

Publication bias in “Expecting to lift a box together makes the load look lighter”

Gregory Francis

Department of Psychological Sciences

Purdue University

phone: 765-494-6934

gfrancis@purdue.edu

17 September 2012

Running head: Publication bias in “expecting to lift a box”

Abstract

Doerrfeld, Sebanz and Shiffrar (2012) reported evidence that perception is shaped by what can be accomplished with other people. Their conclusion was based on consistent findings from four experiments. However, given the effect sizes and sample sizes of the experiments, the probability that all four experiments would reject the null hypothesis is only 0.082 if the effect magnitude is as reported and the experiments were run properly. This low probability suggests that the experiment set contains publication bias and thereby does not provide valid scientific information about the phenomena. Scientists interested in perception and joint actions will need to run new unbiased experiments.

Doerrfeld, Sebanz and Shiffrar (2012) tested whether anticipated effort alters perceptual experience by having individuals judge the weight of a filled basket when they intended to lift the basket either alone or with another person. Four experiments consistently rejected the null hypothesis that the weight judgments were the same across the two intention conditions. The experimental results were interpreted as strong evidence for the validity of the result and confirmation of a theoretical hypothesis that perception is shaped by what can be accomplished with other people.

However, experiments should only reject the null hypothesis at a rate that reflects the power of the experiments. When experiments have low or moderate experimental power, random sampling means that some experiments should not reject the null hypothesis even if the reported effect is true (Francis, 2012a,b,d, in press; Ioannidis & Trikalinos, 2007; Schimmack, in press). The absence of expected null findings indicates publication bias, which makes it impossible to judge whether the reported experiments are valid. As shown below, such is the case for the findings in Doerrfeld *et al.* (2012).

Table 1 shows the key statistical properties of the four experiments. The fourth column reports a measure of effect size (Hedges g) for each experiment. The effect size for Study 1 might seem to differ from the effect sizes of the other three studies, but as Figure 1 shows, the difference is not unusual when considering the range of a 95% confidence interval around the effect size for each experiment. In fact, the standard deviation across the observed effect sizes (0.237) is close to the theoretically predicted value (0.195) based on pooling the effect sizes (Hedges & Olkin, 1985). Given the statistical properties of the effect sizes and given that the experiments used very similar methods and measures, it is appropriate to pool the effect sizes. A meta-analysis that gives more emphasis to

experiments with larger sample sizes, by weighting each experimental effect size by its inverse variance, (Hedges & Olkin, 1985), computes 0.865, $CI_{95}=(0.483, 1.25)$, as the best estimate of the true effect size.

The last column of Table 1 reports experimental power, which is the probability that an experiment rejects the null hypothesis, for the pooled effect size. The sum of the power values, 2.24, is the expected number of times experiments like those in Doerrfeld *et al.* (2012) would reject the null hypothesis (Ioannidis & Trikalinos, 2007). Thus, it is surprising that four experiments rejected the null hypothesis. The probability that four out of four experiments like these would reject the null hypothesis is the product of the power values, which is 0.082. If they were run properly and reported fully, the experimental outcomes in Doerrfeld *et al.* (2012) are quite unusual for the reported effect sizes and sample sizes. The low probability of the experiment set is below the 0.1 criterion commonly used to establish publication bias (Begg & Mazumdar, 1994; Ioannidis & Trikalinos, 2007; Sterne, Gavaghan & Egger, 2000).

Such bias can occur in a variety of ways. One approach is the suppression of non-supportive findings, which is often described as a file-drawer problem, (Rosenthal, 1984); and this kind of bias can dramatically alter the magnitude of a reported effect (Lane & Dunlap, 1978). It may seem unlikely that researchers would deliberately suppress proper experiments that did not show the effect, but there is sometimes a tendency to dismiss negative or null findings as “pilot studies.”

Perhaps even more common are mistakes in sampling and analysis that inflate the rejection rate of the null hypothesis (Simmons, Nelson & Simonsohn, 2011; John, Loewenstein & Prelec, 2012). For example, a common sampling approach starts by

gathering an intermediate set of data and running a statistical analysis to see if the effect of interest is present. If the effect is found, the experiment stops and the finding is reported; but if the effect is not found, additional subjects are recruited and their data is added to the previous set for a new analysis. The cycle of gathering data and testing repeats until finding the effect of interest or the experimenter gives up. Such “data peeking” dramatically inflates the Type I error rate (Strube, 2006), so the test conclusions cannot be trusted.

There is no statistical marker that convincingly reveals the use of a data peeking strategy, but it seems generally consistent with the pattern of results in Doerrfeld *et al.* (2012). In every experiment a notable analytical investigation (not necessarily those in Table 1) produced a p value that was just below the criterion 0.05 value. Study 1 and 2b produced the effects described in Table 1 with p values of 0.048 and 0.046, respectively. In Study 2a, the test reported in Table 1 gives a modest $p=0.028$, but an ANOVA comparing a related interaction produced $p=0.047$. Likewise, in Study 3, the comparison reported in Table 1 gives $p=0.006$, but an ANOVA of a main effect of condition (across four conditions) gives $p=0.045$. Perhaps the sampling process was stopped as soon as a desired statistical significance was found for all of the tests of interest to the authors. Without such a bias it is difficult to imagine how the authors consistently selected just the right sample size so as to barely reject the null hypotheses they were interested in testing.

Equally common may be biased theorizing. For example, it is possible to measure many different variables and then run a variety of analysis and look for post-hoc patterns that make a compelling story. Kerr (1998) described this approach as hypothesizing after the results are known (HARKing). There is no direct way of detecting this practice, but it is notable that Doerrfeld *et al.* (2012) measured several variables that generally seem

unrelated to their main conclusion. For example, each experiment measured weight judgments before lifting the basket and after lifting the basket. Support for the theory was claimed on the basis of the consistent pre-lifting effects. The post-lifting effects were described, but the effects were inconsistent and were treated by the authors as unrelated to the thesis. Given that the outcome was unrelated to the theory, one has to wonder why the post-lifting variable was even measured. Curiously, the presence of null effects for the post-lifting judgments actually makes them more believable than the overly consistent pre-lifting judgments. Whether the post-lifting data support the thesis is something for a subject matter expert to determine. A skeptic might suspect that whichever data set happened to produce a consistent pattern would have been used as evidence for the theory, and such HARKing often produces publication bias.

I describe these possible routes to bias not as an accusation against Doerrfeld *et al.* (2012), who I suspect operated with the best of intentions in designing and reporting their studies, but as a demonstration of how seemingly minor decisions throughout a research project can lead to a biased experiment set. Ultimately, it does not much matter exactly how bias was introduced; the main observation is that the set of experimental results reported by Doerrfeld *et al.* (2012) as evidence for their thesis would be quite rare if they were generated without some form of bias. Since there is no way to know the extent of the bias, readers should be skeptical about the findings and conclusions of the original study.

Anyone wanting to explore the relationship between weight judgments under solo and joint lifting conditions will need to run new experiments. If the effect size is close to the pooled estimate of 0.865, then sample sizes around 22 are required to achieve a power of 0.8. Some biases inflate the effect size estimate, and if the true effect size is half the

value of the pooled estimate, then sample sizes of around 85 (in each group) are required. For such samples, the probability of four out of four such experiments all rejecting the null hypothesis would be 0.41. To achieve a probability of 0.8 for four out of four experiments rejecting the null each experiment would need to have a power of 0.946, which would require 136 subjects in each group.

If such sample sizes seem excessive for establishing scientific evidence for an effect, it is only because current scientific practice in psychology vastly underestimates the uncertainty that persists after an experiment. Figure 1 makes it clear that even statistically significant findings contain much uncertainty about the true magnitude of an effect. Consider Study 2a, which rejected the null hypothesis with $p=0.028$. The 95% confidence interval around the standardized effect size stretches from 0.072 to 2.208. These extremes are equivalent to saying that the effect may essentially be of no practical importance, or that the effect is so large that it could hardly fail to be noticed. In isolation, this experiment provides very little information about the effect. Indeed, convincing scientific evidence in psychology almost always requires a meta-analysis that pools information across experiments. Such meta-analyses can only draw proper conclusions for unbiased experiment sets, so it is critical to identify biased experiment sets that promote a solid set of experimental results.

References

Begg, C. B. & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, **50**, 1088-1101.

Doerrfeld, A., Sebanz, N. & Shiffrar, M. (2012). Expecting to lift a box together makes the load look lighter. *Psychological Research*, **76**, 467-475.

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**(3), 176-178.

Francis, G. (2012c). Replication initiative: Beware misinterpretation. *Science*, **336**(6083), 802.

Francis, G. (2012d). Evidence that publication bias contaminated studies relating social class and unethical behavior. *Proceedings of the National Academy of Sciences*. **109**:E1587.

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*. San Diego, CA: Academic Press.

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.

Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, **2**, 196-217.

Lane, D. M. & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology*, **31**, 107-112.

Rosenthal, R. (1984). *Applied Social Research Methods Series, Vol. 6. Meta-analytic procedures for social research*. Newbury Park, CA: Sage Publications.

Schimmack, U. (in press). The ironic effect of significant results on the credibility of multiple study articles. *Psychological Methods*.

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.

Sterne, J. A., Gavaghan, D. & Egger, M. (2000). Publication and related bias in meta-analysis: Power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology*, **53**, 1119-29.

Strube, M. J. (2006). SNOOP: A program for demonstrating the consequences of premature and repeated null hypothesis testing. *Behavior Research Methods*, **38**, 24-27.

Table 1: Statistical properties of the Doerrfeld *et al.* (2012) experiments. Effect sizes were computed from the reported test statistics.

	N1	N2	Effect size	Power from pooled ES
Study 1	21	22	0.612	0.791
Study 2a	8	8	1.157	0.364
Study 2b	10	10	0.918	0.449
Study 3	10	30	1.056	0.637

Figure 1: Effect sizes and 95% confidence intervals for the key findings in Doerrfeld *et al.* (2012).

