to appear in AStA Advances in Statistical Analysis

What's a multiverse good for anyway?

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

[1] Wilhelm Wundt Institute for Psychology, Leipzig University, Leipzig, Germany

[2] Department of Computer Science, Northwestern University, Evanston, Illinois, United States

[3] Department of Statistics and Department of Political Science, Columbia University, New York, United States.

Correspondence concerning this manuscript should be addressed to Julia M. Rohrer, julia.rohrer@uni-leipzig.de

ORCiD:
JMR: https://orcid.org/0000-0001-8564-4523
JH: https://orcid.org/0000-0001-6826-3550
AG: https://orcid.org/0000-0002-6975-2601

Competing interests: The authors have no relevant financial or non-financial interests to disclose.

Contribution statement: Conceptualization: Julia M. Rohrer, Jessica Hullman, Andrew Gelman; Writing – original draft preparation: Julia M. Rohrer; Writing – review and editing: Julia M. Rohrer, Jessica Hullman, Andrew Gelman
Ethics approval, consent, data, material and or/code availability: Does not apply – no empirical data were collected or presented.
Acknowledgments: No funding was received to assist with the preparation of this manuscript

What's a multiverse good for anyway?

Abstract: 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.

Keywords: multiverse analysis, research degrees of freedom, estimands, meta-science

---

Arguably, the most important step in a multiverse analysis is to consider doing it in the first place. Why would one put in all the effort—what's a multiverse good for anyway? Here we consider and contrast possible answers to this question.

Going back to one of the roots, we start with ideas that one of us had when he first applied the term (Steegen et al. 2016). This conception views multiverse analysis as 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 for multiverse-style analyses, multiple lineages can be traced as different sets of authors in different literatures developed their own approaches at around the same time, such as the literature discussing "vibration of effects" (Palpacuer et al. 2019; Patel et al. 2015), or the "computational solution" to model uncertainty (Young 2018; Young and Holsteen 2017). So our discussion will turn to other potential purposes: the multiverse as a persuasive tool and the multiverse as an (serious) 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 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 (Gelman and Loken 2013).

The original paper proposing the multiverse under this name, Steegen et al. (2016), focuses on p-values (see Fig. 1A). 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 was not designed to demonstrate the optimal way to evaluate a body of evidence. Additionally, one may consider the running example they provide: a re-analysis of an interaction between women's fertility and their relationship status on outcome such as political attitudes originally reported by Durante et al. (2013). The design and the data analysis of the original study could have been improved in many different ways which do not require a multiverse at all.[1] However, 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.

[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.

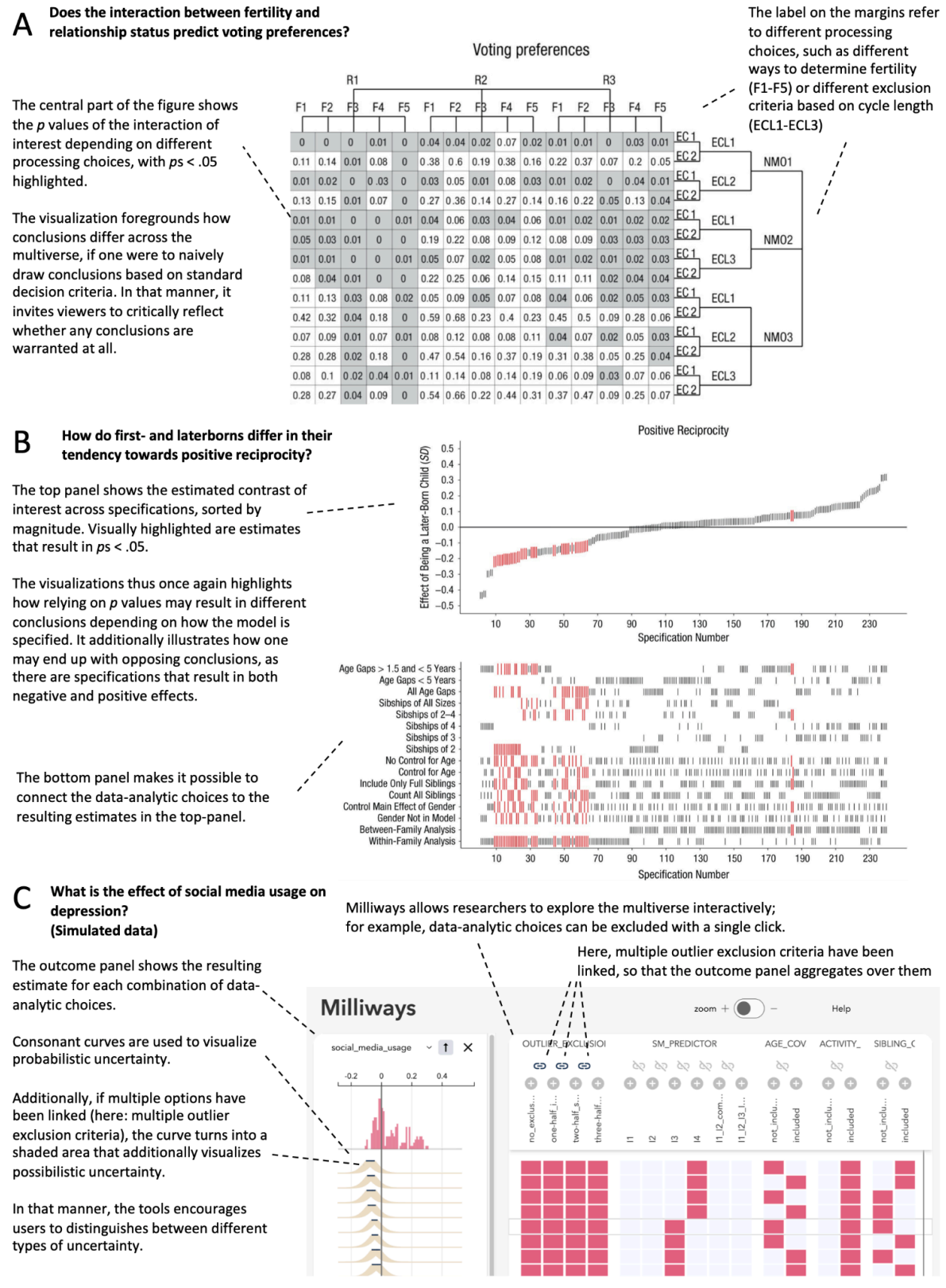

Fig. 1. Different ways to present the results of a multiverse analysis. Panel A adapted from Steegen et al. (2016); panel B adapted from Rohrer et al. (2017); panel C adapted from Sarma et al. (2024).

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 in analysis, 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.

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. For example, in their multiverse they varied whether women who reported that they were "dating or involved with only one partner" were counted as in a relationship (pooled with those who were engaged, living with a partner, or married) or as single (pooled with those not dating). Both coding schemes had been used in previous studies on the topic, which suggests that the theoretical rationale underlying the research is underspecified. If there was a clear understanding of what *precisely* it is about relationship status that is supposed to modify the effects of fertility on political attitudes, that would render one of the two options—or maybe some completely different way to assess relationship status—the appropriate choice. This general idea that the multiverse identifies gaps in our understanding has come up repeatedly since.

The original multiverse paper 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 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." We therefore cannot blame readers who 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; see Fig. 1B for a visualization) 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* considered a narrower estimand more suitable. First, the authors argued that some control variables included by Orben and Przybylski (e.g., negative attitudes towards school, school grades and closeness to parents) should not be included based on the reasoning that they are more likely mediators rather than confounders between well-being and digital technology use. Second, they argued that Orben and Przybylski's operationalization of mental health (combinations of mental health items, including individual items as well as overall scale scores) was suboptimal and instead treated each of the four supposedly validated mental health scales as one possible outcome. Third, the authors argued that negative associations are largely to be

expected for social media use (rather than screen time or digital media use more generally) and for girls (rather than boys), justifying the analyses to be separated by both type of screen activity and gender. The authors conducted their own modified specification curve analysis and in turn concluded that there is indeed a substantial negative association—in 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 modern inquiry is tending, as illustrated by special issues like this one, with the overarching goal "to formalize and advance a rigorous application of Multiverse Analysis" (Tomaselli et al. 2026). 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 implementations of multiverse analyses. For example, 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). To take another 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 yet 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.

There have been promising developments on these matters, including rather recently the development of post-selection inference in multiverse analysis (PIMA; Girardi et al. 2024), which provides a properly formalized framework to flexibly test different types of hypotheses, including tests across all models but also tests for individual models chosen post hoc. In the following, we will focus less on statistical details and more on the underlying conceptual issues that arise when the multiverse is employed as a serious inferential tool.

3.2. The decision on what to include

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 that a multiverse should only vary aspects that are truly arbitrary. If some of the alternatives included are *not* arbitrary, that is, if they are non-equivalent and could be expected a priori to result in different answers, the multiverse may result in "biased estimates." The authors then discuss three types of nonequivalence: measurement nonequivalence (e.g., one way to calculate a composite outcome may result in more valid measurement than another one); effect nonequivalence (e.g., one particular covariate set may target a different effect than another one); and power/precision nonequivalence (e.g., one set of inclusion criteria may result in drastically higher statistical power than another one).

When faced with potential nonequivalence between different ways to analyse the data, Del Giudice and Gangestad suggest that three types of analytic decisions may result. First, the analyst may decide that there are principled reasons to assume equivalence ("Type E decision"). For example, maybe there exists empirical evidence that two ways to calculate the outcome variable result in comparable reliability and validity. Second, the analyst may decide that there are principled reasons to assume nonequivalence ("Type N decision"). For example, an analysis that includes as a covariate a mediator of the focal effect will necessarily target a different effect of interest. Third, the analyst may end up uncertain ("Type U decision"). Two different outcomes may plausibly differ in their validity but we have little idea which one is preferable.

The distinction between the three types of analytic decisions is perfectly sensible on a theoretical level. In practice, the actual decision will strongly hinge on both researchers' domain knowledge and their understanding of statistics. One researcher may conclude that two different outcome measures are *not* equivalent despite empirical evidence that they behave the same in many situations due to certain theoretical assumptions; after all, validity is not an observable metric but a feature of the unobservable measurement process (Borsboom et al. 2004). Another researcher may be unaware of the causal rationale underlying third variable control (Wysocki et al. 2022), thus assuming principled equivalence between different sets of covariates.

Lastly, we might expect that more often than not, researchers in the social sciences end up uncertain about whether two decisions are equivalent or not—after all, there is much that we do not know about the mechanisms that generate our data. In this scenario, Del Giudice and Gangestad suggest exploratory multiverses in which alternatives are examined *separately*,

resulting in a potential multitude of multiverses, increasing the complexity of both the author's and the reader's tasks. In the worst case, in the absence of preregistration, researchers may decide to reduce complexity by determining what is equivalent based on whether or not it supports their result, or whether it is easy to communicate. But even without such strategic selectiveness, researchers may come out at very different assessments of what should and what should not be included.

This bespoke nature of the multiverse—even for a fixed research question, every researcher may end up including different sets of analytic decisions—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 choice is perfectly sensible if the underlying goal is to understand an individual study reported by a specific set of authors. It does, however, turn 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 Giudice 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). Ganslmeier and Vlandas had conducted multiverse analyses pertaining to four topical areas, one of them being welfare state development. In this analysis, they regressed six different measures of welfare state generosity on 18 independent variables representing factors that had been found to be associated with welfare state development. The associational language already suggests that it is somewhat unclear what estimands are being targeted here; each of the independent variables is simultaneously considered part of the "control set" of which all combinations become part of the multiverse.

If the estimands of interest were simple conditional associations, this would result in a multitude of estimands, since there is no reason to expect that the conditional association between two variables remains the same quantity if different third variables are conditioned on. If the estimands of interest were the causal effects of each of the independent variables, randomly varying the included controls would be an incoherent strategy from a causal inference perspective (Wysocki et al., 2022). Auspurg (2025) provided a critical reanalysis, focusing on one clearly defined estimand: the effect of union density on social spending. With this clear target, many of the modeling choices implemented in Ganslmeier and Vlandas' multiverse were no longer justified. And as it turned out, it was precisely those unjustified models that resulted in massive variability in the results, with 60% of the significant coefficients pointing into one direction and 40% pointing into the other direction. In contrast, in the justified models, 97% of the significant coefficients agreed on the direction of the effect.

The Auspurg case turns on which variables should enter the control set; a topic that has received more attention recently in discussions surrounding “bad controls” (e.g., Cinelli et al. 2024). In this spirit, recent simulations on multiverse calibration (Cantone and Tomaselli 2025a expanding on Del Giudice and Gangestad 2021) feature sign reversal due to collider bias which is induced by control for variables that are causally downstream of both the independent and the dependent variable of interest (Elwert and Winship 2014). But even without collider bias, effect nonequivalence can arise.

Consider the case of longitudinal data, in which one may either conduct analyses that control for a baseline measure of the outcome, or calculate change scores and proceed to use these as outcome variable. These two approaches can even lead to estimates of opposite sign, a phenomenon known as Lord’s paradox (Lord 1967; Pearl 2016; Tennant et al. 2023). Nonetheless, a recent multiverse application averages over both approaches. Sorjonen et al. (2026) conduct a multiverse analysis supposedly targeting cross-lagged effects with alternative models—one of them relying on change scores as outcome, another one that is equivalent to baseline adjustment (and a third one that suffers from the aforementioned collider bias). While the authors explicitly note that these specifications answer different questions, they nonetheless combine results from all three meta-analytically, so that the reported quantity becomes an average over distinct estimands rather than an answer to any one question.

In such scenarios of effect nonequivalence, that is, when analyses target different estimands, multiverse analysis 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). There is a certain risk here that both multiverse analysis and the Many Analysts approach lend a guise of rigor and completeness to investigation that in fact would profit from more in-depth theoretical consideration of what research question is supposed to be answered in the first place, and from a clear articulation of the estimand of interest.

But all of this is only an issue if one expects the approach 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 without any awareness of the matter, 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).

Relatedly, researchers may also *deliberately* design a multiverse that combines genuinely distinct estimands—for example, several plausible treatment definitions and several plausible outcomes may be included together. In this case, what the multiverse contributes is largely organizational; it helps lay out all combinations of definitions and ensures they are reported transparently. Of course, additional analytic choices that are considered more arbitrary may also be part of the mix, in which case one may think of the resulting analyses as a nested multiverse containing multiple sub-multiverses, each of which in turn addresses a single distinct but related estimand. In such a scenario, the points we discuss mainly apply to the sub-multiverses; integrating the sub-multiverses into a coherent interpretation results in new challenges on top, as researchers need to discern whether heterogeneity between the sub-multiverses is systematic or possibly a result of mere chance.

Last but not least, a limiting case of the estimands issue arises when the multiverse targets a composite index. Cantone and Tomaselli (2025b) apply a multiverse analysis to formative models of well-being. Such an index has no value defined independently of the modeling choices, because the construct is constituted by them. The variation across specifications therefore reflects almost entirely the open-endedness of the construct, rather than estimation uncertainty for a fixed target. In line with the idea that the what-to-include decision drives apparent variability, Cantone and Tomaselli find that a conceptually distinct subset of specifications–those built on principal component analysis, which import a reflective-measurement logic ill-suited to formative indexes–inflate the spread of results. In contrast, statistical details such as the choice of aggregation function matter little.

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 is already challenging. For example, researchers struggle to properly interpret displays of inferential uncertainty like confidence intervals (Hoekstra et al. 2014), such as by confusing inferential uncertainty with outcome variability (Zhang et al. 2023). A multiverse analysis confronts readers with uncertainty captured by variation among many individual results, each of which come with their own quantified uncertainty. There is little reason to think that misunderstandings will be lessened by the introduction of this less familiar source of variation.

Considering a related issue in how researchers make sense of numbers, 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 and simplistic dichotomous conclusions 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 note 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 because multiverses tend to cover a large range of results, which results in interpretations along the lines of "the quantity of interest could possibly take on substantial negative values, or it could be essentially zero, or it could take on substantial positive values." Given that such non-committal results to substantive questions are rarely satisfying, it is difficult to blame readers (or even authors) for succumbing to the lure of the probabilistic interpretation.[2]

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; see Fig.1C) 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. How should models be weighted?

Another question that arises once one has settled on what to include in the multiverse: how should the different specifications be weighted? This point of contention is illustrated by the debate surrounding the multiverse-style computational framework for multimodel analysis suggested by Young and Holsteen (2017). This debate cuts across multiple concerns previously discussed, as it may be construed either as a statistical debate about inference or a debate about (causal versus predictive) estimands, and as it also touches on concerns about whether results that are presented appropriately communicate uncertainty.

---

[2] One idea to justify the more attractive probabilistic interpretation relies on a Many Analysts design: multiple groups of researchers generate model specifications and the frequency with which a given decision is made serves as an estimate of its probability across the space of defensible analyses (see Breznau et al. 2022 for one example that pairs a Many Analysts approach with supplemental multiverse analysis). However, such empirical calibration gives rise to new open questions, such as the conditions under which analysts' choices may be considered representative of the space of interest. Thus, this approach relocates rather than resolves the illusion of probability.

In this framework, a preferred estimate is supplemented by a multimodel analysis that results in a distribution of estimates across different model specifications. This modeling distribution is considered analogous to the sampling distribution and together, they combine into a total standard error. In analogy to the *t*-statistic, a robustness statistic can be calculated from them to draw inferences about robustness.

Importantly, this inferential framework builds on the raw, unweighted distribution of estimates. This (deliberate) choice was strongly criticized by Slez (2019), who argued that if the multimodel analysis is used for inferential purposes, it turns into a problem of model averaging. For this, in turn, established Bayesian frameworks exist (Slez 2019; Yao et al. 2018). And from this perspective, an unweighted approach cannot be defended; fit-based weighting is preferable (Slez 2019).

A model-averaging approach to multiverse analysis with fit-based weights was subsequently implemented by Semken and Rossell (2022), who developed Bayesian specification curve analysis. As an empirical illustration, they re-analyzed the association between adolescent well-being and digital technology use in the same data as Orben and Przybylski (2019), whose work we discussed in section 2 as an example for the multiverse as a persuasive tool that fails to persuade at least some readers. Broadly in line with Twenge et al.'s specification curve (re-)analysis (2022), Semken and Rossell concluded that the associations were heterogeneous across treatment, gender, and, additionally, depended on whether well-being was assessed by adolescents themselves or their parents.

However, the notion that estimates *should* be weighted has been contested. In his response to Slez (2019), Young (2019) argued that the appropriateness of fit-based weighting depends on the analysis goal—which loops back to the question of estimands (section 3.3). If the goal is mere prediction, model fit is a relevant criterion. But if the goal is to estimate treatment effects, that is, if the estimand is causal, model fit is simply the wrong metric; control variables earn their place not by being correlated with the outcome but by their role in the causal net. In such a scenario, fit-based weighting discounts relevant confounders that are only weakly correlated with the outcome but rewards "bad controls", such as posttreatment variables whose inclusion can improve fit substantially but results in biased effect estimates. As it happens, another critical re-analysis that we discussed before (section 3.3) empirically supports this argument: in Auspurg's (2025) re-analysis of the multiverse analyses conducted by Ganslmeier and Vlandas (2025), a-priori unjustified models were consistently *favored* by a model fit criterion.

Young further argues that, in practice, the fit-based weighted approach championed by Slez (2019) boils down to the selection of the single model with the highest model fit, strongly discounting alternative models. This directly goes against Young and Holsteen's original motivation of transparently showing which underlying model assumptions are necessary to support one's preferred results—and once again highlights that different researchers may use multiverse methods with different purposes on their mind.

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

Despite the many potential issues and controversy, 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.

Additionally, as agentic AI tools make it trivially easy to identify alternative analysis decisions and implement a large space of possible analyses, multiverse analyses can appear to be a necessary format for making sense of empirical results. Early work on autonomous AI analysts documents the potential for researchers to "prompt-hack" analysis agents to produce a desired result (Bertran et al. 2026; Miao et al. 2026), leading to proposals to use agents to vet new results by identifying plausible alternative decisions and summarizing where the reported result falls in the induced distribution (Miao et al. 2026). This approach, however, risks further popularizing the misinterpretations discussed above. Taking a more formal approach, Fishman and Sekeres (2026) promote the necessity of multiverse analysis by modeling automated analysis as reducing the cost of privately searching over correlated pieces of evidence. Imposing more stringent versions of conventional standards, like alpha thresholds, simply induces more search. To maintain effective editorial screening, researchers should therefore instead disclose their search process, at a larger scale than has been typical of sensitivity analyses in social scientific fields.

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 as a serious inferential tool 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 emphasize the need to 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 (Steegen et al. 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). This formula assumes that one observes n independent draws from the population of interest, each classified without error. In real life, these assumptions tend to fail: clustering and other dependencies shrink the effective sample size; misclassified cases add uncertainty; and dropout and nonresponse mean that the people one observes may differ from those one does not observe. Thus, the rule will understate uncertainty. At best, it's a lower bound on error—and sometimes not even that, as misclassification, dropout and nonresponse not only add uncertainty, but also potentially add bias. But it is still a useful baseline, as long as one does not 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 is focused on the narrow goal of getting a "correct" p-value; it accepts the premise that the scientific question is a binary detection problem answered by a thresholded p-value. More broadly, instead of focusing on such a narrow goal, researchers may more usefully invest time and energy into interpreting p-values alongside other considerations (such as related prior evidence, plausibility of mechanism, and real world costs and benefits) that are currently often treated as subordinate to it (McShane et al. 2019).

The multiverse starts from the premise that the analysis could have been different. This distinguishes it from multiple testing correction, which requires one to know 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 needs neither: rather than reconstructing the actual or counterfactual analysis choices, it simply enumerates reasonable choices and carries them out. In this sense, it is not forensic. Taken to its conclusion, this view suggests treating the set of specifications not as a collection of separate results, but as related estimates that could, in principle, be modeled jointly—pointing ultimately toward some form of multilevel model, in which the variability across analytic choices becomes a quantity to be estimated.

While we cannot predict precisely what the future holds for multiverse analysis,we remain convinced that at the very least, it will remain 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
Bertran, M., Fogliato, R., Wu, Z.S.: Many AI analysts, one dataset: Navigating the agentic data science multiverse. Proc. Natl. Acad. Sci. U. S. A. 123, e2606495123 (2026). https://doi.org/10.1073/pnas.2606495123
Borsboom, D., Mellenbergh, G., van Heerden, J.: The concept of validity. Psychol. Rev. 111, 1061–1071 (2004). https://doi.org/10.1037/0033-295X.111.4.1061
Breznau, N., Rinke, E.M., Wuttke, A., Nguyen, H.H.V., Adem, M., Adriaans, J., Alvarez-Benjumea, A., Andersen, H.K., Auer, D., Azevedo, F., Bahnsen, O., Balzer, D., Bauer, G., Bauer, P.C., Baumann, M., Baute, S., Benoit, V., Bernauer, J., Berning, C., Berthold, A., Bethke, F.S., Biegert, T., Blinzler, K., Blumenberg, J.N., Bobzien, L., Bohman, A., Bol, T., Bostic, A., Brzozowska, Z., Burgdorf, K., Burger, K., Busch, K.B., Carlos-Castillo, J., Chan, N., Christmann, P., Connelly, R., Czymara, C.S., Damian, E., Ecker, A., Edelmann, A., Eger, M.A., Ellerbrock, S., Forke, A., Forster, A., Gaasendam, C., Gavras, K., Gayle, V., Gessler, T., Gnambs, T., Godefroidt, A., Grömping, M., Groß, M., Gruber, S., Gummer, T., Hadjar, A., Heisig, J.P., Hellmeier, S., Heyne, S., Hirsch, M., Hjerm, M., Hochman, O., Hövermann, A., Hunger, S., Hunkler, C., Huth, N., Ignácz, Z.S., Jacobs, L., Jacobsen, J., Jaeger, B., Jungkunz, S., Jungmann, N., Kauff, M., Kleinert, M., Klinger, J., Kolb, J.-P., Kołczyńska, M., Kuk, J., Kunißen, K., Kurti Sinatra, D., Langenkamp, A., Lersch, P.M., Löbel, L.-M., Lutscher, P., Mader, M., Madia, J.E., Malancu, N., Maldonado, L., Marahrens, H., Martin, N., Martinez, P., Mayerl, J., Mayorga, O.J., McManus, P., McWagner, K., Meeusen, C., Meierrieks, D., Mellon, J., Merhout, F., Merk, S., Meyer, D., Micheli, L., Mijs, J., Moya, C., Neunhoeffer, M., Nüst, D., Nygård, O., Ochsenfeld, F., Otte, G., Pechenkina, A.O., Prosser, C., Raes, L., Ralston, K., Ramos, M.R., Roets, A., Rogers, J., Ropers, G., Samuel, R., Sand, G., Schachter, A., Schaeffer, M., Schieferdecker, D., Schlueter, E., Schmidt, R., Schmidt, K.M., Schmidt-Catran, A., Schmiedeberg, C., Schneider, J., Schoonvelde, M., Schulte-Cloos, J., Schumann, S., Schunck, R., Schupp, J., Seuring, J., Silber, H., Sleegers, W., Sonntag, N., Staudt, A., Steiber, N., Steiner, N., Sternberg, S., Stiers, D., Stojmenovska, D., Storz, N., Striessnig, E., Stroppe, A.-K., Teltemann, J., Tibajev, A., Tung, B., Vagni, G., Van Assche, J., van der Linden, M., van der Noll, J., Van Hootegem, A., Vogtenhuber, S., Voicu, B., Wagemans, F., Wehl, N., Werner, H., Wiernik, B.M., Winter, F., Wolf, C., Yamada, Y., Zhang, N., Ziller, C., Zins, S., Żółtak, T.: Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. Proc. Natl. Acad. Sci. U. S. A. 119, e2203150119 (2022). https://doi.org/10.1073/pnas.2203150119
Cantone, G.G., Tomaselli, V.: Characterisation and calibration of multiversal methods. Adv. Data Anal. Classif. 19, 989–1021 (2025)(a). https://doi.org/10.1007/s11634-024-00610-9
Cantone, G.G., Tomaselli, V.: On the coherence of composite indexes of well-being: Multiverse analysis for formative models of measurement. Socioecon. Plann. Sci. 102, 102314 (2025)(b). https://doi.org/10.1016/j.seps.2025.102314
Cinelli, C., Forney, A., Pearl, J.: A crash course in good and bad controls. Sociol. Methods Res. 53, 1071–1104 (2024). https://doi.org/10.1177/00491241221099552
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
Elwert, F., Winship, C.: Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable. Annu. Rev. Sociol. 40, 31–53 (2014). https://doi.org/10.1146/annurev-soc-071913-043455
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)
Fishman, N., Sekeres, G.: Editorial Screening when Science is Cheap, https://njw.fish/static/papers/agentic_specification.pdf, (2026)
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
Girardi, P., Vesely, A., Lakens, D., Altoè, G., Pastore, M., Calcagnì, A., Finos, L.: Post-selection Inference in Multiverse Analysis (PIMA): An inferential framework based on the sign flipping score test. Psychometrika. 89, 542–568 (2024). https://doi.org/10.1007/s11336-024-09973-6
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
Hoekstra, R., Morey, R.D., Rouder, J.N., Wagenmakers, E.-J.: Robust misinterpretation of confidence intervals. Psychon. Bull. Rev. 21, 1157–1164 (2014). https://doi.org/10.3758/s13423-013-0572-3
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
Lord, F.M.: A paradox in the interpretation of group comparisons. Psychological bulletin. 68, 304 (1967)
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.: Abandon statistical significance. Am. Stat. 73, 235–245 (2019). https://doi.org/10.1080/00031305.2018.1527253
Miao, J., Pritchard, J.K., Zou, J.: The agentic garden of forking paths, http://arxiv.org/abs/2607.01507, (2026)
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
Palpacuer, C., Hammas, K., Duprez, R., Laviolle, B., Ioannidis, J.P.A., Naudet, F.: Vibration of effects from diverse inclusion/exclusion criteria and analytical choices: 9216 different ways to perform an indirect comparison meta-analysis. BMC Med. 17, 174 (2019). https://doi.org/10.1186/s12916-019-1409-3
Patel, C.J., Burford, B., Ioannidis, J.P.A.: Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. J. Clin. Epidemiol. 68, 1046–1058 (2015). https://doi.org/10.1016/j.jclinepi.2015.05.029
Pearl, J.: Lord's paradox revisited – (oh Lord! Kumbaya!). J. Causal Inference. 4, (2016). https://doi.org/10.1515/jci-2016-0021
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)
Semken, C., Rossell, D.: Specification analysis for technology use and teenager well-being: Statistical validity and a Bayesian proposal. Journal of the Royal Statistical Society Series C: Applied Statistics. 71, 1330–1355 (2022)
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
Slez, A.: The difference between instability and uncertainty: Comment on young and Holsteen (2017). Sociol. Methods Res. 48, 400–430 (2019). https://doi.org/10.1177/0049124117729704
Sorjonen, K., Melin, B., Melin, M., Persson, B.N.: Multiverse analyses can be used to evaluate cross-lagged panel network models: An example with psychological flexibility. Acta Psychol. (Amst.). 264, 106468 (2026). https://doi.org/10.1016/j.actpsy.2026.106468
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
Tennant, P.W.G., Tomova, G.D., Murray, E.J., Arnold, K.F., Fox, M.P., Gilthorpe, M.S.: Lord's 'paradox' explained: the 50-year warning on the use of 'change scores' in observational data, http://arxiv.org/abs/2302.01822, (2023)
Tomaselli, V., Finos, L., Cantone, G.G., Altoè, G.: Call for Papers: The Role of Multiverse Analysis in Statistical Modelling and Applications, https://link.springer.com/journal/10182/updates/27818336, (2026)
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. 224, 103512 (2022). https://doi.org/10.1016/j.actpsy.2022.103512
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
Yao, Y., Vehtari, A., Simpson, D., Gelman, A.: Using stacking to average Bayesian predictive distributions (with discussion). Bayesian Analysis. 13, 917–1003 (2018)
Young, C.: Model uncertainty and the crisis in science. Socius. 4, 237802311773720 (2018). https://doi.org/10.1177/2378023117737206
Young, C.: The difference between causal analysis and predictive models: Response to "comment on young and holsteen (2017)." Sociol. Methods Res. 48, 431–447 (2019). https://doi.org/10.1177/0049124118782542
Young, C., Holsteen, K.: Model Uncertainty and Robustness: A Computational Framework for Multimodel Analysis. Sociol. Methods Res. 46, 3–40 (2017). https://doi.org/10.1177/0049124115610347
Zhang, S., Heck, P.R., Meyer, M.N., Chabris, C.F., Goldstein, D.G., Hofman, J.M.: An illusion of predictability in scientific results: Even experts confuse inferential uncertainty and outcome variability. Proc. Natl. Acad. Sci. U. S. A. 120, e2302491120 (2023). https://doi.org/10.1073/pnas.2302491120