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

Journal of Mathematical Psychology
in press

Gregory Francis[1]

Department of Psychological Sciences

Purdue University

703 Third Street

West Lafayette, IN 47907-2004

May 25, 2013

Key words: Hypothesis testing, statistics, publication bias, scientific publishing

Running head: Focus on the biases that matter

[1]E-mail: gfrancis@purdue.edu; phone: 765-494-6934. I thank each of the commentary authors for sharing their thoughts about the target article. I also thank E. J. Wagenmakers for inviting the target article and organizing the commentaries.

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 non-significant 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 suggest 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 .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 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 .4.

Each set of five experiments can produce a pooled effect size, and the solid line curve in Figure 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 Figure 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 the true power is .4, if we estimated the power from the reported

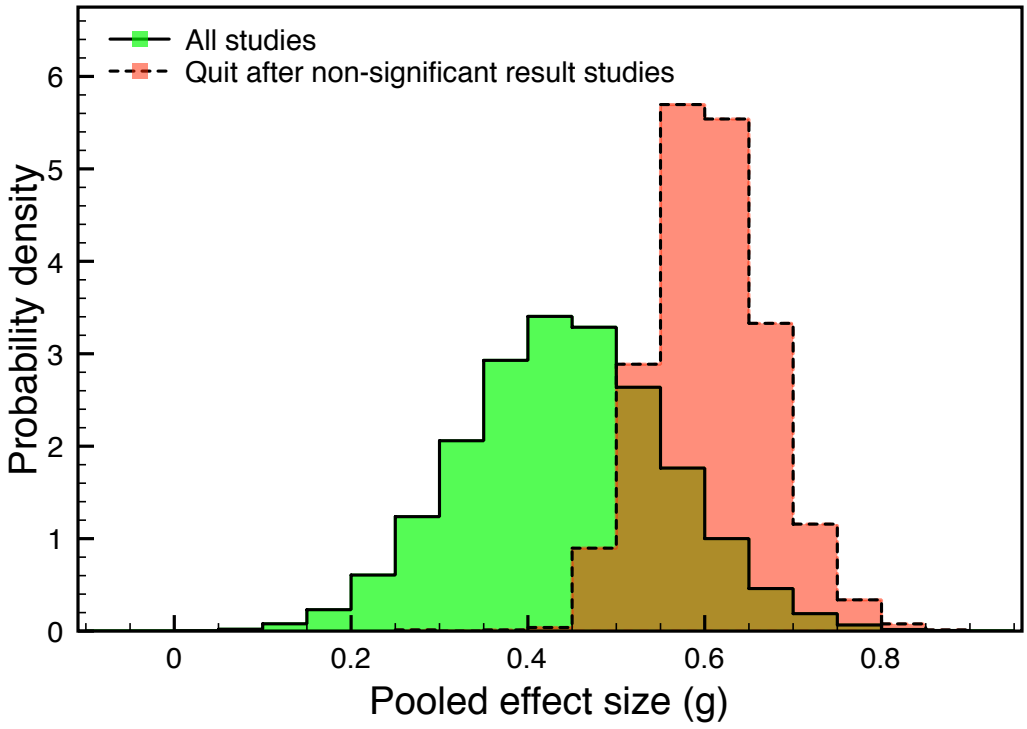

Figure 1: Distributions of pooled effect sizes for 100,000 simulated sets of five experiments. The solid line distribution is for all experiment sets. The dashed line distribution is for the subset of experiment sets that satisfy the quit-after-nonsignificant-result (QANSR) process described by Morey.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

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 provides 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.

My thoughts about how to make scientific arguments with statistics is 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

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

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

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 quite low for the other five experiments. Indeed, the consistency test computes P_c=0.008 and concludes that the experiment set appears to be biased.

As I noted in the target article, the appearance of bias does not necessarily mean that every experiment in the set is invalid. In this particular case, a scientist might decide that Experiment 1 is sound because its large sample size provides a good investigation into the phenomena. In other cases, such as where all the experiments have similar power values, it is difficult to see how anything could be salvaged from a biased experiment set (and this is the case for many of the analyzed experiment sets I have reported). There will surely be in-between cases where it is not clear whether there is data worth salvaging. When in doubt, my recommendation is to be cautious and gather new data, but reasonable scientists can have different interpretations at the margins.

The situation is more complicated when experiment effect sizes measure different phenomena

and should not be pooled. For example, in the Topolinsky and Sparenberg (2012) experiments that were analyzed in the target article, I am generally inclined to believe, at least from a statistical perspective, that the mere exposure finding reported in Experiment 1 (p=0.001, power=0.922) is valid. However, I remain skeptical about the theoretical conclusion based on this experiment and the other reported findings. Each of the other four experimental findings has an estimated power value less than 0.6, so the probability of all the experiments rejecting the null is quite low. The belief in results from an experiment on one topic does not transfer to the entire set; and uncertainty does not average. A scientific argument based on a perfect pattern of statistical significance that combines information from investigations of different effects accumulates uncertainty; so a good argument requires a high confidence in each individual result.

It is worth pointing out that the magnitude of bias, which Simonsohn correctly notes the consistency analysis does not provide, is largely irrelevant for salvaging data from an apparently biased set. What matters is the quality of an experiment by itself, not whether it is part of a set of five or five hundred other, low-power, published experiments (we never know about unpublished experiments). These basic ideas for salvaging data seem to be different from a method for identifying evidential value using p-curves (Simonsohn, Nelson & Simmons, in press). Using the web site described in Simonsohn *et al.* (in press), the p-curve analysis concludes that the experiments in Table 1 were intensely p-hacked (biased) but that there is no evidential value in the set. It draws the same conclusion (the numbers for the analysis are unchanged) if the sample size for Experiment 1 is increased to 5000 in each group. In terms of identifying bias, the p-curve analysis seems to share some similarities with the consistency test, which makes our apparent disagreement rather odd and unfortunate.

2.2 Not all biases mean the same thing

Many of Simonsohn's concerns reflect a misunderstanding about different types of bias. Simonsohn seems to believe that all biases are the same and that essentially all experiment sets are biased (and thus that we learn nothing from the consistency test). The reality is much more nuanced. Some forms of bias only impose restrictions on what kinds of inferences can be drawn from an analysis.

For example, a scientist that investigates many different topics with independent experiments might report only those results that are statistically significant. One could apply the consistency analysis to the experiments with a theoretical link being that they were all produced by the same

scientist. If the power values were relatively low, the analysis might very well conclude that the findings are inconsistent and thereby appear to be biased. Indeed, we should be skeptical that a scientist could always produce significant results with low powered studies. However, I would argue that this is a bias that no one cares about because scientists do not claim that their reported findings are representative of their skills to produce significant results. As noted earlier in this reply, scientists fundamentally care about biases that misrepresent effect sizes and replication rates for those effects. The former concern does not apply here because there is no common effect size across the independent experiments, and each individual experiment should provide an unbiased estimate of its specific effect. The latter concern does not apply here because people who inferred the scientist's reported success rate as being representative of his scientific skills misunderstood the scientific process. Good scientists generally make lots of mistakes, but they catch the mistakes and improve their methods until they have solid findings. The final solid findings are not undermined by the earlier mistakes. More generally, empirical scientific reports are about characterizing effects, not about describing the abilities of scientists. A bias that influences the former is a serious problem, but a bias that influences the latter is mostly irrelevant.

The same kind of reasoning applies to bias in my published reports of publication bias. Simonsohn notes that there is a connection across the reports, but this connection only promotes an improper or meaningless inference. He states, "The critiques by Francis...are by the same author, published the same year, conducting the same statistical test, to examine the exact same question (do social psychologists report all their failed studies?)." The key confusion is the part in parentheses. If that question was really the intent of my investigations into publication bias, then Simonsohn's concerns about cherry picking, controlling the rate of false-positives, and selective reporting would be valid. To the contrary, the bias investigations, as I have used them, are explicitly not examining the question Simonsohn proposes. That question is interesting and important, but the selective nature of my analyses prohibit me from answering it. However, such a prohibition does not prevent me from answering (likewise interesting and important) questions about specific experiment sets. Personally, I think the questions about specific experiment sets are more important than the broad question about whether some scientists produce biased data. The specific question is more important because I care about scientifically measured effects more than I care about the behavior of scientists.

Just to finish out this thought, Simonsohn notes that authors whose work I have critiqued could

take a similar attitude and claim that although there is bias in the set, each individual experiment stands on its own. The statement is correct, but it would require a notable shift in an author's interpretation of their data. Topolinsky and Sparenberg (2012), for example, would need to give up the claim that their experiments supported their theoretical conclusion about clockwise rotation and novelty. Without such a conclusion I am not sure the experiments are worth publishing, and readers might wonder why four unrelated experiments were published together. In contrast, the motivation for Francis (2012a,b) to report on multiple apparently biased experiments sets was not because those sets suggested a common theory, instead the motivation was to demonstrate a statistical method that offered insight about interesting experiment sets.

2.3 What do we learn?

Simonsohn finishes his comment by pondering what we might be learning from the critiques I have published. For the most part I think his comments are simply irrelevant for the analyses I have performed. Still, the question is a good one, so let me give my own answer.

First, the analyses provided an explanation for how some very surprising results might have been published. Like many scientists I was initially baffled how Bem (2011) could report seemingly convincing evidence for precognition. At first glance, the experimental findings seemed of equal quality to other findings in psychology. It was a relief to discover (Francis, 2012a) that the consistency test suggested that the set of precognition experiments appears to be biased (other people came to the same conclusion through a scrutiny of Bem's methods).

Second, my relief turned to dismay when I realized that other specific experiment sets also appeared to be biased, even though the reported findings seemed much more plausible than precognition. This discovery indicates that in some important cases scientists are not making good arguments for the validity of their empirical findings and/or their theoretical positions. Contrary to what has been widely believed, repeated successful replication is not always a good way to make a scientific argument. Because of statistical uncertainty, the rate of replication success needs to match the experimental power. Although obvious in hindsight, this relationship was a surprise to me, and it has seemingly been a surprise to other scientists as well.

Third, by identifying bias within a specific experiment set, the consistency test draws scrutiny about both that effect and other effects that depend on the apparently biased finding. Scientists care about individual effects and specific theories because they provide a foundation for future work

and recommendations that can improve people's lives. Identifying apparently biased experiment sets saves people time and effort that might otherwise have been spent chasing exaggerated effects.

Fourth, I am pretty sure that the published analyses were a surprise to the authors whose work I have critiqued. I believe these scientists wanted to produce unbiased data and that they operated with the best of intentions. I hope that these authors will recover from the sting of being criticized.

Fifth, until a couple of years ago, I was largely unaware how power, replication, effect sizes, meta-analysis, and Bayesian methods were related to general scientific practice. I now understand that many people have been discussing these issues for quite some time, but the broad field seems not to have listened to their concerns. I think the consistency test provides a means to make the point very explicit by showing how ignoring these issues undermines the quality of specific scientific investigations.

Overall, I think we have learned a lot from the consistency test analyses.

3 Johnson

Johnson (2013) expresses skepticism about the consistency test, which partly reflects his general attitude about the logic of hypothesis testing. He raises concerns about cherry picking and controlling the Type I error rate, but since the target article and the reply to Simonsohn already covered these issues I will not discuss them further. Johnson raises one novel claim: that the application of the consistency test to Topolinsky and Sparenberg (2012) was done incorrectly. I disagree with Johnson's claim, but the topic deserves further discussion because it is possible to misapply the consistency test, and it might be fruitful to discuss the process.

The crux of Johnson's complaint is that my application of the consistency analysis picked the wrong statistics to form the basis of the power analysis. As Johnson notes, it can sometimes be difficult to identify which statistics correspond to an author's theoretical conclusions. This difficulty already suggests that scientists are sometimes not making clear arguments, but I think Topolinsky and Sparenberg (2012) were generally clear about which statistics were important for their theoretical claims.

For Experiment 1 in Topolinsky and Sparenberg (2012), I used two different statistics (t-tests from independent groups) that measured preference for novel versus old items. One statistic was for participants who turned cranks counterclockwise and the other statistic was for participants who

turned cranks clockwise. Both statistics were significant (in opposite directions), but the latter statistic had a p value not much smaller than 0.05 and so a post hoc power value close to one half. Johnson suggests that the relevant statistic was actually the interaction between these two measurements, which was very strong and had a post hoc power value close to one.

To identify the relationship between the data and the theoretical position, it is often necessary to read the text. The title of the Topolinsky and Sparenberg (2012) paper is "Turning the hands of time: Clockwise movements increase preference for novelty." The title makes it quite clear that it is not the interaction in Experiment 1 that matters but the finding that participants who turned a crank clockwise gave higher preference ratings for novel items than old items. The abstract is very specific on this point: "Supporting this hypothesis, participants who turned cranks counterclockwise preferred familiar over novel stimuli, but participants who turned cranks clockwise preferred novel over old stimuli, reversing the classic mere exposure effect (Experiment 1)." The interaction is related to this effect, but an interaction that showed only an effect for counterclockwise rotation but not clockwise rotation would be inconsistent with the title and theoretical claims. To base the power on only the interaction from Experiment 1 misrepresents the arguments presented by Topolinsky and Sparenberg (2012).

Johnson does not have any disagreement about the selection of the statistics in Experiments 2 and 4, and I agree with his comment that Topolinsky and Sparenberg (2012) somewhat muddle the relationship to their theoretical claims by including seemingly superfluous moderating variables in Experiment 2. Nevertheless, I think the statistics we both selected are the appropriate ones relative to the theoretical claims.

In Experiment 3 Topolinsky and Sparenberg (2012) measured two dependent variables. Because one sample of subjects provided two scores, it appears to be impossible to estimate the probability of a random sample showing the desired results for both variables (one would need to know the correlation between variables across participants). Johnson chooses to treat the measures as independent and multiply their separate power values. This approach gives an estimated minimum power. My approach uses only the smaller individual power value, which is more generous to Topolinsky and Sparenberg (2012) because it provides an estimated maximum power value.

So, Johnson and I agree on two of the experiments, I am more generous in estimating power for one experiment, and Johnson provides an interpretation contrary to the original authors for another experiment. I think the consistency analysis reported in the target article holds up quite

well.

4 Vandekerckhove, Guan, and Styrcula

Vandekerckhove *et al.* (2013) correctly point out that applying the consistency test as a filter to remove apparently biased experiment sets might actually leave a field more biased, because experiment sets with very overestimated effect sizes will be consistent. They conclude that the consistency analysis is useless as a tool for meta-analysis. Vandekerckhove *et al.* raise some important issues, but their conclusion goes too far.

First, as Vandekerckhove *et al.* noted just before their depressing conclusion, the consistency test remains useful for individual audits. The inability to detect some experiment sets with bias does not render meaningless the cases where it does indicate bias.

Second, Vandekerckhove *et al.* may have been asking too much of the consistency test: it cannot turn straw into gold. In their simulations, all of the experiments in a simulated multi-study paper were subject to a file drawer bias. Thus, filtering out the experiment sets that appear to be inconsistent leaves the meta-analysis with biased effect sizes from consistent experiment sets. If bias is so pervasive that it includes 100% of experimental findings, then there is no kind of filter that is going to salvage the data. A perfect bias filter would simply end up with no experiments to pool.

A more reasonable scenario is that a field of study includes some biased experiment sets and some unbiased experiment sets. Figure 2 plots simulated meta-analysis pooled effect sizes from experiment sets that varied the proportion of biased sets. Each experiment used a two-sample, two-tailed, t-test and drew samples from normal distributions centered on zero and 0.4 (the true effect size) with standard deviations of one. Unbiased experiments took samples of size $n_1 = n_2 = 40$. Ten such experiments defined a set that were published and could be analyzed with the consistency test.

For a proportion of experiment sets, bias was introduced in two ways. First, each biased experiment followed an optional stopping rule, where sampling started with $n_1 = n_2 = 10$ and continued in steps of one up to a sample size of 40. At each sample the data was analyzed to test for significance, and sampling stopped when significance was found. In addition, experiments that never produced a significant result were put in a file drawer and did not contribute to the published

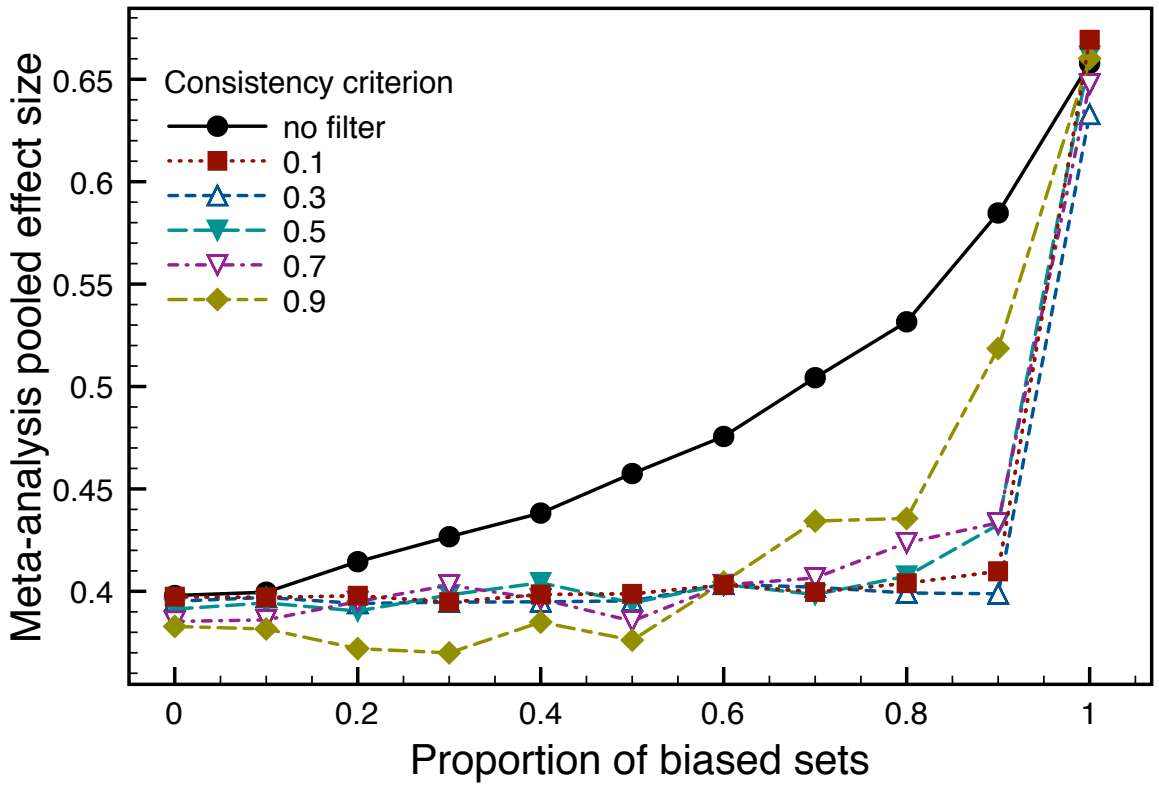

Figure 2: Pooled effect size estimates from a large set of experiments as a function of the proportion of biased experiment sets. Separate curves indicate filtering out experiments sets that appear inconsistent by the given criterion. The pooled effect size is highly over-estimated when all experiments are biased, but the effect size estimate is more accurate (0.4) when some of the experiment sets are unbiased and apparently biased experiment sets are filtered out by the consistency test.

experiment set. This combination of biases is favorable for the consistency test because these joint approaches tend to produce sets where every experiment just barely rejects the null hypothesis. The consistency test identified such biased sets about 94% of the time.

Each data point in Figure 2 is based on 1000 experiment sets that had the indicated proportion of biased experiment sets. The consistency test was applied to each set, and experiment sets that did not pass the test (whether truly biased or not) were removed from a subsequent meta-analysis that pooled effect sizes across all remaining experiments. The solid line in Figure 2 shows the influence of having more biased sets without filtering by the consistency test. As noted in the target article and elsewhere, the biased sets inflate the pooled effect size. The other lines in Figure 2 correspond

to filtering by the consistency test with different criteria for P_c. The data points on the far right are when all experiment sets are biased, which is similar to the case considered by Vandekerckhove *et al.* As they noted, the consistency test filter provides no benefit and possibly exaggerates the effects of bias. However, if some of the experiment sets are not biased (any proportion but one), the consistency-filtered experiments give an improved estimate of the true effect size (0.4), and this is true even for the conservative criterion of 0.1. More liberal criteria, above 0.5, do introduce problems, but are still better than no filter at all.

I do not want to make too much of the results in Figure 2. The success of the filtering process depends on the ability of the consistency test to identify bias. With the combined optional stopping and file drawer biases, the consistency test can filter out a good number of biased experiment sets and thereby allow the unbiased sets to dominate the meta-analysis. For other settings the conservative nature of the consistency test will mean that many biased experiment sets will contribute to the meta-analysis and thereby lead to an overestimated pooled effect size. Of course, scientists do not know whether they are working with totally biased or unbiased data sets, so there are no clear markers to indicate whether the consistency test is an effective filter or not. It is unlikely that any kind of filter is going to allow a meta-analysis to correct for all possible sources of bias. Fundamentally, the best way to produce good scientific data is to run unbiased experiments.

5 Ioannidis

I appreciated the insights and clarifications provided by Ioannidis's (2013) comment about the properties and pitfalls of exploring bias. The closest we come to a disagreement is about the name for the analysis. I picked the term "consistency" because it captures the fundamental basis of the analysis. Ioannidis prefers the term "Test for Excess Significance" (TES), which captures the practical comparison. Since the test was developed by Ioannidis and Trikalinos (2007), I defer to their judgment about the name. Too avoid confusion, I continue to use the term consistency test in this reply, but my future work will refer to the analysis as TES.

I wanted to make one comment on the relationship between bias and estimated effect sizes. Ioannidis notes that various biases that produce an excess of statistically significant results also tend to inflate the summary effect size. In general this is true, but there is a case that was pointed out to me by Jeff Miller (University of Otago) where the reported effect size and power are both

underestimated because non-stochastic effects are treated as sampling variability. Consider a two-sample, two-tailed, t-test that compares a control (no effect) and an experimental group. Suppose that, unknown to the experimenter, most (75%) of the scores (e.g., females) in the experimental group do not show an effect, but 25% (males) show a very large effect (say, a standardized score of 10). Finally, suppose that an experimenter takes samples of size $n_1 = n_2 = 20$, always makes sure to have a 75% and 25% split between the females and males, but ignores the subgroups in the analysis by combining their data. Such experiments tend to produce a Hedges's g effect size of around 0.75 and an estimated power of around 0.64.

These calculations dramatically misrepresent reality. The probability of an experiment like this rejecting the null hypothesis is close to 0.9. This rejection rate is much higher than the estimated power because the effect for males is not stochastic, its influence is present in every sample. The effect size and power calculations suppose that the variability due to a male score of 10.2 and a female score of 0.2 is the result of random sampling from a common population defined by a single normal distribution. In reality some of the variability between scores is due to sampling from different distributions. Another way to describe the situation is that the t-test analysis used to test for significance is inappropriate because the data do not follow the assumptions of the analysis (likewise the g effect size is an inappropriate and inaccurate description of the population). Repeated experiments with this kind of inappropriate data analysis will trigger the TES. In 1000 simulated sets with ten experiments each, 60% of the sets generated a P_c value below the 0.1 criterion, even though there was no file drawer or optional stopping. Thus, one interpretation of violating the TES/consistency analysis is that the sample data contains (possibly unknown) variables that violate the assumptions of the significance test. As many statisticians have emphasized, scientists need to look at their data and not just blindly apply significance tests.

6 Gelman

I appreciated the comments from Gelman (2013) about hypothesis testing and the pitfalls involved in gathering empirical data. The closest we might come to a disagreement is on whether the consistency test is weak compared to other available information. I do agree that a scrutiny of methods can identify areas of concern regardless of statistical issues. One troublesome characteristic of journal articles in psychological science is incomplete reporting of methods. For example, authors

almost never report how the sample size was selected, and it appears common to not report all measured variables. In such an environment something like the consistency test might be the only option available.

Gelman takes a stronger stance than me against hypothesis testing. My own views on the issue are changing on a monthly basis, and his observation about data contributing to a decision regardless of hypothesis testing seems reasonable. My current view is that data analysis should focus on describing data, and no hypothesis test is required. In contrast, model testing and identification can benefit with some form of hypothesis testing (probably a Bayesian one) in order to identify and test quantitative theories. An experimental psychologist who focuses on measurement precision and testing of quantitative models will have little motivation to produce biased data.

7 Conclusions

The discussion has been fruitful and highlighted the benefits, limitations, and difficulties in investigating bias with the consistency test. In my view, the discussion indicates that the consistency/TES analysis is likely to provide a beneficial check on scientific investigations and publishing practices. Applying the analysis to specific sets of experiments is warranted because such sets are the focus of scientific arguments. Ultimately, the most important outcome may be to identify and address problematic methods in current practice. With such a focus we can identify alternative methods that can tap the full potential of psychological science to gather important information about human behavior, develop theories to predict such behavior, and identify ways to use that knowledge to benefit society.

References

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, **100**, 407–425.

Francis, G. (2012a). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin & Review*, **19**, 151–156.

Francis, G. (2012b). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review*, **19**(6), 975–991. doi: 10.3758/s13423-012-0322-y

Francis, G. (2012c). The psychology of replication and replication in psychology. *Perspectives on Psychological Science*, **7**(6), 580-589. doi: 10.1177/1745691612459520

Francis, G. (2013). Replication, statistical consistency, and publication bias. *Journal of Mathematical Psychology*, http://dx.doi.org/10.1016/j.jmp.2013.02.003.

Gelman, A. (2013). Interrogating p-values. *Journal of Mathematical Psychology*, http://dx.doi.org/10.1016/j.jmp.2013.03.005.

Ioannidis, J. P. A. (2013). Clarifications on the application and interpretation of the test for excess significance and its extensions. *Journal of Mathematical Psychology*, http://dx.doi.org/10.1016/j.jmp.2013.03.002.

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

Johnson, V. E. (2013). Comments on "Replication, statistical consistency, and publication bias." *Journal of Mathematical Psychology.*

Morey, R. D. (2013). The consistency test does not-and cannot-deliver what is advertised: A comment on Francis (2013). *Journal of Mathematical Psychology*, http://dx.doi.org/10.1016/j.jmp.2013.03.004.

Simonsohn, U. (2012) It does not follow: Evaluating the one-off publication bias critiques by Francis (2012a,b,c,d,e,f), *Perspectives on Psychological Science*, **7**(6), 597–599.

Simonsohn, U. (2013). It really just does not follow, comments on Francis (2013). *Journal of Mathematical Psychology*.

Simonsohn, U., Nelson, L. & Simmons, J. (in press). *P*-curve: A key to the file drawer. *Journal of Experimental Psychology: General*.

Topolinski, S. & Sparenberg, P. (2012). Turning the hands of time: Clockwise movements increase preference for novelty. *Social Psychological and Personality Science*, **3**, 308–314.

Vandekerckhove, J., Guan, M. & Styrcula, S. A. (2013). The consistency test may be too weak to be useful: Its systematic application would not improve effect size estimation in meta-analyses. *Journal of Mathematical Psychology*, http://dx.doi.org/10.1016/j.jmp.2013.03.007.