STAT8801, April 17, 2017

Statistics and the Reproducibility Crisis

In May 2016, Nature reported that “more than 70% of researchers have tried and failed to reproduce another scientist’s experiements, and more than half have failed to reproduce their own experiments. ... 52% of those surveyed agreed that there is a significant ‘crisis’ of reproducibility.”

What have you heard about this ‘crisis’? What are some possible reasons for it? In particular, what might be some statistical reasons?

The National Academy of Sciences recently released a new report on ”Fostering Integrity in Research,” which “shines a spotlight on how the research enterprise as a whole creates incentives that can be detrimental to good research” (MPR news).

That is, it’s not just outright fraud! A paper “Scientists behaving badly” surveyed researchers about these behaviors. What percent would you guess did these things?

- falsifying data
- not disclosing involvement with companies who would benefit
- failing to present data that contradict own research
- changing design, methods, or results, in response to pressure from a funding source
- publishing same data or results in two publications
- withholding details of methodology
- using inadequete or inappropriate research designs
- dropping data based on a gut feeling that they were inaccurate

In the Nature study, “More than 60% said each of two factors – pressure to publish and selective reporting – always or often contributed.” We’ll talk mostly about selective reporting today.

Here’s recommendation 8 from the NAS report: “To avoid unproductive duplication of research and to permit effective judgments on the statistical significance of findings, researchers should routinely disclose all statistical tests carried out, including negative findings. Research sponsors, research institutions, and journals should support and encourage this level of transparency.”

In the Nature paper, “Nearly 90% ... ticked ‘More robust experimental design,’ ’better statistics,’ and ’better mentorship’ ” as ways to improve reproducibility.

Here's a list of recommendations from an article "False-Positive Psychology." They say, "It is rare, and sometimes impractical, for researchers to make all these [analysis] decisions beforehand. Rather, it is common (and accepted practice) for researchers to explore various analytic alternatives, to search for a combination that yields 'statistical significance,' and to then report only what 'worked.'" They call these potential choices "researcher degrees of freedom."

1. Authors must decide the rule for terminating data collection before data collection begins and report this rule in the article.
2. Authors must collect at least 20 observations per cell or else provide a compelling cost-of-data-collection justification.
3. Authors must list all variables collected in a study.
4. Authors must report all experimental conditions, including failed manipulations.
5. If observations are eliminated, authors must also report what the statistical results are if those observations are included.
6. If an analysis includes a covariate, authors must report the statistical results of the analysis without the covariate.

Why is your recommendation important? What could happen if you don't do it?

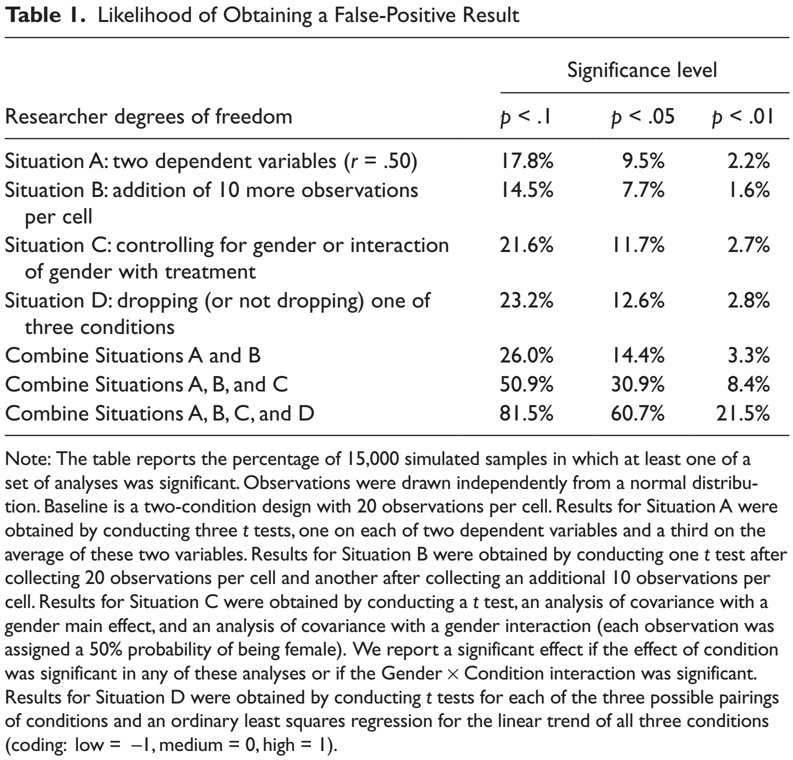
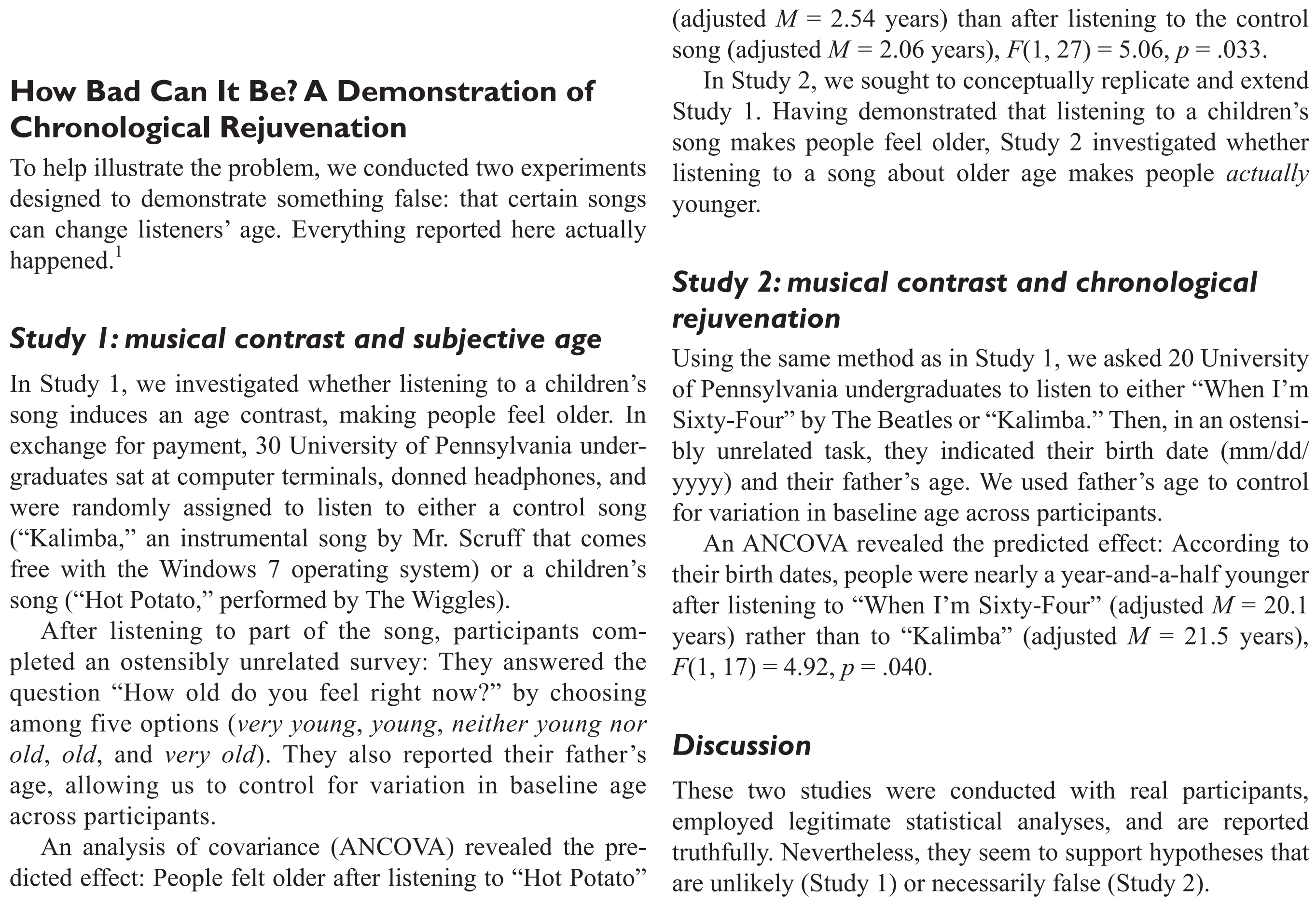


Table 1. Likelihood of Obtaining a False-Positive Result

	Significance level		
Researcher degrees of freedom	$p < .1$	$p < .05$	$p < .01$
Situation A: two dependent variables ($r = .50$)	17.8%	9.5%	2.2%
Situation B: addition of 10 more observations per cell	14.5%	7.7%	1.6%
Situation C: controlling for gender or interaction of gender with treatment	21.6%	11.7%	2.7%
Situation D: dropping (or not dropping) one of three conditions	23.2%	12.6%	2.8%
Combine Situations A and B	26.0%	14.4%	3.3%
Combine Situations A, B, and C	50.9%	30.9%	8.4%
Combine Situations A, B, C, and D	81.5%	60.7%	21.5%

Note: The table reports the percentage of 15,000 simulated samples in which at least one of a set of analyses was significant. Observations were drawn independently from a normal distribution. Baseline is a two-condition design with 20 observations per cell. Results for Situation A were obtained by conducting three t tests, one on each of two dependent variables and a third on the average of these two variables. Results for Situation B were obtained by conducting one t test after collecting 20 observations per cell and another after collecting an additional 10 observations per cell. Results for Situation C were obtained by conducting a t test, an analysis of covariance with a gender main effect, and an analysis of covariance with a gender interaction (each observation was assigned a 50% probability of being female). We report a significant effect if the effect of condition was significant in any of these analyses or if the Gender × Condition interaction was significant. Results for Situation D were obtained by conducting t tests for each of the three possible pairings of conditions and an ordinary least squares regression for the linear trend of all three conditions (coding: low = -1, medium = 0, high = 1).

This example is from that same paper.



How Bad Can It Be? A Demonstration of Chronological Rejuvenation

To help illustrate the problem, we conducted two experiments designed to demonstrate something false: that certain songs can change listeners' age. Everything reported here actually happened.[1]

Study 1: musical contrast and subjective age

In Study 1, we investigated whether listening to a children's song induces an age contrast, making people feel older. In exchange for payment, 30 University of Pennsylvania undergraduates sat at computer terminals, donned headphones, and were randomly assigned to listen to either a control song ("Kalimba," an instrumental song by Mr. Scruff that comes free with the Windows 7 operating system) or a children's song ("Hot Potato," performed by The Wiggles).

After listening to part of the song, participants completed an ostensibly unrelated survey: They answered the question "How old do you feel right now?" by choosing among five options (*very young*, *young*, *neither young nor old*, *old*, and *very old*). They also reported their father's age, allowing us to control for variation in baseline age across participants.

An analysis of covariance (ANCOVA) revealed the predicted effect: People felt older after listening to "Hot Potato" (adjusted M = 2.54 years) than after listening to the control song (adjusted M = 2.06 years), $F(1, 27) = 5.06$, $p = .033$.

In Study 2, we sought to conceptually replicate and extend Study 1. Having demonstrated that listening to a children's song makes people feel older, Study 2 investigated whether listening to a song about older age makes people *actually* younger.

Study 2: musical contrast and chronological rejuvenation

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), $F(1, 17) = 4.92$, $p = .040$.

Discussion

These two studies were conducted with real participants, employed legitimate statistical analyses, and are reported truthfully. Nevertheless, they seem to support hypotheses that are unlikely (Study 1) or necessarily false (Study 2).

Spoiler alert: The researcher chose their analysis to get this result.
How do you think they did it? What did they not report?

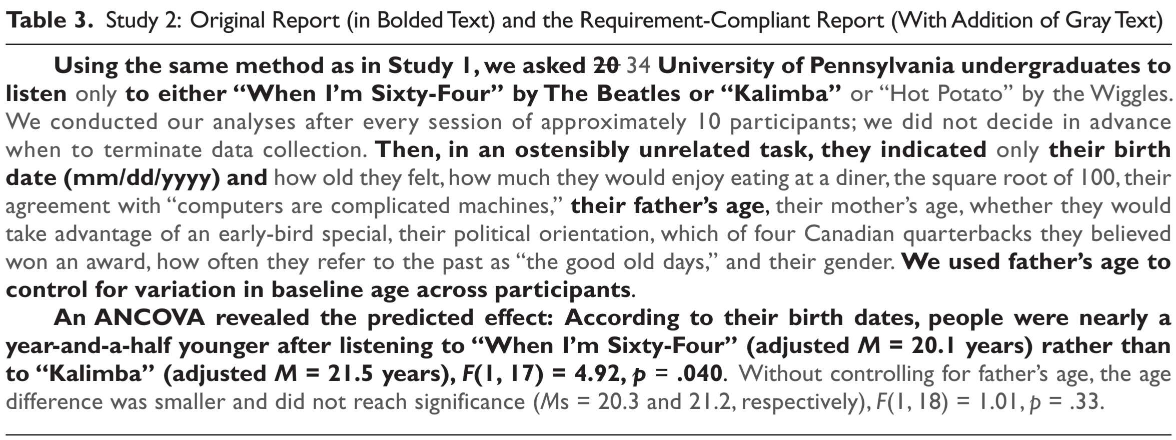


Table 3. Study 2: Original Report (in Bolded Text) and the Requirement-Compliant Report (With Addition of Gray Text)

Using the same method as in Study 1, we asked ~~20~~ 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 (*M*s = 20.3 and 21.2, respectively), *F*(1, 18) = 1.01, *p* = .33.

The reported probabibilities from "Scientists Behaving Badly":

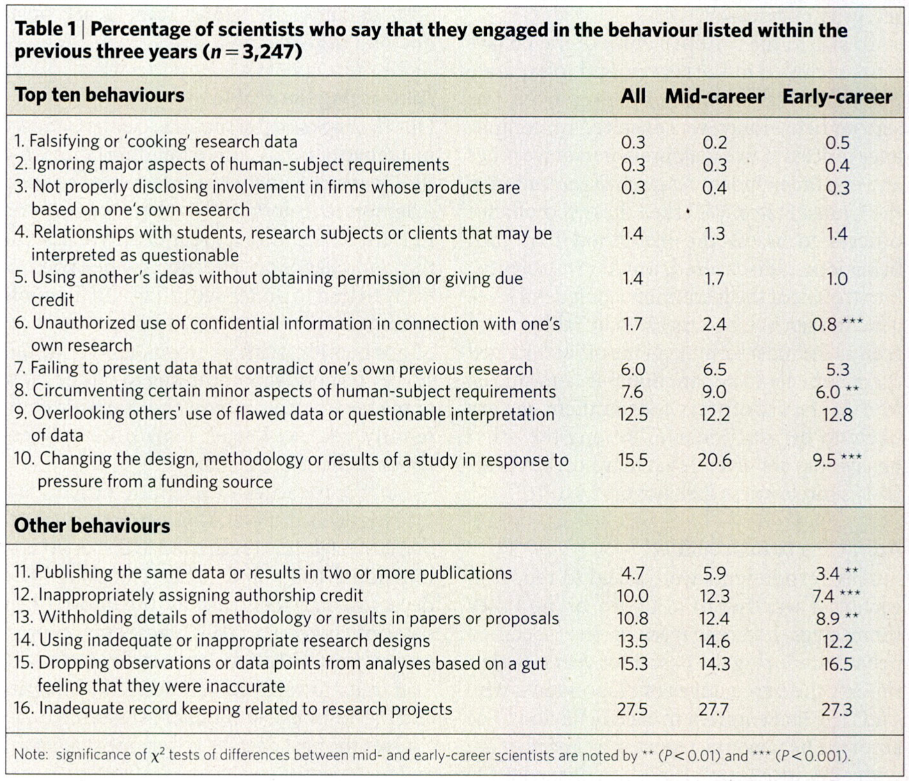


Table 1 | Percentage of scientists who say that they engaged in the behaviour listed within the previous three years ($n = 3{,}247$)

Top ten behaviours	**All**	**Mid-career**	**Early-career**
1. Falsifying or 'cooking' research data	0.3	0.2	0.5
2. Ignoring major aspects of human-subject requirements	0.3	0.3	0.4
3. Not properly disclosing involvement in firms whose products are based on one's own research	0.3	0.4	0.3
4. Relationships with students, research subjects or clients that may be interpreted as questionable	1.4	1.3	1.4
5. Using another's ideas without obtaining permission or giving due credit	1.4	1.7	1.0
6. Unauthorized use of confidential information in connection with one's own research	1.7	2.4	0.8 ***
7. Failing to present data that contradict one's own previous research	6.0	6.5	5.3
8. Circumventing certain minor aspects of human-subject requirements	7.6	9.0	6.0 **
9. Overlooking others' use of flawed data or questionable interpretation of data	12.5	12.2	12.8
10. Changing the design, methodology or results of a study in response to pressure from a funding source	15.5	20.6	9.5 ***
Other behaviours			
11. Publishing the same data or results in two or more publications	4.7	5.9	3.4 **
12. Inappropriately assigning authorship credit	10.0	12.3	7.4 ***
13. Withholding details of methodology or results in papers or proposals	10.8	12.4	8.9 **
14. Using inadequate or inappropriate research designs	13.5	14.6	12.2
15. Dropping observations or data points from analyses based on a gut feeling that they were inaccurate	15.3	14.3	16.5
16. Inadequate record keeping related to research projects	27.5	27.7	27.3

Note: significance of χ^2 tests of differences between mid- and early-career scientists are noted by ** ($P < 0.01$) and *** ($P < 0.001$).

Fostering Integrity in Research. National Academies of Sciences, Engineering, and Medicine. 2017. https://doi.org/10.17226/21896

Top Scientists Revamp Standards To Foster Integrity In Research. Richard Harris, MPR News Morning Edition, 11 Apr 2017. http://www.npr.org/sections/health-shots/2017/04/11/523406710/top-scientists-revamp-s

1,500 scientists lift the lid on reproducibility: Survey sheds light on the 'crisis' rocking research. Monya Baker. Nature 533, 452–454 (26 May 2016) http://dx.doi.org/10.1038/533452a

Scientists behaving badly. Brian C. Martinson, Melissa S. Anderson, Raymond de Vries, Nature 435, 737–738 (9 June 2005). http://dx.doi.org/10.1038/435737a

False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. Joseph P. Simmons, Leif D. Nelson, Uri Simonsohn, Psychological Science 22(11) 1359–1366 (17 Oct 2011). http://dx.doi.org/10.1177/0956797611417632

STAT8801, April 17, 2017

Name: ______________________________

What did you learn today about the reproducibility crisis that was new or interesting?

STAT8801, April 17, 2017

Name: ______________________________

What did you learn today about the reproducibility crisis that was new or interesting?