Too Much Success for Recent Groundbreaking Epigenetic Experiments

Gregory Francis
Department of Psychological Sciences, Purdue University, West Lafayette, Indiana 47906, and Brain Mind Institute, École Polytechnique Fédérale de Lausanne, Lausanne 1015, Switzerland
ORCID ID: 0000-0002-8634-794X (G.F.)

ABSTRACT An article reporting statistical evidence for epigenetic transfer of learned behavior has important implications, if true. With random sampling, real effects do not always result in rejection of the null hypothesis, but the reported experiments were uniformly successful. Such an outcome is expected to occur with a probability of 0.004.

INDEPENDENT replications of empirical findings are critical for the development of science (*e.g.*, Prinz *et al.* 2011; Collins and Tabak 2014; McNutt 2014), but there are difficulties in interpreting replications of *statistical* findings. Due to random sampling, not every experiment will produce a successful statistical outcome, even if an effect actually exists. If the statistical power of a set of experiments is relatively low, then the absence of unsuccessful results implies that something is amiss with data collection, data analysis, or reporting (Ioannidis and Trikalinos 2007; Francis 2012, 2013, 2014). Here, I apply these ideas to a recent study reporting epigenetic transfer of olfactory conditioning (Dias and Ressler 2014) that has been hailed as both groundbreaking and puzzling (Hughes 2014; Szyf 2014; Welberg 2014).

The claim for epigenetic transfer is based on behavioral and neuroanatomical findings. The first experiment (coded as "Figure 1a" in Table 1) is representative of the behavioral studies. One group of male mice was subjected to fear conditioning in the presence of the odor acetophenone. Compared to the offspring of unconditioned control mice, the offspring of the conditioned mice exhibited significantly enhanced sensitivity to acetophenone as measured by the fear-potentiated startle ($P = 0.043$). A *post hoc* power calculation suggests that a replication experiment using the same sample sizes is estimated to produce a statistically significant outcome ($P < 0.05$) only 51% of the time if the effect is similar to what was reported in the original experiment. Nine other behavioral experiments explored variations of the finding (using different odors, generations, mouse strains, and developmental contexts). As defined by Dias and Ressler (2014), success in those experiments usually involved rejecting the null hypothesis, but for some experiments success was based on a predicted null result or a pattern of significant and nonsignificant results. I estimated success probabilities for experiments like these with standard power calculations or simulated experiments that used the reported sample sizes, means, and standard deviations. For all of these calculations, the hypothesis tests of the original findings were assumed to be appropriate and valid for the data (*e.g.*, the data were sampled from populations having normal distributions with homogeneity of variance). R scripts for estimating the probabilities are provided with this article's supplemental material.

Table 1 lists the sample sizes, the inferences that defined success, and the estimated probability of such outcomes for each experiment. I followed Dias and Ressler (2014)'s treatment of the experiments as being statistically independent, so the probability of a set of 10 behavioral experiments like these all succeeding is the product of the probabilities: 0.023. This value is an estimate of the reproducibility of the statistical outcomes for these behavioral studies. Its low value suggests that the outcomes deemed by Dias and Ressler (2014) as support for their claim are unlikely with experiments similar to the ones they reported. It is important to recognize that such a low probability is not a necessary outcome for all possible experiment sets. When a reported experiment set includes unsuccessful results (as it should if the probabilities are modest), the excess success analysis estimates the probability of producing the observed or a greater

doi: 10.1534/genetics.114.163998
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.163998/-/DC1.
Address for correspondence: Department of Psychological Sciences, 703 Third St., Purdue University, West Lafayette, IN 47906. E-mail: gfrancis@purdue.edu

Table 1 Probability of success for experiments like those in Dias and Ressler (2014)

Experiment	Type	Sample sizes	Reported inference	Probability of success
Figure 1a	Behavior	16, 13	$\mu_1 \neq \mu_2$	0.512
Figure 1b	Behavior	7, 9	$\mu_1 = \mu_2$	0.908
Figure 1c	Behavior	11, 13, 19	ANOVA, $\mu_1 \neq \mu_2$, $\mu_2 \neq \mu_3$, $\mu_1 \geq \mu_3$	0.662
Figure 1d	Behavior	10, 11, 8	ANOVA, $\mu_1 = \mu_2$, $\mu_2 \neq \mu_3$	0.712
Figure 2a	Behavior	16, 16	$\mu_1 \neq \mu_2$	0.663
Figure 2b	Behavior	16, 16	$\mu_1 \neq \mu_2$	0.928
Figure 3g	Neuroanatomy	38, 38, 18	ANOVA, $\mu_1 \neq \mu_2$, $\mu_2 \neq \mu_3$	0.782
Figure 3h	Neuroanatomy	31, 40, 16	ANOVA, $\mu_1 \neq \mu_2$, $\mu_2 \neq \mu_3$	≈1.00
Figure 3i	Neuroanatomy	6, 6, 4	ANOVA, $\mu_1 \neq \mu_2$, $\mu_2 \neq \mu_3$	0.998
Figure 4a	Behavior	8, 12	$\mu_1 \neq \mu_2$	0.675
Figure 4b	Behavior	8, 11	$\mu_1 \neq \mu_2$	0.545
Figure 4g	Neuroanatomy	7, 8	$\mu_1 \neq \mu_2$	0.999
Figure 4h	Neuroanatomy	6, 10	$\mu_1 \neq \mu_2$	0.974
Figure 4i	Neuroanatomy	23, 16	$\mu_1 \neq \mu_2$	0.973
Figure 4j	Neuroanatomy	16, 19	$\mu_1 \neq \mu_2$	≈1.00
Figure 5a	Behavior	13, 16	$\mu_1 \neq \mu_2$	0.600
Figure 5b	Behavior	4, 7, 6, 5	ANOVA, $\mu_1 \neq \mu_2$, $\mu_3 \neq \mu_4$	0.775
Figure 5g	Neuroanatomy	6, 4, 5, 3	ANOVA, $\mu_1 \neq \mu_2$, $\mu_3 \neq \mu_4$, $\mu_1 = \mu_3$	0.892
Figure 5h	Neuroanatomy	4, 3, 8, 4	ANOVA, $\mu_3 \neq \mu_4$, $\mu_1 = \mu_3$	0.824
Figure 6a	Neuroanatomy	12, 10	$\mu_1 \neq \mu_2$	0.574
Figure 6c	Neuroanatomy	12, 10	$\mu_1 = \mu_2$	0.901
Figure 6e	Neuroanatomy	8, 8	$\mu_1 \neq \mu_2$	0.681

The reported inferences were those used by Dias and Ressler (2014) to support their theoretical claims. The probability of success for such inferences is estimated by *post hoc* power calculations or simulated experiments. Experiments are labeled according to the data figures in Dias and Ressler 2014.

number of successful outcomes. For example, if 3 of the 10 behavioral experiments reported in Dias and Ressler (2014) had been unsuccessful, then the probability of producing seven or more successful outcomes would be estimated as 0.65, which would not raise any concerns. R code for the calculation is provided in the Supporting Information, File S1 with this article.

Dias and Ressler (2014)'s argument for epigenetic transfer of conditioning was bolstered by 12 neuroanatomical experiments, with the first one (marked as "Figure 3g" in Table 1) being representative. Staining indicated that the offspring of mice fear conditioned with acetophenone had larger acetophenone-responding glomeruli in the olfactory bulb compared to both the offspring of mice without conditioning and to the offspring of mice conditioned to a different odor. Experimental success required a significant ANOVA and a significant contrast between the experimental group and each of the control groups. The probability of a successful outcome (estimated by simulated experiments as 0.782) differs from the ideal value of one because the test between mice conditioned to different odors has only modest experimental power due to the relatively small sample size for one of the groups ($n = 18$). Other neuroanatomical studies compared staining of odor-responding glomeruli in different brain regions and in different mouse strains, generations, and developmental contexts. Similar to the behavioral studies, every reported experiment produced a pattern of significant and nonsignificant findings deemed to provide support for the theoretical claims. The probability of experiments like these being so successful is the product of the appropriate probabilities listed in Table 1, which is 0.189. Although better than for the behavioral experiments, this analysis indicates only a one in five chance of successfully replicating the full set of neuroanatomical findings reported in Dias and Ressler (2014) with effects and sample sizes similar to the original report.

The claim that olfactory conditioning could epigenetically transfer to offspring is based on successful findings from both the behavioral and neuroanatomical studies. If that claim was correct, if the effects were accurately estimated by the reported experiments, and if the experiments were run properly and reported fully, then the probability of every test in a set of experiments like these being successful is the product of all the probabilities in Table 1, which is 0.004. The estimated reproducibility of the reported results is so low that we should doubt the validity of the conclusions derived from the reported experiments.

How could the findings of Dias and Ressler (2014) have been so positive with such low odds of success? Perhaps there were unreported experiments that did not agree with the theoretical claims; perhaps the experiments were run in a way that improperly inflated the success and type I error rates, which would render the statistical inferences invalid. Researchers can unintentionally introduce these problems with seemingly minor choices in data collection, data analysis, and result interpretation. Regardless of the reasons, too much success undermines reader confidence that the experimental results represent reality.

Even if some of the effects prove to be real, the findings reported in Dias and Ressler (2014) likely overestimate the effect magnitudes because unreported unsuccessful outcomes usually indicate a smaller effect than reported successful outcomes. Scientists planning to design experiments that replicate the significant behavioral findings in Dias and Ressler (2014) might find it prudent to halve the pooled effect size value from 1.0 to 0.5. To show statistical

significance with a power of 0.8 for a difference of means, such a replication experiment requires sample sizes of 64 in each group, which is four times the size of the largest experimental samples used by Dias and Ressler (2014). Importantly, even for such high power experiments, one would not expect all studies to produce successful outcomes. For proper experiments, the rate of experimental success has to match the characteristics of the experiments, effects, and analyses. Scientific claims based on hypothesis tests from a set of experiments require either highly powered successful experiments or pooling across both successful and unsuccessful experiments.

Note added in proof: See Dias and Ressler 2014 (pp. 453) and Churchill 2014 (pp. 447–448) in this issue for a related work.

Literature Cited

Churchill, G. A., 2014 When are results too good to be true? Genetics 198: 447–448.

Collins, F., and L. A. Tabak, 2014 NIH plans to enhance reproducibility. Nature 505: 612–613.

Dias, B. G., and K. J. Ressler, 2014 Reply to Gregory Francis. Genetics 198: 453.

Dias, B. G., and K. J. Ressler, 2014 Parental olfactory experience influences behavior and neural structure in subsequent generations. Nat. Neurosci. 17: 89–96.

Francis, G., 2012 Too good to be true: publication bias in two prominent studies from experimental psychology. Psychon. Bull. Rev. 19: 151–156.

Francis, G., 2013 Replication, statistical consistency, and publication bias. J. Math. Psychol. 57: 153–169.

Francis, G., 2014 The frequency of excess success for articles in *Psychological Science*. Psychon. Bull. Rev. http://link.springer.com/article/10.3758/s13423–014–0601-x

Hughes, V., 2014 Epigenetics: the sins of the father. Nature 507: 22–24.

Ioannidis, J. P. A., and T. A. Trikalinos, 2007 An exploratory test for an excess of significant findings. Clin. Trials 4: 245–253.

McNutt, M., 2014 Reproducibility. Science 343: 229.

Prinz, F., T. Schlange, and K. Asadullah, 2011 Believe it or not: How much can we rely on published data on potential drug targets? Nat. Rev. Drug Discov. 10: 712–713.

Szyf, M., 2014 Lamarck revisited: epigenetic inheritance of ancestral odor fear conditioning. Nat. Neurosci. 17: 2–4.

Welberg, L., 2014 Epigenetics: a lingering smell? Nat. Rev. Neurosci. 15: 1.

Communicating editor: M. Johnston

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.163998-/DC1

Too Much Success for Recent Groundbreaking Epigenetic Experiments

Gregory Francis

DOI: 10.1534/genetics.114.163998

File S1

R code for estimating probabilities

Available for download as a .zip file at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.163998/-/DC1