# Checking the counterarguments confirms that publication bias contaminated studies relating social class and unethical behavior

Gregory Francis
Department of Psychological Sciences
Purdue University
gfrancis@purdue.edu
23 May 2012

In their reply to my criticism of the published findings in Piff *et al.* (2012a), Piff *et al.* (2012b) proposed three counterarguments and one observation. I will rebut the counterarguments and try to clarify the observation.

The analysis in Francis (2012c) was a modification of an approach proposed by Ioannidis and Trikalinos (2007). The basic argument was that the power values for the experiments in Piff *et al.* (2012a) were usually (with one exception) close to 0.5, so the probability that all seven experiments would reject the null hypothesis was so small that one doubted that the experiments were run properly and reported fully. Piff *et al.* (2012b) made three counterarguments to suggest that the analysis in Francis (2012c) was invalid. In fact, an investigation of the counterarguments only leads to further confirmation of the analysis in Francis (2012c).

**Counterargument (1).** *The power analysis used in Francis (2012c) requires a large sample of studies*

As Piff *et al*. (2012b) noted, Ioannidis and Trikalinos (2007) developed their power analysis for application to relatively large sets of experimental findings. If an experiment set has both rejections and non-rejections of the null hypothesis, then the analysis does require a fairly large number of experiments to be able to draw firm conclusions. However, the findings in Piff *et al.* (2012a) are an extreme pattern of findings: *every* experiment rejected the null hypothesis. Under such an extreme pattern, a small number of experiments will suffice to make a convincing argument for publication bias.

Consider a more familiar situation. I tell you that I have a fair coin that will show heads or tails equally often. I flip the coin many times and tell you that it showed heads seven times. If the coin is fair, the probability of seven heads (out of an implied seven flips) is $(0.5)^7$=0.0078125. Following the logic of hypothesis testing, it is reasonable for you to conclude that the coin is not actually fair. The analysis in Francis (2012c) is analogous, with a flip showing heads being replaced by a report of rejecting a null hypothesis and with the concept of a fair coin being replaced by unbiased experimental findings. The numbers are a bit different because the power values in Piff *et al.* (2012a) were a bit larger than 0.5 (and for one experiment much larger), but the logic is the same.

The number of experiments does matter for this kind of analysis, but in the opposite direction implied by Piff *et al.* (2012b). If you are trying to decide whether a coin is biased, then it will be difficult to use hypothesis testing to make a correct conclusion after only two coin flips because the probability of two heads from a fair coin is 0.25, which is not so rare. Likewise, it is difficult to find evidence of publication bias from a small number of experiments. Given the power values for the experiments in Piff *et al.* (2012a), seven experiments is more than enough.

**Counterargument (2).** *The power analysis in Francis (2012c) inappropriately assumed homogeneity in effect sizes*

When valid, a pooling of effect sizes across experiments gives the best estimate of the common effect size (Francis, 2012a,b). However, Piff *et al.* (2012b) are correct that such pooling might be inappropriate for their findings given their widely different measurement techniques. Indeed, this is why the analysis in Francis (2012c) did not pool the standardized effect sizes. The power values in Table 1 of Francis (2012c) are not based on a pooled effect size. Each power value is calculated from the effect size for an experiment, and this calculation is sometimes called observed or post-hoc power. Counterargument 2 of Piff *et al.* (2012b) simply does not apply to the analysis in Francis (2012c).

**Counterargument (3).** *The power analysis in Francis (2012c) inappropriately used observed power*

Piff *et al.* (2012b) correctly noted that effect sizes computed from experimental data are estimates and that confidence intervals can characterize the variability of these estimates. They noted that if one takes the upper limit of the pooled effect size's confidence interval, then the product of the power probabilities could be as high as 0.881. (Given counterargument 2, it is odd that they used a pooled effect size, but one gets a similar number by taking the upper limit of the confidence interval for each experiment's effect size.) However, this calculation is not a proper way of building a confidence interval for any statistic (Kelley, 2007), including the probability of rejections for an experiment set.

A proper investigation of the uncertainty about the probability of experiment set rejections (ESR) can be found with simulated experiment sets. For each experiment, the simulation sampled an effect size value from the distribution of effect sizes suggested by the experimental data in Piff *et al.* (2012a). Using the samples sizes from Piff *et al.* (2012a) and the sampled effect sizes from all seven experiments, the power analysis in Francis (2012c) was applied to estimate the probability that all seven experiments would reject the null hypothesis (the product of the power values). Out of 100,000 such simulated experiment sets, the 2.5% quantile of the ESR probability was 0.0000584 while the 97.5% quantile was 0.107. The distribution was quite skewed, with a median value of 0.004 and a 99% quantile of 0.161. The 0.881 value noted by Piff *et al.* (2012b) can happen, but it is exceedingly rare. Moreover, just as statisticians running hypothesis tests do not generally compute confidence intervals of *p* values (they are bigger than you might think!, see Cumming, 2012), so too one does not draw inferences from a confidence interval of the ESR

probability. The analysis in Francis (2012c) properly drew an inference based on the probability of the observed data. What the simulation analysis does tell us is that the calculation of the ESR probability is quite robust to variations in the estimated effect size. That is, the bias in Piff *et al.* (2012a) is so convincing that any reasonable miss-estimation of effect size matters very little.

There is a somewhat related issue regarding the properties of observed power. A common way of measuring statistical power is to use the experimentally estimated effect size as the basis for a power calculation. This was the approach used by Francis (2012c), but there are other possibilities. For example, Gillett (1994) suggested computing expected power, which requires a distribution of plausible effect sizes and weights the resulting power for each effect size by the probability of that effect size. Integration of the weighted powers gives the expected power. This approach is not commonly used (because one is never sure how to define the effect size distribution), and it tends to produce smaller (though possibly more accurate) estimates of power than the standard approach. Using this calculation in the power anlaysis of Francis (2012c) would only further reduce the ESR probability for the findings in Piff *et al.* (2012a).

Piff *et al.* (2012b) are correct in noting that observed power has sometimes been misapplied, because people have tried to use it as evidence of no effect for experiments that did not reject the null hypothesis. There are philosophical problems with this kind of approach that largely reflect limitations of hypothesis testing. However, the analysis in Francis (2012c) explored experiments where this criticism cannot be levied because every experiment in Piff *et al.* (2012a) rejected the null hypothesis.

There is a more general concern about observed power, which is that it may be a poor estimate of true power. Indeed, observed power can be systematically biased (Yuan & Maxwell, 2005) such that true power is underestimated when true power is bigger than 0.5 and overestimated when true power is smaller than 0.5. When true power equals 0.5, observed power follows a uniform distribution between zero and one. The problem is that the symmetry of the estimated effect size distribution around the true value leads to an asymmetry in power estimates because a shift in one direction causes a larger change in power than an equivalent shift in the other direction (due to the differing areas in the tails of the sampling distribution).

To insure that power underestimation was not responsible for the conclusion of publication bias, additional simulated experiments were run to explore the false alarm rate of the power analysis. The experiment set emulated the six one-sample experiments in Piff *et al.* (2012a), which by themselves were enough to draw a conclusion of publication bias (one other experiment had a larger effect size, but its power is less than one so the set of seven is always less probable than the set of six). Each experiment set consisted of eleven experiments (one-sample *t*-tests) with an effect size chosen from a normal distribution with a mean 0.177 and standard deviation 0.0418 that mimicked the effect sizes reported by Piff *et al.* (2012a). Sample sizes were chosen to produce a true power of 0.53, which is slightly bigger than the power values in Piff *et al.* (2012a). To avoid unreasonably large sample

sizes, any effect size less than 0.05 was reset to the value 0.05. Sample data were drawn from a normal distribution with a mean of the experiment's effect size and a standard deviation of one. With these choices, the set of eleven experiments averaged 5.85 rejections of the null hypothesis.

The power analysis used by Francis (2012c) was then applied to the experiment set under the case where all eleven experiments were fully reported (no bias). That is, the effect size of each experiment was computed and used to produce the observed power. Since typically not all eleven experiments rejected the null hypothesis, a Fisher's exact test was used to compute the probability that the experiments would produce the observed (or more) rejections of the null hypothesis given the post hoc power values. Publication bias was concluded if this probability was less than 0.1 (Ioannidis & Trikalinos, 2007; Francis, 2012a,b).

Out of 100,000 such unbiased simulated experiment sets, only 1,211 concluded evidence for publication bias. The false alarm rate is 0.01211, so fears that underestimation of power might lead to false reports of publication bias are unfounded, at least for experiments with properties similar to those reported in Piff *et al.* (2012a).

Moreover, the test is extremely conservative even when bias does exist. From the same simulations, experiment sets were created with a file drawer bias that did not report the experiments that failed to reject the null hypothesis. Applying the same kind of power analysis to these biased sets concluded evidence of bias in 21,450 experiment sets. The low hit rate (0.2145) is because biased experiment sets often report uncommonly large effect sizes, which then lead to overestimates of the true power values. The bottom line is that the power analysis only catches the most egregious cases of publication bias.

**Observation.** *Piff et al. (2012b) do not know the source of bias in their experiments*

Piff *et al.* (2012b) explained that selective reporting was not an issue because they ran seven experiments and published them all. They also reported that a scrutiny of their experimental methods did not reveal any indication of bias. I do not doubt the honesty of Piff *et al.* or their intention to run valid scientific experiments. I suspect that bias crept into their experiments without their intention or realization. There are two broad ways for bias to contaminate a set of experiments.

The first way is by more frequently publishing findings that reject the null hypothesis than publishing findings that do not reject the null hypothesis. This is usually described as a file drawer bias, where findings that do not reject the null hypothesis are deliberately suppressed. Piff *et al.* (2012b) are clear that they did not intentionally suppress such studies, but it can happen unintentionally. For example, researchers sometimes run low powered pilot experiments to explore various methodological details. In some situations these variations make no difference for the experimental effect, and researchers are effectively running multiple experiments. In such a case researchers may not feel it is appropriate to report the pilot experiments, so the reported low powered experiments end

up rejecting the null hypothesis more frequently than is appropriate. Likewise, if an experiment has multiple measures of behavior and researchers only report the measures that reject the null hypothesis (or describe non-rejecting measures as being unrelated to the effect of interest), then there is effectively a suppression of experimental findings. There are several variations of these basic ideas that all lead to something like a file-drawer bias.

The second way to produce an experiment set with bias is to run the experiments incorrectly. Valid hypothesis testing techniques require that researchers take a sample of a fixed size from a population. Any deviation from a fixed sample size can lead to too frequent rejections of the null hypothesis. For example, suppose a researcher gathers data from a random sample of $n$=20 subjects and runs a $t$-test. Suppose the $p$ value comes out to 0.07, which is not below the 0.05 criterion that is commonly used in psychology. Undaunted, the researcher gathers data from a randomly selected additional 5 subjects so that he now has a total of $n$=25. With the new data set he computes $p$=0.04, so he rejects the null hypothesis and reports his result. But there is something like multiple testing going on here, and the second hypothesis test is invalid. The sample is not a fixed $n$=25 because the researcher would have stopped at $n$=20 if the first test had rejected the null hypothesis. This methodological approach is called optional stopping (or data peeking). If the null hypothesis is true, optional stopping increases the Type I error rate. If the null hypothesis is false, optional stopping increases the frequency of rejecting the null hypothesis relative to the power of the final experiment. If an experimenter is willing to keep adding subjects, the probability of rejecting the null hypothesis with this method is effectively 1.0 even if the null hypothesis is true. (See Kruschke, 2010 for a further discussion of optional stopping and how Bayesian methods can avoid some of these difficulties.)

Optional stopping (in various forms) appears to be widely practiced by experimental psychologists (John *et al.*, 2012). It is insidious because it seems to fit naturally with the idea that larger samples are always better in statistics, but that idea is true only for hypothesis tests with a fixed sample size. Optional stopping also seems to fit in naturally with the idea that an experimenter should continue to gather data until finding a definitive result, but this idea is inconsistent with the foundations of hypothesis testing. Simmons *et al.* (2011) discuss some other invalid methodological choices that can increase the frequency of rejecting the null hypothesis.

Biases like optional stopping lead to such serious misconceptions about replication that some researchers appear to not believe the conclusions of a standard power analysis. Consider the experiments reported by Piff *et al.* (2012a). For the first three experiments one might suppose that the researchers had little idea of the effect size and picked samples sizes based on intuition or convenience. If this is the case, then it is rather remarkable that they happened to pick a sample size for each experiment that was just big enough to reject the null hypothesis; but even more remarkable is what happened for the last three experiments. Having already run the earlier experiments, Piff *et al.* should have had a pretty good idea that the effect size was no larger than 0.2 (and probably smaller). Even if they were optimistic, a power analysis with this effect size would have recommended a sample size of $n$=156 to insure a power of 0.8. Instead, they picked sample sizes of 108, 195 and 90,

which would give power values of 0.66, 0.87, and 0.59, respectively. Why would anyone deliberately design an experiment to have a power of around 0.6? The variability in sample sizes and power values suggests that Piff *et al.* (2012a) did not perform a power analysis when designing their experiments, which begs the question: how did they select their sample sizes? We may never know for sure (even the researchers may not know because sometimes data collection is passed off to other members of the lab), but optional stopping seems like a possibility. It would explain why the selected sample sizes were often just big enough to reject the null hypothesis.

## Conclusions

To a certain extent, the exact source of the bias does not matter when deciding how to interpret the findings in Piff *et al.* (2012a). The findings appear to be biased and so we should not consider the reported experiment set to be a valid scientific investigation of the relationship between social class and ethical behavior. Of course, it is possible that there really was no bias and the findings in Piff *et al.* (2012a) just had the extremely bad misfortune of appearing to have bias. However, the probability of the reported data pattern is very low if the experiments were unbiased. Regardless, the scientific interpretation remains clear: the *appearance* of publication bias is enough to cause us to interpret the findings in Piff *et al.* (2012a) as invalid. Any other interpretation is a refutation of the principles of hypothesis testing (in which case one doubts the findings of Piff *et al.* based on skepticism of their own hypothesis tests).

From a scientific point of view, the conclusion of publication bias is a cautious, but mild, criticism about the effect itself. The claim is not that the null hypothesis is true (the effect does not exist), but only that the reported experiments are not valid. The hypothesized relation between social class and ethical behavior may, or may not, be true. Only new experiments that are free of bias can determine the validity of the claim. Given the appearance of publication bias, the findings in Piff *et al.* (2012a) should not be part of the evaluation of the effect.

The evidence of publication bias in Piff *et al.* (2012a) is convincing, and the counterarguments raised by Piff *et al.* (2012b) do not alter this conclusion. I have no doubt that the bias was unintentional, and I suspect that whatever methods introduced the bias are not restricted to Piff *et al.* but are commonly used by other researchers. The field of psychology needs to fundamentally alter how it gathers and draws conclusions from experimental data.

## References

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.

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). The same old New Look: Publication bias in a study of wishful seeing *i-Perception*, **3**, 176-178.

Francis, G. (2012c). Evidence that publication bias contaminated studies relating social class and unethical behavior. *Proceedings of the National Academy of Sciences USA*, doi: 10.1073/pnas.1203591109.

Gillett, R. (1994). Post hoc power analysis. *Journal of Applied Psychology*, **79**, 783-785.

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

John, L. K., Loewenstein, G. & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological Science*, **23**, 524-532.

Kelley, K. (2007). Confidence intervals for standardized effect sizes: Theory, application, and implementation. *Journal of Statistical Software*, **20**. http://www.jstatsoft.org/v20/a08/

Kruschke, J. K. (2010). Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, **1**(5), 658-676. doi:10.1002/wcs.72

Piff, P. K., Stancato, D. M., Côté, S., Mendoza-Denton, R., Keltner, D. (2012a). Higher social class predicts increased unethical behavior. *Proceedings of the National Academy of Sciences USA*, **109**, 4086–4091. doi/10.1073/pnas.1118373109

Piff, P. K., Stancato, D. M., Côté, S., Mendoza-Denton, R. & Keltner, D. (2012b). Reply to Francis: Cumulative power calculations are faulty when based on observed power and a small sample of studies. *Proceedings of the National Academy of Sciences USA*. doi:10.1073/pnas.1205367109

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.

Yuan, K.-H. & Maxwell, S. (2005). On the post hoc power in testing mean differences. *Journal of Educational and Behavioral Statistics*, **30**, 141-167.