i-PERCEPTION

a Pion publication

i-Perception (2012) volume 3, pages 176 –178

dx.doi.org/10.1068/i0519ic ISSN 2041-6695 perceptionweb.com/i-perception

i-Comment

The same old New Look: Publication bias in a study of wishful seeing

Gregory Francis

Department of Psychological Sciences, Purdue University, West Lafayette, IN, USA; email: gfrancis@purdue.edu

Received 16 March 2012, in revised form 21 March 2012, published 22 March 2012

Abstract. A recent study reported evidence of "wishful seeing," where observers reported seeing a desired object as being closer than other objects. A statistical analysis of the experimental findings reveals evidence of publication bias in the study, so the existence of wishful seeing remains unproven.

1 Introduction

Balcetis and Dunning (2010) reported five experiments with evidence that desirable objects were judged closer than other objects; an effect they termed "wishful seeing." Every experiment rejected the null hypothesis, thereby indicating evidence of the effect, and such replication across experiments is often taken as evidence that an effect is robust. However, this interpretation is valid only if the experiments have high statistical power, which is the probability that an experiment will reject the null hypothesis. If all experiments reject the null hypothesis despite having relatively low power, then the correct interpretation is that there was a publication bias that over-reports positive findings (Ioannidis and Trikalinos 2007; Francis 2012, in press).

2 Power analysis

Table 1 lists the sample sizes, standardized effect size, and power of each experimental finding in Balcetis and Dunning (2010) that investigated wishful seeing. The application of a meta-analytic method (Hedges and Olkin 1985), which pools the effect sizes across the experiments, reveals that the best estimate of the effect of wishful seeing is $g^* = 0.537$. The last column of Table 1 shows the power of each experiment to detect this pooled effect size. It is noteworthy that the two studies with the smallest samples sizes have power values less than one half.

Table 1. Statistical properties of the Balcetis and Dunning (2010) experiments on wishful seeing. Effect sizes were computed from the reported *t*-tests.

Description	N1	N2	Effect size	Power from pooled ES
Study 1	47	43	0.418	0.712
Study 2a	61	60	0.513	0.834
Study 2b	42	47	0.423	0.706
Study 3a	20	20	1.063	0.381
Study 3b	26	26	0.626	0.476

The sum of the power values (3.11) is the expected number of times these experiments should reject the null hypothesis. The probability that all five experiments would reject the null hypothesis is the product of the power values (0.076), which is below the 0.1 threshold that is frequently used to indicate evidence of publication bias (Begg and Mazumdar 1994; Ioannidis and Trikalinos 2007). Another way to describe this finding is that the reported experiments are not self-consistent. Given the reported effect and sample sizes, it is not believable that there would be so many rejections of the null hypothesis if the experiments were run properly and reported fully. The proper interpretation of the experimental findings is that they are non-scientific or anecdotal.

It might be tempting to argue that the probability of the Balcetis and Dunning (2010) experiments

is not much below the criterion, so maybe there is hope that future experiments could make the findings more believable. Although it is mathematically possible, such a situation is unlikely because Ioannidis (2008) notes that most experiments overestimate the true effect size, so the above analysis probably overestimates the true power of the experiments. Even if this were not true, new experiments are unlikely to change the conclusion of publication bias. If a new experiment rejects the null hypothesis with a similar effect size, then the product of the power values for all experiments can only be less than the product for the experiments in Table 1. If a new experiment fails to reject the null hypothesis, it will usually have a smaller effect size than what is shown in Table 1. The pooled effect size across experiments will be smaller, which will reduce the power of all experiments. Once publication bias has been found, it is difficult to remove.

3 Interpretation

There are two broad explanations of how publication bias could have contaminated the findings in Balcetis and Dunning (2010). First, they may have run, but not reported, additional experiments that did not reject the null hypothesis. This type of "file drawer problem" could happen because the authors deliberately suppressed some findings or because reviewers or the editor insisted that the null/negative findings be removed from the manuscript. Something similar to the file drawer problem can also occur for experiments that have multiple measures but report data from only a subset of the measures.

The second broad explanation is that the experiments in Balcetis and Dunning (2010) were run improperly in a way that caused an elevated rejection rate for the null hypothesis. One invalid approach is to start with a relatively small set of subjects and run a hypothesis test. If the null hypothesis is not rejected, additional subjects are recruited and the test is repeated. This procedure is continued until the null hypothesis is rejected or the experimenter gives up. It may seem like good scientific practice to gather data until a research question is settled, but analyzing such data sets as if they were gathered with a fixed sample size leads to a dramatic increase in the rejection of the null hypothesis, regardless of whether it is true or false (Strube 2006). There are several other experimental methods that also produce too many rejections of the null hypothesis. Simons, Nelson and Simonsohn (2011) describe how some of these techniques can ensure that almost every experiment rejects the null hypothesis, regardless of whether it is true or false. This too frequent rejection of the null hypothesis will show up as publication bias.

There is no way of telling which of these broad approaches, and it could be both, were used by Balcetis and Dunning (2010). In a similar way, now that the data are known to be contaminated with publication bias, there is no way to determine whether the null hypothesis is true or false. Researchers interested in wishful seeing are advised to ignore the findings in Balcetis and Dunning (2010) and run new experiments without bias.

A third possible explanation of the pattern of data in Table 1 is that the studies measured different effect sizes, in which case the meta-analytic pooling is improper. For example, some experiments measured estimates of distance, while experiment 3a measured accuracy of distance related actions (tossing a beanbag to a target). The effect of wishful seeing as expressed by the accuracy of tosses might alter the reported effect size. However, the reported effect sizes are inconsistent with this explanation. The action of tossing a beanbag might scale the overall magnitude of the measurement variable, but it will also introduce an additional noise term to the experimental measurements, which will increase the standard deviation. A change in scale will not alter the standardized effect size, but a larger standard deviation will decrease the effect size. Based on this analysis, one might expect that experiment 3a will have a smaller standard deviation than the other experiments, but Table 1 shows that experiment 3a has the largest effect size of all of the experiments.

4 Conclusions

The study of Balcetis and Dunning (2010) is one of several new studies (eg, Balcetis and Lassiter 2010) that have revived ideas of the New Look theorists from the 1950s. The New Look approach argued that an observer's motivations and desires could alter perceptual experience, but the New Look findings were ultimately rejected because of poor methodology. If the publication bias in Balcetis and

Dunning (2010) is present in similar studies, then the empirical efforts to revive the ideas of the New Look theory may suffer from variations of the methodological problems of the past.

References

Balcetis E, Dunning D, 2010 "Wishful seeing: More desired objects are seen as closer" *Psychological Science* **21** 147–152

Balcetis E, Lassiter G D, 2010 *Social psychology of visual perception* (New York: Psychology Press)

Begg C B, Mazumdar M, 1994 "Operating characteristics of a rank correlation test for publication bias" *Biometrics* **50** 1088–1101 doi: 10.2307/2533446

Francis G, 2012 "Too good to be true: Publication bias in two prominent studies from experimental psychology" *Psychonomic Bulletin & Review* **19** 151–156 doi: 10.3758/s13423-012-0227-9

Francis G, in press "Publication bias in "Red, Rank, and Romance in Women Viewing Men" by Elliot et al. (2010)" *Journal of Experimental Psychology: General*

Hedges L V, Olkin I, 1985 *Statistical methods for meta-analysis* (New York: Academic Press)

Ioannidis J P A, 2008 "Why most discovered true associations are inflated" *Epidemiology* **19** 640–648 doi:10.1097/EDE.0b013e31818131e7

Ioannidi, J P A, Trikalinos T A, 2007 "An exploratory test for an excess of significant findings" *Clinical Trials* **4** 245–253 doi:10.1177/1740774507079441

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 doi: 10.1177/0956797611417632

Strube M J, 2006 "SNOOP: A program for demonstrating the consequences of premature and repeated null hypothesis testing" *Behavior Research Methods* **38** 24–27 doi:10.3758/BF03192746