A false-positive error in search in selective reporting: A refutation of Francis

Running head: Refutation of Francis

Emily Balcetis(1), email: emilybalcetis@nyu.edu & David Dunning(2), email: dad6@cornell.edu
1. Department of Psychology, New York University, New York, New York
2. Department of Psychology, Cornell University, Ithaca, New York

Abstract
Francis' statistical method for ferreting out selective reporting fails when underlying questionable assumptions inherent in that report are relaxed or other reasonable choices are made. We suggest that surveying the literature for suspicious studies will, of course, lead to research that looks suspicious just by chance. Francis fails to correct for these problems.

Introduction

Francis claims to find evidence for selective reporting in our 2010 *Psychological Science* article suggesting that people see desirable objects as closer than non-desired ones (Balcetis & Dunning, 2010). As such, he calls our major conclusion into question. We respond that Francis' claims are inappropriate and overstated.

Our Analysis

Francis (2012a) bases his conclusions on a technique (Ioannidis and Trikalinos 2007; see also Begg and Mazumdar 1994) testing the likelihood that researchers would find a run of statistically significant findings given the underlying statistical power of their studies. Francis concludes that our studies were too underpowered to allow for a run of five significant findings. That is, our effect sizes and sample sizes were so small that we had a chance of merely .076 (or 7.6%) to produce the unanimously successful studies we reported. Thus, *j'accuse*! Francis concludes that there must be other studies out there, most likely showing null results, they we did not report.

The appropriateness of Francis' method rests on a strong assumption that is likely wrong. Francis assumes that there is a uniform effect size across our studies. This would be an appropriate assumption if we had used the exact same operationalization of independent and dependent variables in each study. However, this assumption is not appropriate when operationalizations of independent or dependent variables vary. When they do, effect sizes will vary because some instantiations of the independent variable will be stronger or more valid than others. Effect sizes may also vary because some dependent measures are more sensitive to underlying psychological states. With experimental variations, effect sizes across studies were likely to be *heterogeneous* rather than *homogeneous*.

In our five-study package, operationalizations of desirability (the independent measure) varied a great deal. For instance, water became more desirable to the participants we made thirsty compared to those whose thirst we

quenched. Some target objects were more desirable because they carried financial value. Some brown objects were made more desirable because they were shaped like a chocolate truffle rather than dog feces. Measures of the dependent variable of distance estimation also differed substantially. Some studies required verbal reports of numerical estimates whereas others required action-based responses.

It is already well-known that Francis' (2012a) method is often misleading in the heterogeneous case. As Ioannidis and Trikalinos (2007) themselves clearly state: "Applying the test ignoring genuine heterogeneity is ill-advised" (p. 246). Thus, we redid Francis' power analysis relaxing his strong assumption. Instead of assuming a common effect size, we assumed that the best estimate of a study's effect size was the one reported for that specific study (as listed in his Table 1). Running the analysis this way, we find that the chance of achieving five significant results is .116. To indicate selective reporting, that figure should be smaller than .100. Thus, when the appropriate test is conducted given the heterogenous nature of the effect sizes, Francis' test fails.

His analysis also fails when one makes a different, yet quite appropriate, choice in how a common effect size is calculated. Francis calculated a common effect size giving weight to the number of participants in each study. This is a perfectly appropriate choice, but probably more so in the homogeneous situations Francis assumes. When operationalizations differ, it is also appropriate to give each study equal weight. This is because one does not know which studies contain the most valid or representative instantiations of independent and dependent variable. Thus, we redid Francis' analysis using a common effect size that weighted each study equally. The resultant effect size was slightly bigger (.609 rather than .537). Using it, we found that the chance that all five studies achieved significance was .163, which again did not reach the .100 threshold.

But Francis makes an additional egregious error in his critique. He implies that selective reporting calls the very validity of our central conclusion into question. However, this is not the usual inference researchers draw in cases where selective reporting is found. Researchers are cautioned against inferring the null hypothesis. Rather, they are instructed to conclude that the central hypothesis is likely true but that its magnitude—its effect size—has been overstated (Ioannidis 2008). Thus, even if we conceded Francis' central assertion, the sin we would be most likely guilty of is overstating the magnitude with which desirability influenced perceived distance, not falsely claiming that the effect exists.

But, full disclosure: There was, in fact, one study in this research program that we did not report in our 2010 article (Balcetis & Dunning, 2010). In that study, we followed almost exactly the methods reported in our Study 3B. We found that participants stood further away from brown objects fashioned to look like truffles than those fashioned to look like dog feces, presumably because they saw the desirable truffles as closer. The result, however, was only marginally significant, $t(61) = 1.91$, $p = .061$, $d = .50$, two-tailed. We included this study in our original submission to the journal, but the editor and reviewers stated our approach to studying potential emotional influences needed amendment. Thus, we completely re-ran the study with new emotion measures and reported this new study as Study 3B in the published paper. What is the effect of not reporting that study? It appears

that our published work overstates the effect size of our central finding, assuming a common effect size, by .002 (g^* = .537 in the published studies versus .535 for all studies). However, the statistical significance of our central hypothesis would have strengthened if we had included this sixth study, Stouffer's Z rising from 5.20 to 5.52, $p < .0001$. Now that all our studies have been reported, Francis should be relieved to discover that the chance of achieving 5 significant results out of 6 studies now lies at .225 (.362 if we weight studies equally in computing a common effect size).

Conclusion

In sum, we find that Francis' statistical claims that we selectively reported our findings fail to survive alternative ways of conducting his test. His claims fail when we relax an assumption that is most likely to be false in our studies. It fails when we make different but appropriate choices in calculating effect size. The conclusions he reaches even if selective reporting is stipulated are incautious at best and overblown at worst.

And one last note. It appears that Francis might be scouring the literature widely for potential instances of selective reporting across a number of diverse research topics (Francis 2012b, in press). Although we endorse vigilance when it comes to scientific rigor, it appears that Francis has forgotten that a wide-ranging scrutiny for suspicious articles violates a central tenant of rigor. If one examines dozens or hundreds of articles, one will—just by chance—find some that spuriously fit criteria for suspiciousness. The fact that, in the pursuit of rigor, Francis fails to discuss or correct for this issue is astonishing. That he does not, in light of this issue, conduct analyses to see if his conclusions are robust (as we did above) is astonishing.

Francis attempts to test if researchers are cherry-picking the studies they include in their articles, but in so doing may have fallen prey to the very phenomenon in which he is interested. In effect, Francis' project becomes ironic. In the pursuit of routing out "false-positive" findings, Francis makes himself susceptible to the same methodological sin he purports to oppose: making false positive claims in the pursuit of favored conclusions.

References

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

Begg C B, Mazumdar M, 1994 "Operating characteristics of a rank correlation test for publication bias" Biometrics 50 1088-1101

Francis G, 2012a "The same old New Look: Publication bias in a study of wishful seeing. i-Perception.

Francis G, 2012b "Too good to be true: Publication bias in two prominent studies from experimental psychology" Psychonomic Bulletin & Review 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

Ioannidis J P A, 2008 "Why most discovered true associations are inflated" Epidemiology 19 640-648

Ioannidi, J P A, Trikalinos T A, 2007 "An exploratory test for an excess of significant findings" ClinicalTrials 4 245-253