RESEARCH · RESEARCH INTEGRITY

P-Hacking, HARKing and Questionable Research Practices

What p-hacking, HARKing and other questionable research practices are, how common they are, why they produce false findings, and how preregistration and full reporting prevent them.

ModQu research article · 735 words · 7 references (APA 7th edition)

Outright fraud is rare in research. A much more common threat to reliable findings comes from practices that sit in a grey zone: not clearly dishonest, often done without any intent to mislead, but capable of making weak or false results look convincing. These are known as **questionable research practices**. This article explains the two best known, p-hacking and HARKing, and what can be done to prevent them.

P-hacking

Many studies judge a result by its **p-value**, with values below 0.05 conventionally treated as statistically significant. **P-hacking** means trying different ways of collecting or analysing data until a significant result appears, and then reporting that version as if it had been the plan all along. Common forms are shown below.

Common forms of p-hacking

Each choice is defensible alone; trying many and reporting one is not

Practice	What happens	Safeguard
Optional stopping	Stop when p drops below 0.05	Fix sample size in advance
Choosing outcomes	Report the measure that worked	Name the main outcome first
Selective exclusions	Drop data points that hurt	Set exclusion rules in advance
Flexible controls	Add or drop covariates	Pre-specify the model
Subgroup fishing	Test groups until one differs	Report every test run

Drawing on Simmons et al. (2011) ModQu.com

Figure 1. Five common forms of p-hacking and how to guard against them. Table by ModQu, drawing on Simmons et al. (2011).

Each choice may be defensible on its own. The problem is that trying many of them and reporting only the one that works greatly increases the chance of a false positive. Simmons et al. (2011) showed through simulations that combining just a few such choices could raise the false-positive rate from the nominal 5 per cent to more than 60 per cent. Gelman and Loken (2014) pointed out that the same problem arises even when researchers do not consciously try several analyses, because choices made after seeing the data are shaped by what the data show. They called this the **garden of forking paths**.

P-hacking leaves traces in the published record. Masicampo and Lalande (2012) found that p-values just below 0.05 appeared in psychology journals more often than would be expected. Head et al. (2015), mining p-values from a large body of published research, found evidence that p-hacking was widespread, although they judged that it probably did not change the overall conclusions of most meta-analyses they examined.

HARKing

Kerr (1998) introduced the term **HARKing**, for hypothesising after the results are known. A researcher looks at the data, notices an unexpected pattern and then writes the paper as though that pattern had been predicted from the start. The result looks like a successful test of a hypothesis when it was in fact an exploratory finding. Because any data set contains some chance patterns, HARKing turns noise into apparently confirmed theory, and it hides from readers how many other patterns were examined and discarded.

HARKing should not be confused with exploration. Noticing an unexpected pattern and reporting it as a new hypothesis for future testing is good practice. The problem lies only in presenting it as a prediction.

How common are these practices?

John et al. (2012) surveyed about 2,000 academic psychologists, using a method that rewarded honest answers, and found that many admitted to at least one questionable practice. Common examples included failing to report all of a study's measures, deciding whether to collect more data after checking the results and reporting only the studies that produced the desired result. Many respondents regarded these practices as defensible, which suggests that they were part of normal working habits rather than deliberate deception.

Why they matter

Questionable practices make published effects look larger and more reliable than they are. They help explain why many findings fail to replicate and why the literature in some fields contains more positive results than real effects could produce. They also waste resources, as other researchers and organisations build on results that do not hold.

Prevention

The most effective safeguard is to separate planning from analysis. **Preregistration** means recording the hypotheses, sample size, outcomes and analysis plan in a time-stamped public registry before the data are collected (Nosek et al., 2018). Analyses that follow the plan can then be reported as confirmatory, and anything else is clearly labelled as exploratory. Other safeguards include deciding sample sizes in advance, reporting all measures, conditions and exclusions, and sharing data and code so that others can check the analysis.

The same principles apply in business research. An A/B test checked every day and stopped as soon as it looks significant, or a survey analysis that searches subgroups until one shows a difference, is p-hacking whatever the setting.

Conclusion

P-hacking and HARKing do not require bad intentions to cause harm. They exploit the flexibility present in every study. Making plans in advance, reporting everything and labelling exploratory findings honestly are simple steps that greatly improve the reliability of results.

References

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102*(6), 460–465. https://doi.org/10.1511/2014.111.460

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

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524–532. https://doi.org/10.1177/0956797611430953

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

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of p values just below .05. *Quarterly Journal of Experimental Psychology, 65*(11), 2271–2279. https://doi.org/10.1080/17470218.2012.711335

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632