Journal of Mathematical Psychology 57 (2013) 190–195

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

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

Commentary

We should focus on the biases that matter: A reply to commentaries

Gregory Francis

Department of Psychological Sciences, Purdue University, 703 Third Street, West Lafayette, IN 47907-2004, United States

HIGHLIGHTS

- The commentaries provided valuable perspectives on the bias analysis.
- Most of the criticisms reflect misunderstandings about types of bias.
- These types of analyses should focus on the biases that matter to scientists.

ARTICLE INFO

Article history:
Available online 29 July 2013

The commentaries on the target article (Francis, 2013) discussed challenging ideas and new ways of characterizing the issues around bias and the consistency test. I have organized my reply by author with the more negative commentaries being addressed first. I do not try to address every point in every commentary, especially if I feel the point has been addressed elsewhere. Instead I have focused on what I judged were the most important issues raised in each commentary.

1. Morey

Morey (2013) presents three arguments against the consistency test. I counter that these arguments often do not focus on the types of bias that are important for science.

1.1. Bias as a process or as an outcome

Morey argued that the intention of the consistency analysis is improper because bias is an aspect of a process rather than a state of the data. This is an interesting observation about bias, but I think it only confuses the discussion. Bias in a statistical sense is related to systematic misestimation of a value. Research psychologists appear to be especially interested in bias related to two values: how often experiments reject the null hypothesis (replicability) and the magnitude of an effect size. Some scientific processes lead to biased measures of these values. For example, a file drawer (where nonsignificant findings are suppressed) can overestimate both the replication rate and the effect size. Scientific investigations that use optional stopping (stop gathering data when statistical significance has been found) can dramatically overestimate replicability but do not have much bias for the effect size (Francis, 2012c). Morey's description of bias suggests that experimental results can be biased even if they unbiasedly estimate the variables of interest.

Morey's description of bias does not seem like a useful one for a practicing scientist, and it is for this reason that the target article focused on consistency rather than bias. Consistency is a property of the data and experiment sets, and I think it is justifiable to ask for experiment sets to be consistent, relative to some criterion. Unbiased (in the statistical sense) experiment sets are almost always consistent, and biased experiment sets are sometimes inconsistent (depending on the process that produces the statistical bias). The consistency test sets a modest standard for experiment sets and detects some instances of bias.

1.2. Concluding bias when it does not exist

Morey points out that experiments are often planned in a sequential method with previous results influencing the properties (and existence) of additional experiments. He claims that this approach will often trigger the consistency test even when there is no bias. Morey describes a quit-after-nonsignificant-result (QANSR) process where a researcher runs multiple experiments, stops with the first nonsignificant experiment, and publishes all the findings. As he notes, "If the true power is known to be 0.4 or less, then examining experiment sets of 5 or greater will *always* lead to a significant result, even when there is no publication bias". The final part of the statement is incorrect.

We have to talk about bias relative to the measures scientists care about: replicability and effect size. By definition, a set of five or more low power experiments generated under the QANSR process presents a biased representation of replicability. If a scientist practices QANSR but does not inform readers about that strategy, then readers have a false sense about the replicability

E-mail address: gfrancis@purdue.edu.

0022-2496/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmp.2013.06.001

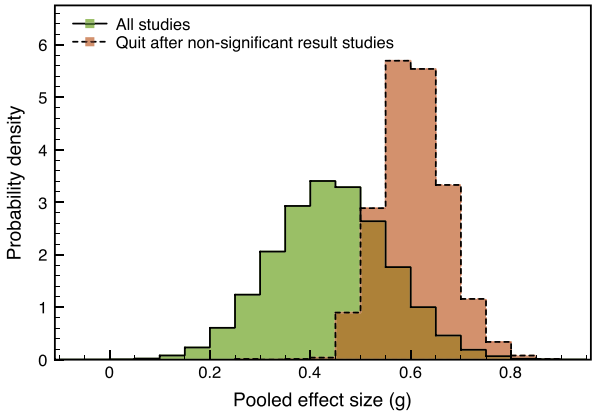

Fig. 1. Distributions of pooled effect sizes for 100,000 simulated sets of five experiments. The solid line distribution is for all experiment sets. The dashed line distribution is for the subset of experiment sets that satisfy the quit-after-nonsignificant-result (QANSR) process described by Morey. The QANSR based distribution is biased relative to the true effect size of $\delta = 0.448$.

of the experimental findings. As long as the scientist is up front about the process, then perhaps there is little harm to such a misrepresentation. However, even though all of the investigations are reported, the QANSR process also introduces a bias for the effect size. To demonstrate this bias consider a population where the true standardized effect size for a difference of means is $\delta = 0.448$. Suppose a researcher runs five experiments with control and experimental groups having $n_1 = n_2 = 30$ and runs a two-sample, two-tailed, t-test. For such tests the true power is 0.4.

Each set of five experiments can produce a pooled effect size, and the solid line curve in Fig. 1 shows the probability density function of the pooled effect size. The function is estimated from 100,000 simulated experiment sets. As expected, this distribution is roughly centered on the true effect size. The dashed curve in Fig. 1 describes the distribution of pooled effect sizes that is estimated from only those experiment sets that satisfy the QANSR process. That is, the first four experiments were statistically significant and the fifth experiment was not significant. Since the power of each experiment is low, such experiment sets are quite rare; only 1538 sets met the QANSR requirement. Most of these sets dramatically overestimate the pooled effect size (the mean is $\overline{g}^* = 0.607$). Such an outcome is expected because when the true power is 0.4, the only way the first four experiments can produce significant results is when the (randomly chosen) samples dramatically overestimate the true effect size. Thus, the QANSR approach described by Morey produces a biased effect size, so it is appropriate that the consistency test indicates bias. (Note, this analysis supposes that we know that the true power is 0.4, if we estimated the power from the reported effect sizes we might not be able to detect the bias.)

One could consider other types of sequential experiment planning schemes, but my intuition is that they will behave in a similar way. Some schemes will properly estimate the true effect size, and such experiment sets are unlikely to trigger the consistency test. Other schemes will be biased and sometimes trigger the consistency test.

1.3. Evidence

Morey's final criticism is that the consistency test does not provide a proper type of evidence for bias. I concede that some of my previous reports used the term "evidence" in a non-precise way. I also concede that there is some incongruity between my call for experimentalists to use Bayesian methods while simultaneously using frequentist logic for the consistency test. In general, I feel that Morey raises a fair point, and I am grateful for the feedback.

Table 1
A hypothetical experiment set that appears to be biased, but where Experiment 1 may have data worth saving.

Exp.	$n_1 = n_2$	t	p	Hedges's g	Power
1	100	3.00	0.003	0.42	0.94
2	20	2.05	0.05	0.64	0.34
3	25	2.07	0.04	0.58	0.41
4	18	2.05	0.05	0.67	0.31
5	30	2.11	0.04	0.54	0.48
6	27	2.12	0.04	0.57	0.44

My thoughts about how to make scientific arguments with statistics are evolving, and I am not sure that there is a single approach that works in every situation. I have asked several Bayesian experts to help develop a Bayesian version of the consistency test, but they have not reported success. I am not convinced that a Bayesian approach is impossible, but it apparently is not straightforward to apply Bayesian principles to this situation.

In general, I agree with Morey's criticisms about p values being misinterpreted, but I think that properly interpreted p values can provide information that helps to promote a scientific argument. The simulations in the target article show that the consistency test is very conservative, so we may be operating in situations where default Bayesian and frequentist approaches provide essentially equivalent analyses.

Despite our differences, Morey and I agree that what is really needed are changes in scientific practice to reduce publication bias. I see the consistency test as a means of profiling the issues about bias and motivating people toward better practice. I will be delighted if scientific practice improves so much that the consistency test becomes useless.

2. Simonsohn

I was disappointed to see that Simonsohn's (2013) comment is essentially a repetition of the arguments presented in Simonsohn (2012). I feel that the target article addressed those concerns, so I will not repeat the same counterarguments. It may be that my counterarguments have not convinced Simonsohn because he believes that the consistency test investigates a quite different topic than what it actually explores. His criticisms of the consistency test are generally valid relative to the topic he thinks it explores, but invalid relative to how the test has actually been used. Before discussing these differences, the next section considers a topic where we are not so far apart.

2.1. What to do with seemingly biased data?

Must we ignore data that appears to be biased? My answer has often been "yes" because the burden of proof is on the original authors to make a strong case, and it is difficult to make a strong statistical argument with apparently biased data sets. Simonsohn argues that such an attitude is imprudent because the biased data may still have evidential value. As I explained in the target article, I am not opposed to efforts to salvage findings from biased experiment sets, but such approaches need to be justified.

I can think of an approach that may be useful in some situations. Suppose there are six experiments (two-sample, two-tailed t-tests) that investigate the same effect. Table 1 summarizes the (entirely made up) statistics for this set of experiments. Every experiment rejects the null hypothesis, but Experiment 1 does so handily, while Experiments 2–6 just barely meet the typical criterion for statistical significance. When applying the consistency test, the pooled effect size comes out as $g^* = 0.5$ and the final column of Table 1 shows the power of each experiment to reject the null for such a pooled effect size. Although the power is very large for Experiment 1 (due to its large sample size), power is