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%).