Scientific "freedom" and the Fountain of Youth

October 19, 2011

“Chronological rejuvenation” is psychological jargon for the Fountain of Youth, that elusive tonic that, when we find it, will reverse the aging process. Though many of us would welcome such a discovery, most of us also know it’s a fantasy, a scientific impossibility.

So imagine my surprise when I came across this passage while browsing the journal *Psychological Science*. I include all the methodological detail because they are important:

Using the same method as in Study 1, we asked 20 University of Pennsylvania undergraduates to listen to either “When I’m Sixty-Four” by The Beatles or “Kalimba.” Then, in an ostensibly unrelated task, they indicated their birth date (mm/dd/yyyy) and their father’s age. We used father’s age to control for variation in baseline age across participants. An ANCOVA revealed the predicted effect: According to their birth dates, people were nearly a year-and-a-half younger after listening to “When I’m Sixty-Four” adjusted M = 20.1 years) rather than to “Kalimba” (adjusted M = 21.5 years), F(1, 17) = 4.92, p = .040.

If you’re like me, you had to reread the passage: It doesn’t say that subjects *felt* a year and a half younger; it says that they actually *were* a year and a half younger. That is, the data support the scientists’ unbelievable hypothesis that listening to music about old age can make us significantly younger.

Unbelievable indeed. The authors of the report—Joseph Simmons and Uri Simonsohn of Penn and Leif Nelson of Berkeley—certainly don’t believe this finding, although they did really run the experiment as described, and they used accepted practices for reporting and analyzing their data. They ran the experiment to demonstrate a serious flaw in the usual way that behavioral science data are collected and reported—a flaw that (they claim) allows scientists to prove that “anything” is a significant finding. The scientists are not charging malfeasance or malicious intent, but they do argue that current methods lead inevitably to self-serving intellectual dishonesty, producing a lot of “false positives”—results that appear statistically valid but in fact are not. Discovering the Fountain of Youth is a deliberately absurd example, but the authors believe that false positives in behavioral science are “vastly more likely” than the 5 percent that’s generally acknowledged.

How does this happen? The scientists ran a computer simulation of 15,000 actual data samples, and identify the culprit as “researcher degrees of freedom.” This simply means that, in the course of running and reporting an experiment, scientists make a number of decisions that can skew results. These decisions—and the rules governing them—make it “unacceptably easy” to publish bogus evidence for literally any hypothesis. One example is the decision about the sample size. Typically, a researcher will recruit a sample—say 20 subjects—and run the experiment. At that point, the research tests the data for significance. If the result is significant, terrific: The researcher stops collecting data and reports the result. So far so good, but what if the result does not reach significance after 20 subjects? In that case, the researcher has the option of adding another group of subjects—say 10 more—and testing the data

again. And again, and again. According to the authors' computer simulation, this seemingly small degree of freedom increases the false-positive rate by 50 percent. A recent survey found that about seven in ten psychological scientists admit to making such interim decisions in their research.

Manipulating sample size is just one of the decision-making "freedoms" that the authors highlight in their *Psychological Science* paper. Another is flexibility in including (and reporting) dependent variables. If a researcher designs an experiment with two dependent variables, he or she can decide to test just one, or the other, or both, increasing the likelihood of producing at least one significant result. Similarly, flexibility in controlling for gender can dramatically boost false positives, as can dropping (or not dropping) one of three experimental conditions. What's more, the authors note, scientists often use all these freedoms in the same experiment, a practice that would lead to a stunning 61 percent false positive rate. In other words, a researcher with all the best intentions is *more likely than not* to falsely detect a positive result just by using the accepted practices in the field.

They consider this estimate conservative. Researchers often test and choose among more than two dependent variables; exclude subsets of subjects as "outliers"; consign early data to "pilot studies"; and so on.

So what's to be done? The authors offer a simple solution to the problem of false-positive publication, including requirements for researchers and guidelines for journal reviewers. The requirements all aim for more transparency in reporting data and methods of analysis—listing all variables, for example, and deciding on sample size before data collection begins. Here, for illustration, is the way the authors would revise their own report on the Fountain of Youth study:

Using the same method as in Study 1, we asked 34 University of Pennsylvania undergraduates to listen only to either "When I'm Sixty-Four" by The Beatles or "Kalimba" or "Hot Potato" by the Wiggles. We conducted our analyses after every session of approximately 10 participants; we did not decide in advance when to terminate data collection. Then, in an ostensibly unrelated task, they indicated only their birth date (mm/dd/yyyy) and how old they felt, how much they would enjoy eating at a diner, the square root of 100, their agreement with "computers are complicated machines," their father's age, their mother's age, whether they would take advantage of an early-bird special, their political orientation, which of four Canadian quarterbacks they believed won an award, how often they refer to the past as "the good old days," and their gender. We used father's age to control for variation in baseline age across participants. An ANCOVA revealed the predicted effect: According to their birth dates, people were nearly a year-and-a-half younger after listening to "When I'm Sixty-Four" (adjusted M = 20.1 years) rather than to "Kalimba" (adjusted M = 21.5 years), $F(1, 17) = 4.92$, $p = .040$. Without controlling for father's age, the age difference was smaller and did not reach significance (Ms = 20.3 and 21.2, respectively), $F(1, 18) = 1.01$, $p = .33$.

This detailed report is about twice as long as the one above, and will no doubt seem burdensome to many scientists. It also will not stop blatant cheaters, the authors concede, but that is not the point. The goal, they conclude is to reduce the "self-serving interpretation of ambiguity, which enables us to convince ourselves that whichever decisions produced the most publishable outcome must have also been the most appropriate."

Wray Herbert's book, *On Second Thought*, has recently been published in paperback. Excerpts from his

two blogs—“We’re Only Human” and “Full Frontal Psychology”—appear regularly in *Scientific American Mind* and in *The Huffington Post*.