When Are Results Too Good to Be True?

Gary A. Churchill
The Jackson Laboratory, Bar Harbor, Maine 04609

SCIENCE is in crisis. Many published studies appear to be irreproducible (Prinz *et al.* 2011; Begley and Ellis 2012). What can be done? How concerned should we be?

Is the reproducibility crisis something new? Suppose we randomly sampled studies from the past 100 years of scientific literature and attempted to replicate their findings. I expect that many of these studies would fail to replicate. We must be careful not to conflate "irreproducible" with "false." The experiments may require specific conditions that are difficult to reproduce; the original studies may have been underpowered; or they may have addressed a hypothesis that turned out to be false. "False" studies are part and parcel of the scientific method in which falsifiable hypotheses are repeatedly put to the test. Many potentially transformative studies have been published and later discredited (*e.g.*, Fleischmann and Pons 1989); others have stood the test of time and have become integrated into the fabric of knowledge (*e.g.*, Luria and Delbruck 1943).

Something has changed, however: the industry of science has grown exponentially, at a rate of ~4% per year as measured in the number of articles published (Larsen and Von Ins 2010). At this rate of growth, the doubling time is ~15 years, which means that very soon more scientific publications will have appeared in the 21st century than in all of prior history. This growth has resulted in fierce competition for coveted spots in high-profile journals. Being good, or even very good, is no longer good enough. One also needs to be lucky.

Nature magazine receives ~10,000 submissions annually and (necessarily) rejects >90% of them. Suppose that all of the 10,000 studies submitted to *Nature* in a year had evaluated false scientific hypotheses. The *p*-values reported would be uniformly distributed between zero and one[1]. Some of these would be significant by chance. If the *Nature* editors evaluated these studies based solely on the reported *p*-values, the probability that at least one of these lucky papers would report "$p < 0.0001$" is 67%. Of course, it takes more than an impressive *p*-value to be published in *Nature*. A truly outstanding article will have something novel, even surprising, to say.

To understand the impact of the novelty criterion, consider the following three experiments (see Greenhouse 2012). In the first experiment, a music expert claims that she can distinguish a score written by Mozart from one written by Haydn. Presented with 10 scores in a double-blinded and randomized order, she identifies each one correctly. In the second experiment, a tea-drinking lady claims that she can tell if the milk or the tea was poured first into the cup. Given 10 carefully prepared cups of tea in randomized order, she correctly identifies each one. Finally, an inebriated customer at a bar claims he can predict the outcome of a coin toss. He proceeds to toss the coin and calls heads or tails correctly 10 times in a row.

The *p*-value in each case is 0.001 ($1/2^{10}$). But what conclusions are we to draw? Here we cannot avoid our subjective opinions about the prior plausibility of each claim. In my opinion, the music experiment was not really necessary because the claim is believable *a priori*. In the case of the tea-drinking lady, while I may have initially doubted this unusual talent, the evidence is convincing and I would consider the matter settled. As for the drunken coin tosser, after carefully examining the coin, I would ask him to do it again.[2] This claim is just too hard to believe; a higher standard of evidence seems justified.

Which of these studies, if properly vetted by review, would make an exciting paper? My vote is with "Alcohol induces clairvoyance." Experiments based on hypotheses that are *a priori* implausible can be potentially groundbreaking, but are also the most likely to be false (Ioannidis 2005). Can we determine which are which?

The shortcomings of *p*-values as a measure of evidence are well documented (Berger and Sellke 1987) and continue to inspire debate (Nuzzo 2014). Imagine, then, a statistic

[1]By definition, under the null hypothesis a *p*-value is randomly and uniformly distributed between zero and one. But for this hypothetical exercise, you must suspend your disbelief about the audacity of the scientists who submitted their nonsignificant findings for publication.

doi: 10.1534/genetics.114.169912
Address for correspondence: The Churchill Group, The Jackson Laboratory, 600 Main Street Bar Harbor, Maine 04609. E-mail: garyc@jax.org

[2]Bear in mind that every day one-in-a-million events will happen to ~7000 people (Littlewood 1986).

that provides an optimal and objective measure of evidence based only on the experimental data at hand without recourse to prior beliefs and opinions. Let us call it the "o-value." Using the o-value as our criterion for publication, we would still publish false conclusions. Otherwise, we would have to be so stringent in our evaluation that an excessive number of true findings would be rejected or our requirements for evidence would be so prohibitive that progress would grind to a halt.

A paradoxical consequence of having access to an optimal measure of evidence is the impossibility of distinguishing a true claim from a false one. If that were possible, it would follow that there is additional information in the data, which contradicts our stipulation that the o-value is optimal. We could use that extra information to create an improved o-value, but we would end up in the same conundrum. Even with an ideal measure of evidence, it is impossible to objectively establish truth or falsehood of a claim based solely on the available data. This is a troubling reality.

The Case of Dias and Ressler

A recent publication in *Nature Neuroscience* (Dias and Ressler 2013) put forward a provocative hypothesis that epigenetic inheritance can be modified by olfaction. The paper understandably drew a great deal of attention. In this issue of *GENETICS*, we publish a critique by G. Francis who argues that the evidence presented is "too good to be true." Francis claims that the study could not plausibly have been as successful as reported and that some of the experiments reported should have failed to reject the null hypothesis. These concerns are reminiscent of Fisher's claim that Mendel cooked his data (Hartl and Fairbanks 2007). While Mendel's data may remain the subject of debate, Mendel's laws—appropriately modified to conform to subsequent observations—have stood the test of time.

The *post hoc* power analysis provided by Francis is enlightening, but the evidence required to evaluate the study is entirely contained in the reported *p*-values (Hoenig and Heisy 2001). If the same analysis were applied to a randomly selected study, it might cast doubt on the integrity of the study. But the Dias and Ressler study was selected from among thousands of studies competing for limited publication space by an evaluation process that selects papers with inflated *p*-values. "Extraordinary claims require extraordinary evidence."[3] Skeptical reviewers are going to balk at any sign of weakness in a controversial manuscript. And yet, extraordinary evidence can occur by chance. This suggests that improbable *p*-values are to be expected in controversial, high-profile papers.

Opinions play an important role in deciding what gets published. As illustrated in our hypothetical example of three experiments that produce identical statistical evidence, subjectivity plays a crucial role in our interpretation of that evidence. The proposal that epigenetic inheritance can be modified by olfaction stretches the imagination, but it is not outside the realm of possibility. Expert scientists vetted the experimental procedures, and we must assume that the data reported are accurate and complete. But is the claim true? Without further study, it is a matter of opinion.

Progress in science requires an influx of new ideas balanced by skepticism that compels us to re-examine the evidence. When we seek more evidence to corroborate or refute hypotheses, some will prove to be wrong. We should not subvert this process by reaching for an unattainable ideal of perfectly reproducible studies. Ironically, statistical evidence presented in the original study may be of little help in determining which hypotheses will hold up. Dias and Ressler have proposed an intriguing hypothesis, and they have reported their evidence to support it. Of course, the study should be repeated—perhaps by using different approaches and methods that address the same hypothesis from different angles. The findings of Dias and Ressler warrant further study, and the scientific method compels us to try to topple this hypothesis. Is it true or too good to be true? Only time—and further investigation—will tell.

Note added in proof: See Dias and Ressler 2014 (pp. 453) and Francis 2014 (pp. 449–451) in this issue for a related work.

Literature Cited

Begley, C. G., and L. M. Ellis, 2012 Drug development: raise standards for preclinical cancer research. Nature 483(7391): 531–533.

Berger, J. O., and T. Sellke, 1987 "Testing a point null hypothesis: the irreconcilability of *p* values and evidence" (with discussion). J. Am. Stat. Assoc. 82: 112–139.

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.

Fleischmann, M., and Pons, S. (1989). Electrochemically induced nuclear fusion of deuterium. J. Electroanal. Chem. 261(2, Part 1): 301–308.

Francis, G., 2014 Too much success for recent groundbreaking epigenetic experiments. Genetics 198: 449–451.

Greenhouse, J. B., 2012 On becoming a Bayesian: early correspondences between J. Cornfield and L. J. Savage. Stat. Med. 31: 2782–2790.

Hartl, D. L., and D. J. Fairbanks, 2007 Mud sticks: on the alleged falsification of Mendel's data. Genetics 175(3): 975–979.

Hoenig, J. M., and D. M. Heisy, 2001 The abuse of power: the pervasive fallacy of power calculations for data analysis. Am. Stat. 55(1): 19–24.

Ioannidis, J. P. A., 2005 Why most published research findings are false. PLoS Med. 2(8): e124.

Larsen, P. O., and M. von Ins, 2010 The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. Scientometrics 84: 575–603.

Littlewood, J. E., 1986 *Littlewood's Miscellany.* Cambridge University Press, Cambridge, UK.

Luria, S. E., and M. Delbrück, 1943 Mutations of bacteria from virus sensitivity to virus resistance. Genetics 28(6): 491–511.

Nuzzo, R., 2014 Scientific method: statistical errors. *Nature* 506: 150–152.

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.

[3]Carl Sagan, from the TV series *Cosmos*.

Communicating editor: M. Johnston

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

Reply to Gregory Francis

Brian G. Dias[1,2] and Kerry J. Ressler[1,2,3]
[1]Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, Atlanta, GA 30329
[2]Center for Behavioral Neuroscience, Yerkes National Primate Research Center, Atlanta, GA 30329
[3]Howard Hughes Medical Institute, Chevy Chase, MD 20815

We thank Gregory Francis for his careful reading of our article and associated commentary. We too remain amazed at this phenomenon, and we agree that much remains to be done to understand the underlying mechanisms. However, we have now replicated these effects multiple times within our laboratory with multiple colleagues as blinded scorers, and we fully stand by our initial observations. Here we focus on addressing some of the broader implications of his letter.

1. The principal assertion made by Dr. Francis for his analyses are that "the reported experiments were uniformly successful" in our reported studies. This is a false statement. While we wish that all our behavioral, neuroanatomical, and epigenetic data were successful and statistically significant, one only need look at the Supporting Information in the article to see that data generated for all four figures in the Supporting Information did not yield significant results. We do not believe that these nonsignificant data support our theoretical claims as is suggested. If that were the case, there ought to be have been correspondence between our DNA methylation data in the sperm (Figure 6), and Main Olfactory Epithelium (Figure S6). In addition, we do not observe differences among groups in levels of histone modifications around the M71 gene in sperm (Figure S7). These multiple cases of nonsignificant data were clearly reported by us within the primary paper and the supporting information. Therefore, we strongly reject the assertion that we only presented data that confirmed the hypotheses. We are actively searching for mechanisms that support these robust findings.
2. We opine that the real story and mechanism is likely to be found probing concepts like penetrance. For it is most likely that epigenetic mechanisms might not affect all the germ cells, and understanding the spread of data will give us a more nuanced view of the mechanism.
3. We wholeheartedly disagree with the shadow that Francis and the accompanying commentary casts on our experimental design and data analysis. All experiments conducted were reported in the article, which means that no experimental data were excluded. When one conducts transgenerational studies that are dependent on the vagaries of breeding and husbandry, one uses all that are given and is never wasteful. All experiments were run blind by the experimenter, and data were analyzed in a double-blind fashion as we have emphasized in the article. The one point of agreement between Francis and us is the need for higher sample sizes, and this is something that we would like to address in subsequent work. This said, it must be emphasized that our sample sizes are consistent with what animal behaviorists use.
4. It is also asserted by Dr. Francis that perhaps the manuscript did not undergo rigorous peer review. The manuscript in fact went through numerous rounds of rigorous peer-review at *Nature Neuroscience*, with at least 3 anonymous and critical external reviewing scientists, along with substantial editorial review. The review process resulted in multiple additional experiments and analyses, some positive and some negative, which we believe improved the manuscript.

In summary, while we appreciate some of Francis' concerns, we stand by our results as robust, reproducible, and verified by blinded assessment. We believe our findings withstand the test of 'extraordinary evidence'. Science is built on the synergy between findings from independent research groups. We wholeheartedly welcome this process, and are excited about the direction of this research.

doi: 10.1534/genetics.114.169904