What's a multiverse good for anyway?

Julia M. Rohrer[1], Jessica Hullman[2], and Andrew Gelman[3]

Multiverse analysis has become a fairly popular approach, as indicated by the present special issue on the matter. Here, we take one step back and ask why one would conduct a multiverse analysis in the first place. We discuss various ways in which a multiverse may be employed – as a tool for reflection and critique, as a persuasive tool, as a serious inferential tool – as well as potential problems that arise depending on the specific purpose. For example, it fails as a persuasive tool when researchers disagree about which variations should be included in the analysis, and it fails as a serious inferential tool when the included analyses do not target a coherent estimand. Then, we take yet another step back and ask what the multiverse *discourse* has been good for and whether any broader lessons can be drawn. Ultimately, we conclude that the multiverse *does* remain a valuable tool; however, we urge against taking it too seriously.

Arguably, the most important step in a multiverse analysis is to consider doing it in the first place. But why would one put in all the effort—what's a multiverse good for anyway? Here, we will consider and contrast possible answers to this question. Going back to the roots, we start with ideas that one of us had when he applied the term as part of Steegen et al. (2016). From this angle, a multiverse analysis is mostly a tool for reflection and critique, and less of an approach to primary data analysis or for inferential summaries. But there is little reason to privilege initial authorial intent over subsequent developments and applied practices, and so we will turn to other potential purposes: the multiverse as a rhetorical tool and the multiverse as an (earnest) inferential tool. In the end, we will zoom out to see whether there are any broader lessons for the scientific community—what's the multiverse discourse good for anyway?

1. The multiverse as a tool for reflection and critique

A team of researchers conducts a study, prepares and analyzes their data in a particular way, and reports the results. However, they never preregistered their study, and from their previous publications, it is clear that they may have plausibly analyzed their data in some other way. In

---

Version: February 3, 2026

This manuscript has been submitted to *AStA Advances in Statistical Analysis* in response to the call for papers: The Role of Multiverse Analysis in Statistical Modelling and Applications.

[1] Wilhelm Wundt Institute for Psychology, Leipzig University; julia.rohrer@uni-leipzig.de

[2] Department of Computer Science, Northwestern University

[3] Department of Statistics and Department of Political Science, Columbia University

such a situation, a multiverse highlights that the researchers *could* have conducted and reported many other analyses. And maybe they even *would* have done so, if the data had looked differently, although it is impossible to know for sure, given the lack of working time machines. In any case, different analyses could have led to different results, including different confidence intervals, p-values, and Bayesian inferences. This is the hypothetical space multiverse analysis opens up, motivated by observations that open-ended research practices, not just selective reporting, can undermine our confidence in research findings, including nominal p-values.

The original paper proposing the multiverse under this name, Steegen et al. (2016), squarely focuses on p-values. At the same time, none of the authors is known to be a proponent of p-values. That discrepancy may provide a first hint that their multiverse analysis is not necessarily designed as the optimal way for evaluating evidence. Additionally, if one wanted to optimally evaluate evidence for the hypothesis being scrutinized (an interaction between women's fertility and their relationship status on outcomes such as political attitudes), many improvements to study design and data analysis are possible which do not require a multiverse at all.[1] But what is of interest here is not how to optimally collect and analyze data, but rather what the original authors of the fertility study (Durante et al. 2013) realistically may have done differently. In that manner, the multiverse analysis provides a post-mortem within the context of a literature that heavily focuses on p-values.

So here, the multiverse analysis serves as a critique of an individual study; more generally, one may think of it as a useful heuristic to reflect on published studies. What could the authors plausibly have done differently, and at which conclusions would they have arrived? In this spirit, the multiverse may be taught to raise awareness of the prevalence and consequences of seemingly arbitrary data-analytic decisions, as proposed by Heyman and Vanpaemel (2022), with the added benefit that actually implementing a multiverse analysis provides an excellent opportunity to practice implementing various steps of data cleaning and data analysis. The "garden of forking paths" (Gelman and Loken 2013), which is instantiated by a multiverse analysis, should be seen not as an accusation of research misconduct but as a description and informal model of non-preregistered research workflows (including our own).

Steegen et al. (2016) further noted that the reflective exercise of a multiverse analysis may lead to the conclusion that there is a gaping hole in theory or in measurement, a point that has been picked up repeatedly throughout the multiverse discourse. They also suggested that "more typical multiverses will tend to be smaller" (Steegen et al. 2016) than the 120 to 210 model specifications presented. However, these would be considered rookie numbers at the time of writing, as thousands to billions of models have been included in such analyses (e.g., Ganslmeier and Vlandas 2025; Muñoz and Young 2018; Orben and Przybylski 2019). Part of this has to do with the fact that Steegen et al. chiefly considered the multiverse of data processing steps, whereas many applications now also vary modeling decisions—a possibility that was only foreshadowed by Steegen et al. under the label of the model multiverse (p. 710). But even just

[1] For example, to improve the measurement of fertility, one may treat the variable as continuous rather than dichotomous, and maybe even collect follow-up data on the onset of the next menstruation to arrive at a more precise estimate of ovulation with the help of backwards-counting (Arslan et al. 2023). To improve the measurement of relationship status, one may simply use an item without ambiguous response options.

focusing on variation in data processing, modern multiverses manage to be more massive (e.g., 528 data processing pipelines in Short et al. 2025).

In any case, the idea that the multiverse is mostly a tool for critique—rather than something to implement when initially presenting evidence—is by no means an obvious take-home message for readers of Steegen et al. (2016). In fact, the manuscript closes with the statement that "it should become standard practice to go beyond single data set analysis and to acknowledge the multiverse of statistical results." Readers may thus understandably conclude that this work recommends multiverse analysis as a default approach for data analysis.

2. The multiverse as a persuasive tool

At least one of us (JMR) has used a variation of multiverse analysis (specification curve analysis, Simonsohn et al. 2020) mainly as a persuasive tool to convince readers that certain conclusions are warranted. This almost inverts the function of the multiverse as a tool for critique; now it is meant to pre-emptively alleviate criticism. This can only work if the results are fairly unambiguous across specifications

Previous studies had concluded that birth order differences in the (very broad) Big Five personality traits are negligible; the multiverse study further investigated birth order differences in narrower personality traits such as risk taking and locus of control (Rohrer et al. 2017). Results indicated that for almost all of the traits considered, the vast majority of specifications result in no significant birth order differences.

The multiverse here also served an educational purpose. The potential for cherry-picking significant results in the birth order literature had been noted early on (Harris 2009; Harris 1998), and the multiverse illustrates this clearly, as it shows readers precisely which specifications could be picked to generate virtually any result. In that sense, this is still partly an application that treats the multiverse as a tool for reflection and critique. However, the focus has now shifted, resulting in a clear substantive conclusion rather than a call for more caution.

A multiverse employed for persuasive purposes may be thought of as a ramped-up robustness check. Indeed, Rohrer et al. (2017) started out with focal analyses supplemented by extensive robustness checks, before the first author learned about specification curve analysis. It may be possible to determine a sensible focal analysis (such as the models used in Rohrer et al. 2015 to address the very same research question) and then relegate the multiverse to the supplement. Just as with robustness checks, one could imagine that this usage of multiverse analysis leads to selective reporting, as multiverses that do not support the central conclusions sufficiently are omitted.

One major limitation of the multiverse as a persuasive tool is that it is not always successful at persuading. For some, hundreds of analyses may seem impressive and provide reassurance following some variation of the sure-thing principle ("no matter which decision one deems correct, the results are similar"). For others, results may appear far from homogeneous, and the wealth of analyses offers up many chances to disagree with the authors about whether all analyses really examine the same effect, whether different operationalizations really result in

equally valid measurements, and whether different analyses may simply diverge due to differing precision or power (Del Giudice and Gangestad 2021).

To consider one prominent example of such disagreement, Orben and Przybylski (2019) conducted a specification curve analysis of the association between adolescent well-being and digital technology use across three data sets and concluded that any association is small. This did not persuade Twenge et al. (2022), who doubted some of the data-analytic decisions and conducted their own specification curve analysis of the same data sets, concluding that there is a substantial negative association (for girls, looking at social media use specifically). Such disagreement leads us to the question of the utility of multiverse analysis if one does not just want to persuade readers that a certain conclusion is compelling, but instead make up one's mind in the first place—that is, is it any good for valid inferences in the face of genuine uncertainty about how to best process and analyze the data?

3. The multiverse as a serious inferential tool

Finally, we arrive at the multiverse as an inferential tool. This is the direction into which the approach is developing, as illustrated by special issues like this one, with the overarching goal "to formalize and advance a rigorous application of Multiverse Analysis." This is also where things get thorny, as various issues arise when implementing a multiverse analysis to answer substantive research questions in the presence of serious uncertainty about how to best process and analyse the data.

3.1. Statistical issues

We imagine that many articles included in this special issue will discuss statistical issues that arise in various implications of multiverse analyses. For example, one appeal of specification curve analysis (Simonsohn et al. 2020) is that it returns a *p*-value for the curve as a whole, but this leads to concerns about how to interpret this summary and the more general question of whether the corresponding null hypothesis is sensible and aligned with how researchers will tend to interpret findings (Artner 2022). To consider another example, the *Boba* workflow (Liu et al. 2021) allows users to "refine" an analysis based on model quality, which gives rise to the question of how to draw inferences from the refined analyses, given the complications of post-selection inference. And in some computationally intensive fields, such as EEG analysis, actually implementing all possible universes may be infeasible, which leads to the question of how one can best sample from the multiverse (Short et al. 2025).

3.2. The decision on what to include

On a more fundamental level, when multiverse analyses are conducted to answer research questions, discussions often boil down to one central question: *what* ought to be included in the multiverse? As noted earlier, the original multiverse analysis of Steegen et al. (2016) was limited to only a few of the many data coding and analysis decisions in the target study. The decision of how broadly to define the multiverse often turns out to be a central point of contention. In principle, there seems to be some consensus that all included model

specifications should be justified or justifiable, sometimes with the added more stringent criterion that any differences between them should be arbitrary.

The hard part is determining which analyses are actually justified and which specification choices can actually be considered arbitrary. Sometimes, as with the study of ovulation and voting, one can examine other articles in the subfield or even by the same authors or even other analyses in the same paper to see a range of coding and analysis options. Other times, there is no such clear reference set of comparison studies.

In these scenarios, arguments can arise about what should and what should not be included. For example, Del Giudice and Gangestad (2021) stress how non-equivalence may lead to biased estimates, and then go on to discuss different types of nonequivalence (measurement nonequivalence, effect nonequivalence, power/precision nonequivalence) and possible analytic decisions (Type E decision: principled equivalence, Type N decision: principled nonequivalence, Type U decision: uncertainty). While these distinctions may be perfectly sensible on a theoretical level, in practice, they will strongly hinge on both researchers' domain knowledge and their understanding of statistics. Additionally, we might expect Type U decisions, where the researcher is uncertain about whether the analyses are equivalent or unequivalent, to be common in practice, as social science researchers are often uncertain about the underlying causal model. However, when the researcher wants to explore different possible sets of covariates, they are advised to conduct separate multiverse analyses rather than a single analysis, increasing the complexity of both the author's and the reader's tasks. In the worst case, in the absence of preregistration, researchers may as well determine what is equivalent based on whether or not it supports their result, or whether it is easy to communicate. And even in the absence of such selectiveness, every researcher may end up with their own multiverse.

This "customized" nature of the multiverse was already visible in the analyses provided by Steegen et al. (2016), which explicitly used previous publications of the original authors to justify which data processing choices to implement. This customization is perfectly sensible if the goal is to understand an individual study reported by a specific set of authors. It turns into a problem if the goal is to seriously draw robust inferences in the presence of uncertainty.

3.3. Unclarity about estimands

One aspect of what to include in a multiverse has received particular attention—do the included analyses actually all target the same estimand? That is, do the many analyses even *try* to answer the same research question? If they do not, which roughly corresponds to a situation of effect nonequivalence (Del Guidice and Gangestad, 2021), it becomes unclear what the variability of the various analyses reveals. There is no reason to expect that different research questions result in the same answers.

In this spirit, Auspurg (2025) criticized a previously published set of multiverse analyses that concluded large variability in results (Ganslmeier and Vlandas 2025). According to the critical reanalysis, once a clear target estimand was chosen, many of the included models were no longer justified, and it was precisely those unjustified models that produced a lot of variability. Thus, the original multiverse exaggerates the genuine uncertainty *about the correct answer* by

piling on top uncertainty *about the precise research question*. The very same criticism has been raised against Many Analysts projects in which different teams of researchers analyze the same data to (ostensibly) answer the same research question (Silberzahn et al. 2018); it turns out that the variability in results can be greatly reduced by only including analyses that plausibly target the same estimand (Auspurg and Brüderl 2021). In such scenarios, there is a risk that the multiverse lends a guise of rigor and completeness although it is ill-equipped to substitute for a paucity of theory.

This is mainly an issue if one expects the approach, whether it be the multiverse or many analysts, to help in answering substantive research questions. If, instead, a multiverse analysis is conducted to criticize a study or a line of studies, it may be perfectly sensible to include analyses with different estimands as long as it is plausible that researchers may try out different estimands. They may even do so unwittingly, given that verbalised research questions and theories are often ambiguous (Frankenhuis et al. 2023) and given that researchers apparently lack training on how to translate research questions into precise estimands (Lundberg et al. 2021). From that angle, the precise estimand turns into yet another researcher's degree of freedom (Rohrer and Arel-Bundock 2025), reducing the credibility of published results. Likewise, if a Many Analysts project is thought of as a way to document potential discrepancies in how researchers tackle their data, heterogeneity in estimands is not a problem but rather an insight gained from the whole process (Rohrer et al. 2025).

3.4. How to present the findings

Once one has settled on what to include and conducted the multiverse analysis, how should one present the findings? After all, digesting and making sense of an individual statistical analysis can already be challenging. Many well-known heuristics, such as belief in the "law of small numbers," represent a tendency to suppress variance when faced with statistical information (Tversky and Kahneman 1971). Relatedly, some have argued that the prevalence of null hypothesis significance testing in the replication crisis arises from a pervasive compulsion to see quantities as dichotomous even when doing so is unnecessary (Greenland 2017). It seems unlikely that such issues would simply go away in the context of multiverse analysis. In fact, the opposite may be the case: it may be very hard for readers of a multiverse analysis to integrate uncertainty into their conclusions, given that hundreds or thousands of often rather unstructured results are dumped onto them. Combined with the impression of completeness that a multiverse can convey, misinterpretation may be the rule rather than the exception.

Hall et al. (2022) survey different ways to visualize multiverse analyses, including simple outcome histograms, specification curve plots, and outcome matrices. One general problem they noted is the "illusion of probability": Results that are more frequently returned by the multiverse may be deemed more likely to be correct, which is akin to interpreting variation *probabilistically*. However, such an interpretation may be unwarranted. Not all specified universes may be equally likely to be correct, and statistical dependence among the universes may exaggerate certainty. Instead, Hall et al. argue, variation should be treated *possibilistically*, which means that any result that occurs at all (i.e., even just in a single universe) should be considered a possible outcome. This possibilistic interpretation is only sensible if every universe included is indeed deemed justified; if one rejects one of the universes outright, one

need not consider the outcome it produces possible. But even if one considers all universes justified, the possibilistic interpretation is typically unattractive, and given the high value placed on papers that provide confirmatory evidence, it is difficult to blame readers (or even authors) for succumbing to the lure of the probabilistic interpretation.

There is active work aimed at easing the burden on readers who must reason about the inclusion of each path in the multiverse while not falling back on heuristics. For example, the Milliways visualization tool (Sarma et al. 2024) tries to make it easier for readers to distinguish between possibilistic uncertainty (due to variation in specification) and probabilistic uncertainty (due to estimation uncertainty) through the use of p-box visualizations (Ferson and Siegrist 2012). It also allows readers to subset the multiverse, making it easier to reason about the implications of different choices. However, such interactivity demands that readers take on an active role, and in the end, even with better tools, a multiverse may result in a significant cognitive burden simply due to the high-dimensional nature of the whole endeavor. This burden may be partially carried by authors who guide readers through the multiverse by narratively making sense of the many results; partially, it may remain on the side of the reader, in particular for those readers who do not fully agree with the authors' decisions.

3.5. Some thoughts on the multiverse as a serious inferential tool

Despite the many potential issues, there are multiple reasons why researchers may gravitate towards using the multiverse as a serious inferential tool. Researchers are usually more interested in making (positive) substantive claims, rather than just criticizing something that came before. Multiverse analysis comes with a catchy name and often eye-catching visualizations (such as specification curve plots), it delivers a fitting response to calls for increased robustness across fields of research, and it provides more information and transparency than a single cherry-picked analysis.

Another appeal of the multiverse as an inferential tool is that it ostensibly provides a data-driven way out when faced with various data-analytic decisions. It does not require committing to a supposedly single best way to make sense of the data. This lack of commitment matches a style of research that is aversive to making explicit assumptions. For example, in observational psychological research, researchers are often uncomfortable with explicitly discussing causality (Grosz et al. 2020); but without explicit causal considerations, it is impossible to justify an appropriate set of control variables (Rohrer 2018; Wysocki et al. 2022). Simply implementing all possible combinations of covariates that can be found in the literature may seem like a convenient solution, as it skirts the issue of causal assumptions while satisfying the need for *some* sort of statistical adjustment. Not everybody may deem this solution satisfying, given its lack of conceptual coherence and the resulting unclarity about estimands and assumptions (Auspurg 2025).

And there are also potential downsides to taking the multiverse too seriously. As spelled out by many apt critics, it may result in mistaken conclusions if nonsensical specifications are included; and in general, it may lead readers to misapprehend the uncertainties involved. This is all aggravated by the fact that the sheer *quantity* of analyses may appear very persuasive, resulting in a potentially false sense of certainty. And coherently criticizing a single statistical analysis

already takes a bit of effort; criticizing *hundreds* of analyses may often be impossible, as each specification may suffer from its own issues.

4. What's the multiverse discourse good for anyway?

Trying to build a more solid foundation for the method may lead to interesting insights and developments, regardless of whether the endeavor will ultimately succeed. But let's zoom out a bit. What has the *idea* of multiverse analysis and the ensuing controversy and discussion been good for? We think that the multiverse discourse itself holds some valuable meta-lessons.

First, the same tool may be used for different purposes by different people with diverse interests. For the multiverse, one of us (AG) saw the focus on highlighting uncertainty, up to the point where it leads to the conclusion that not much can be concluded (Steegen et al. 2016). One of us (JMR) has used it to make the point that some conclusions are rather certain (Rohrer et al. 2017). And one of us (JH) has worked on tools that address implementation and communication challenges to better enable authors and readers to do the necessary reasoning about what paths are justified (Sarma et al. 2024; Sarma et al. 2021).

As multiverse analysis evolved from a conversation among methodologists to a practice adopted by substantive researchers, the drift from highlighting uncertainty and criticizing claims to attempting to reduce uncertainty and support claims may be understandable. As has been repeatedly noted, it is the job of statisticians and methodologists to express uncertainty and scrutinize inferences; it is the job of the substantive researcher to, in the face of uncertainty, somehow arrive at justified inferences. These different motivations lead to different perspectives on various tools. For example, p-values may be thought of as tools for criticism—they can indicate that a favored model does not fit the data well (without necessarily being able to support any alternative model), and they can indicate that one should not get too excited about an observed association that turns out to be fully compatible with mere chance. But the preferred usage by substantive researchers is to support their models of the world (and stoke excitement). Some degree of tension may be inevitable.

Second, the multiverse has been a valuable training tool. This extends beyond scenarios in which it is explicitly used as such (e.g., multiverse in the classroom, Heyman and Vanpaemel 2022). Even an applied researcher who "just" wants to answer a substantive research question will have to put in some planning and coding effort to realize the large number of analyses involved.

Third, multiverse analysis has highlighted weak spots in the research process. This includes gaping holes in theory and measurement highlighted by Steegen et al. (2016), as well as problems in specifying estimands and connecting assumptions to analyses, particularly when it comes to causal inferences (Auspurg 2025; Rohrer et al. 2025). The brute force of the multiverse cannot *fix* these issues, and—just as with other science reform approaches (such as many-analyst projects or preregistration)—there is a risk that researchers mistake it for a silver bullet. But, when approached with the right mindset, the multiverse may give useful directions for how to ultimately deflate it (2016).

Fourth, the multiverse has inspired research across disciplinary lines, including contributions from computer scientists (e.g., Liu et al. 2021; Sarma et al. 2024), theorists of inference, and metascientists (e.g., Gelman and Loken 2014; Simonsohn et al. 2020; Steegen et al. 2016), and applied researchers deploying multiverse workflows in substantive domains (e.g., Rauvola and Rudolph 2023; Rohrer et al. 2017; Vuorre and Przybylski 2024). While the multiverse may never be a satisfying alternative for good, domain-specific theorizing, it has led to productive cross-pollination across disciplines, broadening the types of expertise brought to bear on issues of scientific reform.

As an analogy for putting the multiverse in perspective, consider something very simple: the rule that if you have a proportion, you can compute the standard error as sqrt(p*(1-p)/n). In real life, this rule will understate uncertainty because it ignores measurement error, dropout, nonresponse, misclassification, and so forth. At best, it's a lower bound on error, and sometimes not even that. But it's still a useful baseline, as long as you don't take it too seriously.

We feel similarly about the multiverse. The concept of the multiverse is valuable; it's a good way to think about the analyses that could be done. To see this, compare the multiverse to what came before, which was either (a) doing an analysis and taking the nominal p-values as meaningful, or (b) trying to do some sort of multiple testing correction. The problem with (a) is obvious, but (b) has a problem too in that it's focused on the goal of getting a "correct" p-value, which we feel takes users away from useful scientific goals (McShane et al., 2019).

In contrast, the multiverse is open to the idea that your analysis could have been different. Also, unlike multiple testing correction, the multiverse does not require a record of which analyses were done with the data at hand, or speculation of which analyses would have been done had the data been different. The multiverse is not forensic; instead, it accepts that there are many reasonable data coding and analysis choices, and it points ultimately toward some sort of multilevel model.

We cannot tell what the future holds for multiverse analysis. All things considered, we do think that it remains a valuable conceptual tool, a helpful reminder that any analysis could have looked differently, and a process to reason more openly about the legitimate choices involved in data coding and analysis. It can be seen as a way to embrace uncertainty, as long as one keeps in check the urge to turn it into a means to resolve that uncertainty.

References

Arslan, R.C., Blake, K., Botzet, L.J., Bürkner, P.-C., DeBruine, L., Fiers, T., Grebe, N., Hahn, A., Jones, B.C., Marcinkowska, U.M., Mumford, S.L., Penke, L., Roney, J.R., Schisterman, E.F., Stern, J.: Not within spitting distance: Salivary immunoassays of estradiol have subpar validity for predicting cycle phase. Psychoneuroendocrinology. 149, 105994 (2023). https://doi.org/10.1016/j.psyneuen.2022.105994

Artner, R.: Reproducibility, robustness, and test severity, https://lirias.kuleuven.be/retrieve/656755, (2022)

Auspurg, K.: Robustness is better assessed with a few thoughtful models than with billions of regressions. Proc. Natl. Acad. Sci. U. S. A. 122, e2521917122 (2025). https://doi.org/10.1073/pnas.2521917122

Auspurg, K., Brüderl, J.: Has the Credibility of the Social Sciences Been Credibly Destroyed?

Reanalyzing the “Many Analysts, One Data Set” Project. Socius. 7, 23780231211024421 (2021). https://doi.org/10.1177/23780231211024421
Del Giudice, M., Gangestad, S.W.: A traveler’s guide to the multiverse: Promises, pitfalls, and a framework for the evaluation of analytic decisions. Adv. Methods Pract. Psychol. Sci. 4, 251524592095492 (2021). https://doi.org/10.1177/2515245920954925
Durante, K.M., Rae, A., Griskevicius, V.: The fluctuating female vote: politics, religion, and the ovulatory cycle: Politics, religion, and the ovulatory cycle. Psychol. Sci. 24, 1007–1016 (2013). https://doi.org/10.1177/0956797612466416
Ferson, S., Siegrist, J.: Verified computation with probabilities. In: IFIP Advances in Information and Communication Technology. pp. 95–122. Springer Berlin Heidelberg, Berlin, Heidelberg (2012)
Frankenhuis, W.E., Panchanathan, K., Smaldino, P.E.: Strategic ambiguity in the social sciences. Soc. Psychol. Bull. 18, 1–25 (2023). https://doi.org/10.32872/spb.9923
Ganslmeier, M., Vlandas, T.: Estimating the extent and sources of model uncertainty in political science. Proc. Natl. Acad. Sci. U. S. A. 122, e2414926122 (2025). https://doi.org/10.1073/pnas.2414926122
Gelman, A., Loken, E.: The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time∗, http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf
Gelman, A., Loken, E.: Ethics and statistics: The AAA tranche of subprime science. Chance (N. Y.). 27, 51–56 (2014). https://doi.org/10.1080/09332480.2014.890872
Greenland, S.: Invited commentary: The need for cognitive science in methodology. Am. J. Epidemiol. 186, 639–645 (2017). https://doi.org/10.1093/aje/kwx259
Grosz, M.P., Rohrer, J.M., Thoemmes, F.: The Taboo Against Explicit Causal Inference in Nonexperimental Psychology. Perspect. Psychol. Sci. 15, 1243–1255 (2020). https://doi.org/10.1177/1745691620921521
Hall, B.D., Liu, Y., Jansen, Y., Dragicevic, P., Chevalier, F., Kay, M.: A survey of tasks and visualizations in multiverse analysis reports. Comput. Graph. Forum. 41, 402–426 (2022). https://doi.org/10.1111/cgf.14443
Harris, J.R.: The nurture assumption. Bloomsbury Publishing PLC, London, England (1998)
Harris, J.R.: The nurture assumption. Simon & Schuster, New York, NY (2009)
Heyman, T., Vanpaemel, W.: Multiverse analyses in the classroom. Meta-Psychology. 6, (2022). https://doi.org/10.15626/mp.2020.2718
Liu, Y., Kale, A., Althoff, T., Heer, J.: Boba: Authoring and visualizing multiverse analyses. IEEE Trans. Vis. Comput. Graph. 27, 1753–1763 (2021). https://doi.org/10.1109/TVCG.2020.3028985
Lundberg, I., Johnson, R., Stewart, B.M.: What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory. Am. Sociol. Rev. 86, 532–565 (2021). https://doi.org/10.1177/00031224211004187
McShane, B. B. Gal, D., Gelman, A., Robert, C. & Tackett, J. L. (2019). Abandon statistical significance. *American Statistician* 73(S1), 235-245.
Muñoz, J., Young, C.: We ran 9 billion regressions: Eliminating false positives through computational model robustness. Sociol. Methodol. 48, 1–33 (2018). https://doi.org/10.1177/0081175018777988
Orben, A., Przybylski, A.K.: The association between adolescent well-being and digital technology use. Nat Hum Behav. 3, 173–182 (2019). https://doi.org/10.1038/s41562-018-0506-1
Rauvola, R.S., Rudolph, C.W.: Worker aging, control, and well-being: A specification curve analysis. Acta Psychol. (Amst.). 233, 103833 (2023). https://doi.org/10.1016/j.actpsy.2023.103833
Rohrer, J.M.: Thinking Clearly About Correlations and Causation: Graphical Causal Models for

Observational Data. Advances in Methods and Practices in Psychological Science. 1, 27–42 (2018). https://doi.org/10.1177/2515245917745629
Rohrer, J.M., Arel-Bundock, V.: Models as prediction machines: How to convert confusing coefficients into clear quantities, http://dx.doi.org/10.31234/osf.io/g4s2a_v3, (2025)
Rohrer, J.M., Egloff, B., Schmukle, S.C.: Examining the effects of birth order on personality. Proc. Natl. Acad. Sci. U. S. A. 112, 14224–14229 (2015). https://doi.org/10.1073/pnas.1506451112
Rohrer, J.M., Egloff, B., Schmukle, S.C.: Probing Birth-Order Effects on Narrow Traits Using Specification-Curve Analysis. Psychol. Sci. 28, 1821–1832 (2017). https://doi.org/10.1177/0956797617723726
Rohrer, J.M., Smith, G.D., Munafò, M.: What can be learned when multiple analysts arrive at different estimates. Eur. J. Epidemiol. 40, 493–495 (2025). https://doi.org/10.1007/s10654-025-01249-2
Sarma, A., Hwang, K., Hullman, J., Kay, M.: Milliways: Taming multiverses through principled evaluation of data analysis paths. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. pp. 1–15. ACM, New York, NY, USA (2024)
Sarma, A., Kale, A., Moon, M., Taback, N., Chevalier, F., Hullman, J., Kay, M.: multiverse: Multiplexing Alternative Data Analyses in R Notebooks (Version 0.6.2), https://github.com/MUCollective/multiverse, (2021)
Short, C.A., Hildebrandt, A., Bosse, R., Debener, S., Özyağcılar, M., Paul, K., Wacker, J., Kristanto, D.: Lost in a large EEG multiverse? Comparing sampling approaches for representative pipeline selection. J. Neurosci. Methods. 424, 110564 (2025). https://doi.org/10.1016/j.jneumeth.2025.110564
Silberzahn, R., Uhlmann, E.L., Martin, D.P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M.A., Dalla Rosa, A., Dam, L., Evans, M.H., Flores Cervantes, I., Fong, N., Gamez-Djokic, M., Glenz, A., Gordon-McKeon, S., Heaton, T.J., Hederos, K., Heene, M., Hofelich Mohr, A.J., Högden, F., Hui, K., Johannesson, M., Kalodimos, J., Kaszubowski, E., Kennedy, D.M., Lei, R., Lindsay, T.A., Liverani, S., Madan, C.R., Molden, D., Molleman, E., Morey, R.D., Mulder, L.B., Nijstad, B.R., Pope, N.G., Pope, B., Prenoveau, J.M., Rink, F., Robusto, E., Roderique, H., Sandberg, A., Schlüter, E., Schönbrodt, F.D., Sherman, M.F., Sommer, S.A., Sotak, K., Spain, S., Spörlein, C., Stafford, T., Stefanutti, L., Tauber, S., Ullrich, J., Vianello, M., Wagenmakers, E.-J., Witkowiak, M., Yoon, S., Nosek, B.A.: Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. Advances in Methods and Practices in Psychological Science. 1, 337–356 (2018). https://doi.org/10.1177/2515245917747646
Simonsohn, U., Simmons, J.P., Nelson, L.D.: Specification curve analysis. Nat. Hum. Behav. 4, 1208–1214 (2020). https://doi.org/10.1038/s41562-020-0912-z
Steegen, S., Tuerlinckx, F., Gelman, A., Vanpaemel, W.: Increasing Transparency Through a Multiverse Analysis. Perspect. Psychol. Sci. 11, 702–712 (2016). https://doi.org/10.1177/1745691616658637
Tversky, A., Kahneman, D.: Belief in the law of small numbers. Psychol. Bull. 76, 105–110 (1971). https://doi.org/10.1037/h0031322
Twenge, J.M., Haidt, J., Lozano, J., Cummins, K.M.: Specification curve analysis shows that social media use is linked to poor mental health, especially among girls. Acta Psychol. . (2022)
Vuorre, M., Przybylski, A.K.: A multiverse analysis of the associations between internet use and well-being. Technology, Mind, and Behavior. 5, 65 (2024)
Wysocki, A.C., Lawson, K.M., Rhemtulla, M.: Statistical Control Requires Causal Justification. Advances in Methods and Practices in Psychological Science. 5, 25152459221095823 (2022). https://doi.org/10.1177/25152459221095823