Understanding p-hacking and HARKing

S. Stanley Young,[1] Warren B. Kindzierski,[2] Terry Meyer[3]

[1]CGStat, Raleigh, NC 27607, USA; genetree@bellsouth.net
[2]Independent Consultant, St Albert, Alberta T8N5R1, Canada; wbk@shaw.ca
[3]Outcome Based Medicine, Raleigh, NC 27614, USA; terry_g_meyer@yahoo.com

To whom correspondence should be addressed: Warren B. Kindzierski, wbk@shaw.ca

Abstract: 186 words
Main text: 2,658 words
Illustrations: 5 figures
References: 23

Running title: Understanding p-hacking and HARKing

Keywords: questionable research practices, p-hacking, HARKing, false-positive

Abstract

Many claims made by researchers often fail to replicate when tested rigorously. In short, these claims can be wrong. There is a need for students (and citizens) to understand how incorrect claims can come about in research by using questionable research practices. Ask a lot of questions of a data set and make a claim if any p-value is less than 0.05 is p-hacking. HARKing is making up a claim/narrative *after* looking at the data set. Statistical experts know about p-hacking and HARKing, but they appear to be largely silent. Some researchers know too but they ignore the problem. We present a hands-on demo about rolling ten-sided dice multiple times to show how incorrect claims come about. Several individuals executed simulations of p-values with ten-sided dice and show how easily a small p-value can come about by chance for a modest number of questions (rolls of the dice). Notably, small p-values found by one individual were not replicated by other individuals. These simple simulations allow students (and citizens) to better judge the reliability of a science claim when multiple questions are asked of a data set.

Introduction

It has been known for some time that research claims most often fail to replicate. For example, published estimates of studies or reports that fail to replicate in the biomedical field range from 51 to 100%:

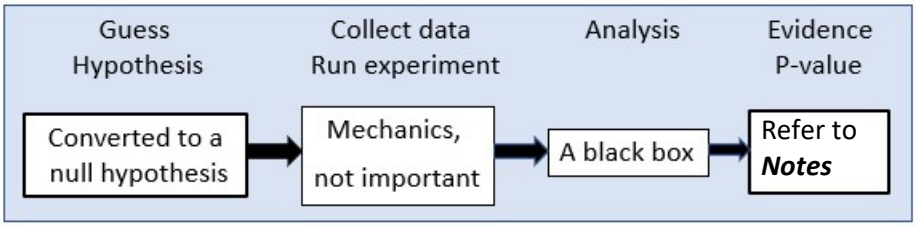
- 41 of 80 studies (51%) examined in the primary care and general medicine field (Glasziou et al. 2008)
- 131 of 257 studies (51%) examined in the clinical psychology, cognitive psychology, cognitive neuroscience, developmental psychology, social psychology, school psychology and various inter-subdisciplinary fields (Hartshorne and Schachner 2012)
- 129 of 238 studies (54%) examined in the fields of neuroscience, developmental biology, immunology, cell and molecular biology, general biology (Vasilevsky et al. 2013)
- 25 of 45 (56%) clinical studies published in high-impact-factor specialty medical journals in 1990-2003 (Ioannidis 2005)
- 52 of 63 studies (78%) examined mostly from the oncology field; but several studies were from the fields of women's health and cardiovascular health (Prinz et al. 2011)
- 52 of 67 studies (89%) examined in the fields of haematology, oncology (Begley and Ellis 2012)
- 52 of 52 studies (100%) examined in the field of nutrition (Young and Karr 2011)

In the 1970s, it was contended that Type A personality (hard-charging, go-getter) caused heart attacks (e.g., O'Flynn-Comiskey 1979). Several negative studies failed to overturn the belief. Now, most agree the claim was wrong (e.g., Myrtek 2001). To better understand how false claims arise, we first comment on what is science and then what is human nature before presenting our own experiment to show how false claims can come about.

Richard Feynman (1918-1988) was awarded the Nobel Prize in physics. He made the simple, but profound statement (Feynman 1965): *…if it disagrees with experiment, it's wrong. In that*

simple statement is the key to science. Refer to a one minute video of him explaining this. Keep in mind that our fear genes dominate rational evaluation genes, a fact noted by H. L. Mencken (1922) *…the whole aim of practical politics is to keep the populace alarmed (and hence clamorous to be led to safety) by menacing it with an endless series of hobgoblins, all of them imaginary*.

Sir Karl Popper, one of the most influential philosophers of science of the 20th century, argued that science advances by the process of deduction alone (Popper 1968). Knowing that science should proceed (or not proceed) by hypothesis confirmed by data – and not by human psychology and the use of fear, we turn to a simple, hands-on simulation of the scientific process. The operation of science is conceptually simple (Figure 1): state a hypothesis (make a guess, as Feynman would say), run an experiment or collect data, then apply Feynman's rule.



Notes: The statistical and science operation of this is a bit convoluted. A statistical hypothesis starts by guessing that "nothing is going on" in an experiment, say factor A does not cause outcome B. This is called the *null hypothesis*. If you run an experiment and get a surprising result – i.e., it looks like A did cause B – you reject the null hypothesis as you have "surprising" evidence that A did cause B.

Figure 1. Outline of science process.

A surprising result in an experiment is measured by a p-value. A p-value is a number calculated from a statistical test that describes how likely (the probability) you are to have found a result in an experiment. It is a number between 0 and 1 and the smaller the number the more surprise (the greater the probability). If the p-value is equal to or more than 0.05, we accept the "nothing is going on" guess. If the p-value is less than 0.05 (which happens 1–in–20 times by chance), then the surprise is enough to tentatively reject the "nothing is going on" guess.

A student should say, wait a minute! A small p-value can happen about 5% of the time by chance. OK. Science requires replication – i.e., more independent experiments with similar results are needed – which may take time to validate. On the other hand, you can ignore the possibility of a chance finding and go ahead and write up and publish your findings.

Here is the first and most common way that researchers can ignore getting a small p-value by chance... you are likely to get, by chance, a "statistically significant" result (i.e., a p-value less than 0.05) by just asking a lot of questions in an experiment. For example, if 60 independent questions are asked in an experiment, there is about a 95% probability of at least one "statistically significant" result...

$1 - 0.95^{60}$ approximately equals 0.95

We used 60 questions as it fits with our upcoming hands-on demo. One sure way to do sound science is to just ask one single question. If the result is a p-value less than 0.05, you have tentative evidence for surprise. Replicate (do the experiment again). If the new p-value from the second experiment is also less than 0.05, you are on the way to confirmation. In science jargon, you have moved from a postulated theory and a hypothesis to the beginnings of a confirmed theory.

Questionable research practices are research behaviors used during the collection and analysis of data and the reporting of results that can increase false-positive findings in the literature (Colquhoun 2017, Andrade 2021). Two of these practices include p-hacking and HARKing. A simple way to explain p-hacking is playing around with data until you find a result that meets the statistical standard for significance, which is p-value less than 0.05 (Hawkes 2018). A simple way to explain HARKing is to play around with data and then Hypothesize After the Results are Known (Kerr 1998). We will show how p-hacking and HARKing can unknowingly (or knowingly) occur in experiments.

An observational study is a type of experiment in which individuals are observed or certain outcomes are measured. There is no attempt made to affect the outcomes. It is typical in an observational study to ask many questions. To be fair, because of costs involved in studying large populations, it makes little sense to just ask one question. It is also typical, and unfortunate, that these types of studies do not clearly state how many questions were asked. It is also not very easy for readers to figure out how many questions were asked in these studies.

Fortunately, there are many statistical strategies to address the multiple question problem. These strategies allow researchers to control the probability of erroneously concluding that a chance finding (i.e., a finding with a p-value less than 0.05) is real. As examples, strategies were provided in the 1990s by Westfall and Young (1993), Shaffer (1995), and Benjamini-Hochburg (1995), and more recently by Dmitrienko and D'Agostino (2018).

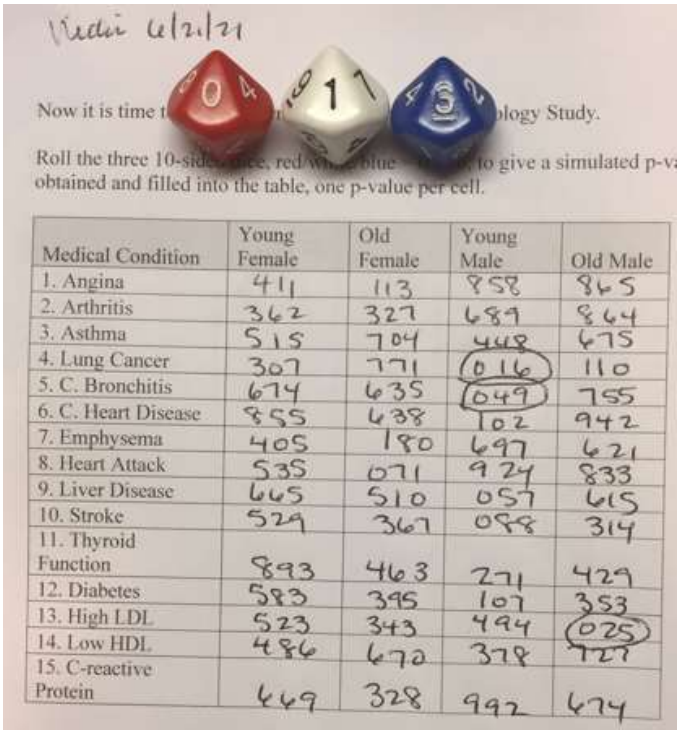
In sound science, the number of questions at issue is explicitly given as part of the study and analysis methods. If you have "skin in the game" you do not want to spend time/money following up on a false lead by using less than sound science principles. Here, industrial research is often better at using sound science principles to develop products or a service than university research. Why? Because a company has skin in the game – it must develop and sell a product or service that works, or it can go out of business.

Our experiment – We ran a hands-on demo having different people each roll three 10-sided dice multiple times to try show the importance of how small p-values occur by chance. The three 10-sided dice represents our data set, and each roll of the dice represented a separate test (question asked) of our data set.

Methods

We used colored dice (1 Red, 1 White, 1 Blue) as show in top of Figure 2. In June and July of 2021, we provided 10 friends with the dice and asked them to do the following:

- roll the three 10-sided dies 60 times
- after each roll, observe the face up digits
- enter the face up digits in a 15 × 4 data table as 0.RWB (R for Red, W for White, B for Blue); for example, the face up digits shown in the top of Figure 2 are 0.016



Nicki 6/21/21

Now it is time t... ...ology Study.

Roll the three 10-sided dice, red/white/blue ... to give a simulated p-va... obtained and filled into the table, one p-value per cell.

Medical Condition	Young Female	Old Female	Young Male	Old Male
1. Angina	411	113	858	865
2. Arthritis	362	327	689	864
3. Asthma	515	704	448	675
4. Lung Cancer	307	771	016	110
5. C. Bronchitis	674	635	049	755
6. C. Heart Disease	855	638	102	942
7. Emphysema	405	180	697	621
8. Heart Attack	535	071	924	833
9. Liver Disease	665	510	057	615
10. Stroke	529	367	088	314
11. Thyroid Function	893	463	271	429
12. Diabetes	583	395	107	353
13. High LDL	523	343	494	025
14. Low HDL	486	670	378	727
15. C-reactive Protein	669	328	992	674

Notes: C. Bronchitis refers to chronic bronchitis, C. Heart Disease refers to chronic heart disease, High LDL refers to high-density lipoprotein cholesterol, Low LDL refers to low-density lipoprotein cholesterol, C-reactive Protein is a protein made by your liver and sent into your bloodstream in response to inflammation.

Figure 2. Result of 60 dice rolls for friend 1 (Nicki).

The data table was set up to simulate medical conditions in a hypothetical observational study. A 'causative factor' in research is a factor that may be responsible for causing something. The data table was set up using 15 fictitious medical conditions (refer to the left column in Figure 2) for an arbitrary causative factor, and four sub-groups of interest (young and old females, and young and old males listed in the four right-hand columns). Our friends were asked to enter their first dice roll result in the top left of the table, second result below it, and so on until all 60 rolls were completed.

When researchers perform an observational study looking at the role of a causative factor in possibly causing a medical condition, they would have collected (or acquired) data for a specific population. We can call this a cohort population. The data might be for observations of participants' behaviors (or exposures) and of various medical conditions occurring over time.

Let's say they want to understand whether red meat consumption is associated with various medical conditions over time. To make it interesting, the population can be divided into sub-groups like those in Figure 2. The assumption here is that any particular sub-group may be at higher risk of a medical condition from the special behavior (or exposure) being studied.

A test can be performed to answer the question whether red meat consumption leads to excess angina (i.e., chest pain caused by reduced blood flow to the heart) in young females. The data is analyzed, and a p-value is obtained for this question. The p-value would be entered into our first box in the data table in Figure 2. The second test asks the question whether red meat consumption leads to excess arthritis in young females using the same data set, and so on until all 60 tests are completed. Multiple testing on the same data set such as this is routine in observational studies in today's world, enabled with the use of modern statistical software.

An important point is made – There is no difference in rolling three 10-sided diced 60 times versus performing 60 statistical tests on a data set using statistical software. There is about a 95% probability of at least one "statistically significant" result occurring by chance in either situation!

Results

The table of dice roll results produced by our first friend Nicki is shown in Figure 2. Notice that her smallest simulated (dice roll) p-value was 0.016. She arranged the dice to show that value before she took the picture. Also, note that she circled three p-values less than 0.05 – 0.016, 0.025 and 0.049.

If she were a researcher and wanted to p-hack, she could possibly write and publish three separate research papers showing these surprising findings. The lead sentence (claim) for each research paper could be:

1. 'red meat is associated with lung cancer in young males'
2. 'red meat is associated with chronic bronchitis in young males'
3. 'red meat is associated with high levels of high LDL (high-density lipoprotein cholesterol) and this has been linked to heart attacks'

Keep in mind that fear genes dominate human psychology, and each claim calls on those genes.

If she were a researcher and wanted to HARK, she could ignore the p-value results greater than 0.05 and try write and publish research papers only presenting and building up stories around the three surprising findings.

As the dice were rolled and the numbers were entered into the 15 × 4 table, there should be some cognitive dissonance (discomfort from holding two conflicting beliefs or values). On the one hand, the dice rolling is producing random numbers. On the other hand, we are seeing p-values that a researcher can use to make a science claim.

Recall, a researcher can ignore the possibility of producing random results when they ask a lot of questions of their data set. All they need to do is ask a lot of questions until they get a result with a p-value to be less than 0.05 so that it is more likely they can publish their findings by p-hacking or HARKing (and not perish)! Worse still, these researchers might not even show their entire results when they publish, just note the "statistically significant" results.

We mentioned replication, essential for sound science. Look at the data table completed by another friend (Barbara) in Figure 3. You will note in Figure 3 that Barbara got two "statistically significant" results – 0.001 and 0.024. If she wanted to p-hack or HARK, she can potentially write only two papers. More importantly, note that she did not replicate any finding by Nicki (compare Figure 3 with Figure 2)!

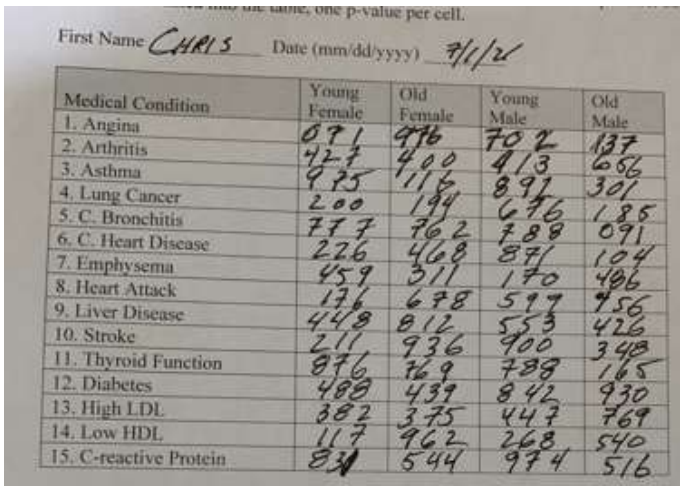
First Name [illegible] Date (mm/dd/yyyy) 7/14/21

Medical Condition	Young Female	Old Female	Young Male	Old Male
1. Angina	.438	.961	459	.649
2. Arthritis	.912	.972	.090	.143
3. Asthma	.733	.438	.124	.321
4. Lung Cancer	.085	.776	.111	.709
5. C. Bronchitis	.692	.237	.162	.293
6. C. Heart Disease	.875	.314	.764	.387
7. Emphysema	.232	.686	.124	.133
8. Heart Attack	.423	.220	.513	.416
9. Liver Disease	.788	.506	.206	.489
10. Stroke	.407	.086	.595	.191
11. Thyroid Function	.732	.497	.861	.001
12. Diabetes	.277	.848	.829	.294
13. High LDL	.769	.825	.936	.660
14. Low HDL	.103	.084	.154	.829
15. C-reactive Protein	.748	.024	.272	.304

Figure 3. Result of 60 dice rolls for friend 2 (Barbara).

It is possible to roll 60 times and not have any p-value less than 0.05. That is expected to happen about 5% of the time. Our friend Chris experienced this – refer to Figure 4. He had no p-values less than 0.05. However, there is indirect good news – Chris did not replicate any findings of Nicki or Barbara. Negative results like this are informative – i.e., the small p-value findings of Nicki and Barbara are nothing more than random results, they are not real!

P-values should be evenly distributed over the interval 0 to 1. It is often informative to view the distribution of p-values for a set of results in an experiment. This can be done using a p-value plot after Schweder and Spjøtvoll (1982). A p-value plot can be constructed by sorting the p-values from smallest to largest and plotting them against the integers 1, 2, 3, . . . , 60 (in this case). Figure 5 shows p-value plots of 60 dice roll results for nine of our friends.



First Name CHRIS Date (mm/dd/yyyy) 7/1/21

Medical Condition	Young Female	Old Female	Young Male	Old Male
1. Angina	071	996	702	137
2. Arthritis	427	400	413	666
3. Asthma	935	116	897	301
4. Lung Cancer	200	144	676	185
5. C. Bronchitis	777	962	788	091
6. C. Heart Disease	226	468	871	104
7. Emphysema	459	311	170	406
8. Heart Attack	176	678	597	956
9. Liver Disease	449	812	553	426
10. Stroke	211	936	900	348
11. Thyroid Function	876	769	788	165
12. Diabetes	488	439	842	930
13. High LDL	382	375	447	769
14. Low HDL	117	962	268	540
15. C-reactive Protein	831	544	974	516

Figure 4. Result of 60 dice rolls for friend 3 (Chris).

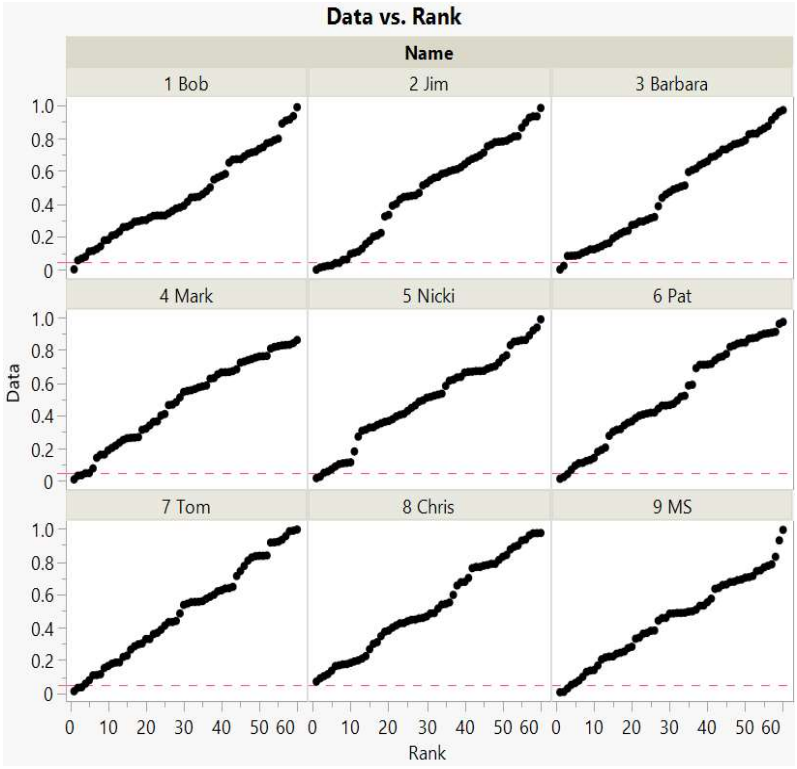


Figure 5. P-value plots for nine simulations of 60 p-values.

Researchers can easily get small p-values by asking a lot of questions (p-hacking). Head et al. (2015) note that p-hacking is endemic in research. Researchers can also HARK. Both are questionable research practices that can lead to false claims in literature. As researchers most often do not attempt to publish negative results, it is possible that false claims become mistaken as fact (canonized) in literature (Nissen et al. 2016).

Lessons learned

It is unfortunately routine and unethical science practice for researchers not to correct an analysis when many questions are asked in an experiment. There are well-known ways to correct an analysis to adjust for the number of questions asked. We provided some sources for making corrections earlier.

As a student (or a citizen), you can often recognize these practices as a research claim may be made with qualifying statements – such that a causative factor may be important for a sub-group (e.g., women). Or a claim is just plain nonsensical – such as coffee causes pancreatic cancer.

Two words have entered the science lexicon – p-hacking and HARKing. Both words are quite harsh, and both indicate what is known as questionable research practices. Getting 60 p-value results in a multiple testing environment and taking small p-values less than 0.05 at face value is p-hacking. Discarding all the p-values greater than 0.05 and making up stories for p-values less than 0.05 is HARKing.

Acknowledgements

No external funding was provided for this study. The cooperation of our friends performing the dice rolls – which included Bob, Jim, Barbara, Mark, Nicki, Pat, Tom, Chris, MS, and Juli – is acknowledged.

References

Andrade C. 2021. HARKing, cherry-picking, p-hacking, fishing expeditions, and data dredging and mining as questionable research practices. *Journal of Clinical Psychiatry* 82(1), 20f13804. https://doi.org/10.4088/JCP.20f13804

Begley CG, Ellis LM. 2012. Raise standards for preclinical cancer research. *Nature* 483, 531–533. https://doi.org/10.1038/483531a

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* 57(1), 125–133. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Colquhoun D. 2017. The reproducibility of research and the misinterpretation of p-values. *Royal Society Open Science* 4(12), 171085. https://doi.org/10.1098/rsos.171085

Dmitrienko A, D'Agostino RB Sr. 2018. Multiplicity considerations in clinical trials. *New England Journal of Medicine* 378(22), 2115–2122. https://doi.org/10.1056/NEJMra1709701

Feynman R. 1965. *The Character of Physical Law*. Cambridge, MA: The MIT Press.

Glasziou P, Meats E, Heneghan C, Shepperd S. 2008. What is missing from descriptions of treatment in trials and reviews? *British Medical Journal* 336(7659), 1472–1474. https://doi.org/10.1136/bmj.39590.732037.47

Hartshorne J.K, Schachner A. 2012. Tracking replicability as a method of post-publication open evaluation. *Frontiers in Computational Neuroscience* 6, 8. https://doi.org/10.3389/fncom.2012.00008

Hawkes N. 2018. Sixty seconds on . . . P-hacking. *British Medical Journal* 362: k4039. https://doi.org/10.1136/bmj.k4039

Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD. 2015. The extent and consequences of p-hacking in science. *PLoS Biology* 13(3), e1002106. https://doi.org/10.1371/journal.pbio.1002106

Ioannidis JPA. 2005. Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association* 294, 218–228. https://doi.org/10.1001/jama.294.2.218

Kerr NL. 1998. HARKing: hypothesizing after the results are known. *Personality and Social Psychology Review* 2(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Mencken HL. 1922. *In Defense of Women*. New York, NY: AA Knopf.

Myrtek M. 2001. Meta-analyses of prospective studies on coronary heart disease, type A personality, and hostility. *International Journal of Cardiology* 79(2−3), 245−251 https://doi.org/10.1016/s0167-5273(01)00441-7

Nissen SB, Magidson T, Gross K, Bergstrom CT. 2016. Publication bias and the canonization of false facts. *eLife* 5, e21451. https://doi.org/10.7554/eLife.21451

O'Flynn-Comiskey AI. 1979. Stress: the type A individual. *American Journal of Nursing* 79, 11, 1956−1958. https://doi.org/10.2307/3424528

Popper KR. 1986. *The Logic of Scientific Discovery*. New York, NY: Harper & Row.

Prinz F, Schlange T, Asadullah K. 2011. Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery* 10(9), 712. https://doi.org/10.1038/nrd3439-c1

Schaffer JP. 1995. Multiple hypothesis testing. *Annual Review of Psychology* 46, 561−584. https://doi.org/10.1146/annurev.ps.46.020195.003021

Schweder T, Spjøtvoll E. 1982. Plots of P-values to evaluate many tests simultaneously. *Biometrika* 69(3), 493–502. https://doi.org/10.1093/biomet/69.3.493

Vasilevsky NA, Brush MH, Paddock H, Ponting L, Tripathy SJ, Larocca GM, Haendel MA. 2013. On the reproducibility of science: Unique identification of research resources in the biomedical literature. *PeerJ* 1, e148. https://doi.org/10.7717/peerj.148

Westfall PH, Young SS. 1993. *Resampling-based Multiple Testing*. New York, NY: John Wiley & Sons.

Young SS, Karr A. 2011. Deming, data and observational studies. *Significance* 8(3), 116−120. https://doi.org/10.1111/j.1740-9713.2011.00506.x