. *Psychological Review*, *115*, 336–356.
Schneider, D. J., Hastorf, A. H., & Ellsworth, P. (1979). *Person perception*. Reading, MA: Addison-Wesley Pub. Co.
Sedikedes, C., & Skowronski, J. J. (1991). The law of cognitive structure activation. *Psychological Inquiry*, *2*, 169–184.
Shen, H. (2013). Mind the gender gap. *Nature*, *495*, 22–24.
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.
Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 word solution. Retrieved from http://ssrn.com/abstract=2160588
Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B*, *13*, 238-214.
Simpson's Paradox (n.d.) In Wikipedia. Retrieved 9/16/15 from https://en.wikipedia.org/wiki/Simpson's_paradox
Snyder, M., & Swann, W. B. (1978). Behavioral confirmation in social interaction: From social perception to social reality. *Journal of Experimental Social Psychology*, *14*, 148–162.

Stanovich, K. E., West, R. F., & Toplak, M. E. (2013). Myside bias, rational thinking, and intelligence. *Current Directions in Psychological Science*, *22*, 259–264.
Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, *69*, 797–811.
Steele, C. M., & Aronson, J. (2004). Stereotype threat does not live by Steele and Aronson (1995) alone. *American Psychologist*, *59*, 47–55.
Tetlock, P. E. (1994). Political psychology or politicized psychology: is the road to scientific hell paved with good moral intentions? *Political Psychology*, *15*, 509–529.
Trope, Y., & Bassok, M. (1983). Information-gathering strategies in hypothesis-testing. *Journal of Experimental Social Psychology*, *19*, 560–576.
Van Hiel, A., Onraet, E., & De Pauw, S. D. (2010). The relationship between social-cultural attitudes and behavioral measures of cognitive style: A meta-analytic integration of studies. *Journal of Personality*, *78*, 1765–1800.
Walton, G. M., & Spencer, S. J. (2009). Latent ability grades and test scores systematically underestimate the intellectual ability of negatively stereotyped students. *Psychological Science*, *20*, 1132–1139.
Walton, G. M., Spencer, S. J., & Erman, S. (2013). Affirmative meritocracy. *Social Issues and Policy Review*, *7*, 1–35.
Washburn, A. N., Morgan, G. S., & Skitka, L. J. (2015). A checklist to facilitate objective hypothesis testing. *Behavioral and Brain Sciences*, *38*, 42–43.
West, T. V., & Kenny, D. A. (2011). The truth and bias model of judgment. *Psychological Review*, *118*, 357–378.
Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLoS One*, *6*, e26828. http://dx.doi.org/10.1371/journal.pone.0026828.
Wolfe, C., & Spencer, S. (1996). Stereotypes and prejudice. *American Behavioral Scientist*, *40*, 176–185.
Yzerbyt, V. Y., Muller, D., & Judd, C. M. (2004). Adjusting researchers' approach to adjustment: On the use of covariates when testing interactions. *Journal of Experimental Social Psychology*, *40*, 424–431.
Zuwerink, J. R., & Devine, P. G. (1996). Attitude importance and resistance to persuasion: It's not not just the thought that counts. *Journal of Personality and Social Psychology*, *70*, 931–944.