International Journal of Psychology

International Journal of Psychology, 2024
DOI: 10.1002/ijop.13229

The multiverse of universes: A tutorial to plan, execute and interpret multiverses analyses using the R package *multiverse*

Martin Götz[1], Abhraneel Sarma[2], and Ernest H. O'Boyle[3]

[1]University of Zurich, Zürich, Switzerland
[2]Northwestern University, Evanston, IL, USA
[3]Indiana University, Bloomington, IN, USA

Even when guided by strong theories and sound methods, researchers must often choose a singular course of action from multiple viable alternatives. Regardless of the choice, it, along with all other choices made during the research process, individually and collectively affects study results, often in unpredictable ways. The inability to disentangle how much of an observed effect is attributable to the phenomenon of interest, and how much is attributable to what have come to be known as *researcher degrees of freedom* (RDF), slows theoretical progress and stymies practical implementation. However, if one could examine the results from a particular set of RDF (known as a *universe*) against a systematically and comprehensively determined background of alternative viable universes (known as a *multiverse*), then the effects of RDF can be directly examined to provide greater context and clarity to future researchers, and greater confidence in the recommendations to practitioners. This tutorial demonstrates a means to map result variability directly and efficiently, and empirically investigate RDF impact on conclusions via *multiverse analysis*. Using the R package *multiverse*, we outline best practices in planning, executing and interpreting of multiverse analyses.

Keywords: Multiverse analysis; Specification curve analysis; Researcher degrees of freedom; Transparency; Robustness.

> Everything affects everything else, and you have to understand that whole web of connections.—M. Mitchell Waldrop (1992, p. 60 f.)

Simmons et al. (2011) coined the term *researcher degrees of freedom* (RDF) to describe the tremendous variety of equally defensible and theoretically justifiable choices at each stage of the research process (see also Breznau et al., 2022; Gelman & Loken, 2013; Götz & O'Boyle, 2023). RDF introduce variance in the means of conducting research even when the goal is constant. That is, two or more researchers starting from the same origin (i.e., shared substantive question) and restricted to viable means of transport only (i.e., theoretically or empirically justified choices) can nevertheless end their journey at vastly different places (i.e., study findings and conclusions) due to RDF. Furthermore, RDF lead to multiple comparisons problems that eventually inflate false-positive rates, thereby obfuscating the actual robustness of a scientific claim, as well as the resulting body of knowledge in the case of its publication (see also Gelman & Loken, 2013; Götz & O'Boyle, 2023; Steegen et al., 2016).

Because of the multiplicative nature of choices at each step of the research process, individual RDF that collectively lead to drastically divergent conclusions do not need to be drastically divergent themselves, such as choosing between qualitative and quantitative methods. Rather, RDF are often ostensibly small discrepancies in the research process, such as one researcher measuring a demographic control variable using categories (e.g., age bands), while another asks for an exact value. These similar RDF at various decision points can individually and collectively affect distal outcomes up to and including the overall retention or rejection of the theory, model, intervention, etc. under investigation. This is a critical

Correspondence should be addressed to Martin Götz, Department of Psychology, Social and Economic Psychology, University of Zurich, Binzmühlestrasse 14/Box 13, CH-8050 Zürich, Switzerland. (E-mail: m.goetz@psychologie.uzh.ch).

We thank Katharina Reher and Jan B. Schmutz for their valuable feedback on earlier versions of this manuscript.

concept, as the statistical technique we demonstrate is simply a means of aggregating and contextualising the results from one collective set of RDF (i.e., a *universe*) from a broader set resulting from the combination of similarly viable RDF (i.e., a *multiverse*).

A multiverse is a finite space containing alternative RDF that, if chosen, would be somehow defensible for a given research effort. For example, if measuring age using a range of ages contradicted theory and norms in say, developmental psychology, it would not be a viable RDF for developmental psychologists. Therefore, any pathways or universes that contained this choice would be excluded from the multiverse. Yet, even when limited to viable universes of minor divergences, a multiverse spanning thousands, if not millions, of unique universes is likely to emerge. For example, a researcher making only three decisions (e.g., outlier treatment, control variable inclusion, outcome variable operationalisation) with four options per decision (e.g., outlier removal, winsorisation, log transformation, robust estimation) will create 4^3 or 64 universes. A fourth decision option would create 256 universes, and a fifth option per decision would generate well over 1000 universes, each potentially adhering to theory, extant literature and research norms.

Despite the variability caused by divergent decisions, researchers typically present only one universe from their actual multiverse. Thus, even highly competent and experienced researchers working with the best of intentions may inadvertently choose and report a universe of perfectly reasonable decisions that nevertheless does not reflect how the effect or phenomenon unfolds in the population, or how the effect would present itself if a different path had been chosen (e.g., Gelman & Loken, 2013; Götz & O'Boyle, 2023; Steegen et al., 2016). Even when researchers attempt to contextualise their preferred universe by presenting one or more alternative universes as robustness checks, the process is non-systematic, far from exhaustive and vulnerable to gamesmanship (e.g., only presenting alternative universes consistent with the preferred universe).

To help researchers uncover, and reason about, the extent their results are based upon decisions in the data analytic process, we present a tutorial on planning, executing and interpreting a multiverse analysis. Our aims are threefold: First, we provide an overview of the theoretical, methodological and pragmatic aspects involved in deriving a sensible set of universes. Second, we demonstrate the versatility of multiverse analysis and ease-of-use of the R package *multiverse* (Sarma et al., 2023), following a recently published study by Willroth et al. (2021). Third, we consider how the vast number of results stemming from a multiverse analysis can be interpreted. We close with a brief discussion of the promises and perils of multiverse analyses.

UNDERSTANDING MULTIVERSE ANALYSIS

Most (social) scientists address research questions with one or a small number of statistical analyses to derive empirical results (e.g., Del Giudice & Gangestad, 2021; Ioannidis, 2008; Simonsohn et al., 2020). Yet, this only represents an excerpt of all defensible statistical analyses and specifications that the researchers could have conducted and presented (e.g., Gelman & Loken, 2013; Götz & O'Boyle, 2023; Steegen et al., 2016). Meta-scientific research has found significant variability introduced at every step of the research cycle. For example, Landy et al. (2020) provided research teams with the same data and asked them to design studies (i.e., materials) to empirically test five hypotheses, only to find stark variation in statistical results with positive, negative, or null effects. The divergent findings were entirely attributable to justifiable but divergent decisions in the design and execution of the research (e.g., Breznau et al., 2022; Schweinsberg et al., 2021; Silberzahn et al., 2018). Similar conclusions were derived from so-called *Many Labs* (e.g., Ebersole et al., 2020) and *Many Analysts* (e.g., Botvinik-Nezer et al., 2020) projects, which found large variations in statistical results and consequent inferences, even when analysing the same datasets and assessing the same hypotheses. We echo Wagenmakers et al. (2022) and others who call for more *Many Analysts*-like projects to allow for more robust science, but we also recognise that these projects are extremely labour- and time-intensive; here, *multiverse analysis* is an efficient and effective means to address the variability introduced by RDF.

The overarching idea of multiverse is simple: A multitude of decisions are made by researchers when studying a phenomenon empirically, with each potentially affecting the results and conclusions.[1] Rather than limiting oneself to reporting a single result based on a single set of defensible, but critically, not the *only* defensible set of decisions, a researcher could transparently analyse and report the entire set of decisions. Multiverse analyses have already been applied to a range of research projects, such as individual-level studies, meta-analyses and multi-team projects (e.g., Breznau et al., 2022; Olsson-Collentine et al., 2023; Singmann et al., 2024). However, regardless of the research question, substantive area or field, or analytic approach, multiverse analysis can efficiently model and aid the interpretation of RDF, and contextualise one chosen universe by giving it the backdrop of alternative universes. In doing so, a multiverse analysis can help researchers gain "a sense of the sensitivity of

[1] In close resemblance to a multiverse analysis, Ioannidis (2008; see also Klau et al., 2023) proposed the *vibration of effects* framework, Young and Holsteen (2017) suggested the *multimodel analysis* and Simonsohn et al. (2020) proposed the *specification curve analysis*.

TABLE 1
An exemplary plan for a multiverse analysis

Parameter	*Parameter #*	*Option #*	*Option*	*Specifics*
Data	1	1	**MIDJA, T1 + T2**	—
Outlier treatment	2	1	**Include all**	—
		2	Winsorise	Winsorise continuous variables at 99%
Transformation	3	1	**Standardise**	*z*-transform continuous variables
		2	Raw	—
Outcome	4	1	**General health, T2**	—
Predictor	5	1	**Change in sense of purpose, T2-T1**	Reliable change index (RCI)
Controls	6	1	**Sense of purpose, T1 (Level)**	—
		2	Exclude	—
	7	1	**General health, T1**	—
		2	Exclude	—
	7	1	Multimorbidity, T1	—
		2	Exclude	—
	9	1	**Gender, T1**	—
		2	Exclude	—
	10	1	**Age, T1**	—
		2	Exclude	—
	11	1	**Education, T1**	—
		2	Exclude	—
Analytic procedure	12	1	**Linear regression**	OLS estimation
		2	Ordinal regression	CLM estimation

Note: Options printed in bold represent original decisions by Willroth et al. (2021), whereas the other options represent alternatives. CLM = cumulative link model; MIDJA = survey of midlife in Japan; OLS = ordinary least squares; T1 = first measurement, T2 = second measurement.

conclusions to [potentially] arbitrary decisions in data preparation and thus of the fragility or robustness of a claimed effect" (Steegen et al., 2016, p. 710). In summary, a well-conceived multiverse analysis (a) considers viable alternative specifications of one's analytic approach to a research question, (b) computationally analyses all resulting universes simultaneously and (c) interprets the resulting informational (over)load.

PLANNING A MULTIVERSE ANALYSIS

As with any scientific method, a carefully crafted plan that translates *conceptual research questions* into *empirical research questions* is the foundation for any multiverse analysis. In the hypothetico-deductive framework, theoretical considerations must guide the formulation of hypotheses, and the respective construction of statistical models to assess them. Thus, an informed theoretical grasp of the phenomenon (e.g., well-being), its respective drivers (e.g., stress) and the population to generalise to (e.g., employees) is a prerequisite. Given that there are a multitude of definitions, conceptualizations and considerations of most phenomena, it is imperative to consider and acknowledge potential alterations in the construction of a multiverse. Thus, profound knowledge of the existing relevant literature is a prerequisite (e.g., Hanel & Zarzeczna, 2023) to avoid considering flawed pathways that could hide or mask the results of interest. More plainly, multiverse analysis does not replace good theory and methodological rigour; it requires them.

Constructing a multiverse first tasks researchers with the identification of relevant decisions along the research cycle that might induce variability in the results (see also Del Giudice & Gangestad, 2021). Each of these decisions, in other words, *parameters*, can entail *options* between two or more equally defensible choices; these are the RDF. Thus, the parameters are the crossroads, and the options are whether to go left, right or straight ahead. To illustrate the use of multiverse analysis, we followed Willroth et al.'s (2021) empirical analysis.[2] These authors investigated whether individuals' perceived sense of purpose in life affected their general health. We created a multiverse that holds the original model specifications or RDF, as well as potential alterations that will be considered as part of a subsequent multiverse analysis. These alterations are shown in Table 1, with the parameters and respective options for each; collectively, they make up the entirety of the universes, the multiverse.

[2] We stress that we selected this work solely for illustrative purposes and not to critique these authors' RDF in any way. The principal reason for selecting Willroth et al. (2021) over others is how commendably they transparently reported data analytic choices, and how accessible their R code was (https://osf.io/brcen/).

Data preparation (Parameters 1–3)

The first parameter to consider is the data source. Naturally, a multitude of options exists here, and aspects of representativeness and generalizability should be the guiding principles (see also Harder, 2020). Beyond standard data collection best practices, the data must not only allow answering the research question using one universe, but also lend itself to modelling alternative RDF. For example, if an RDF is whether to include age in the model, then to test the "age included" option for this parameter, age needs to have been measured. For Parameter 1, consistent with Willroth et al. (2021), we used openly available datasets from the Survey of Midlife in Japan (i.e., MIDJA1 and MIDJA2; Ryff et al., 2011, 2016). Across two measurement waves, the dataset contains the responses of 657 adults regarding sociodemographic (e.g., age), mental health (e.g., sense of purpose) and physical health (e.g., global health) characteristics.

The second parameter is the treatment of outliers. Researchers may choose from multiple options that are empirically or theoretically justified, such as winsorising, or deleting extreme cases (see also Villanova, 2023). While Willroth et al. (2021) did not report any outlier treatment, we winsorised the continuous variables at 99% as a viable option, having inspected their respective univariate distributions. Collectively, these two options constituted our Parameter 2.

The third parameter for data preparation is the potential transformation of data. For example, if there is a significant skew in one or more variables, researchers may apply log transformation, take the square root, or switch to an estimator that is more robust to skewness. Willroth et al. (2021) presented results from standardised (i.e., *z*-transformed) variables in their analyses in lieu of unstandardized ones. For Parameter 3, we modelled both options.

Model building (Parameters 4–11)

The fourth and fifth parameters in Table 1 deal with the operationalisation of the outcomes (Parameter 4) and predictors (Parameter 5). Here, researchers can choose from several potentially theoretically relevant operationalisations to consider; however, this is a key factor in any empirical study. For example, if there is tension in the field about which of the three broad classifications or streams of emotional intelligence measures best capture the construct (e.g., Daus & Ashkanasy, 2005), then this may require alternative measures of emotional intelligence built into the research design as options for this fourth parameter. Similar considerations are required for the outcome variable(s) at Parameter 5. We included only one of Willroth et al.'s (2021) outcomes (i.e., general health) to keep the multiverse internally comparable (Parameter 4), and we kept the main predictor (i.e., change in sense of purpose over time) fixed across all universes (Parameter 5).

The remaining parameters at this step (Parameters 6–11) contain the RDF for the inclusion of control variables. Often, there is less emphasis placed on the measurement or inclusion of control variables (see also Hünermund & Louw, 2023), but in the case of multiverse, researchers must make their rationale regarding control variable inclusion or exclusion explicit to assess the variability of this decision on results. While Willroth et al. (2021) included a fixed set of control variables in their statistical models, we created include/exclude options for all of them. In addition, for the illustrative purposes of this tutorial, instead of modelling multimorbidity (T2) as an outcome variable (Willroth et al., 2021), we decided to include or exclude it as an additional predictor at T1 (Parameters 6–11).

Model analysis (Parameter 12)

Finally, once all these parameters are specified, researchers must map the suitable statistical approaches to run their analyses. Again, a multitude of options quickly emerges in that data can be analysed in a manifest or latent way, using a plethora of statistical estimators that each have different assumptions, and so on (see also Gelman et al., 2020). Willroth et al. (2021) used multiple linear regression and ordinal regression; thus, we also included both options (Parameter 12).

Overall, the product of all options per parameter gives the number of universes in a multiverse; in our case, spanning 512 single universes. This is far from exhaustive, but this sheer—generally exponentially growing—number provides a glimpse into the multiplicity of RDF in a typical case, and how quickly such analyses can become convoluted and open the door to dust bowl empiricism. Yet, it is precisely a well-conceived multiverse analysis that can help researchers acknowledge this hidden variability, and allow them to embrace it (see also Gelman & Loken, 2013; Götz & O'Boyle, 2023; Steegen et al., 2016).

EXECUTING A MULTIVERSE ANALYSIS

With the multiverse assembled, we move on to its implementation using the *multiverse* package (Sarma et al., 2023) in R (R Development Core Team, 2024). Notably, other software for multiverse-style analyses exists, such as Boba (Liu et al., 2021; for a high-level comparison, see Sarma et al., 2023). Because of its ease of use, dependence on R and versatility in using the manifold packages available in R, we focus only on the *multiverse* package.

General overview of the *multiverse* package

The *multiverse* package is designed to closely follow the paradigms prevalent in R, and to reduce the amount of additional code that users are required to write to declare alternatives in a multiverse analysis.[3] *Multiverse* allows users to proceed stepwise through their analysis and to declare alternatives at any step. Users can declare alternative analysis paths by replacing any sub-expression[4] of the R code with a set of alternative sub-expressions. This is made possible in *multiverse* by extending the R syntax—code corresponding to different alternative analysis paths is declared using a special *branch()*-function.

In addition, analyses are declared within a dedicated execution environment, known as a multiverse object. Thus, users must declare a new multiverse object variable at the beginning of their analysis. When implementing a multiverse using R Markdown (Xie et al., 2020), the code must be specified using multiverse code blocks (as opposed to R code blocks). When using R scripts, the code must be declared within an *inside()*-function. Unlike R, which executes any code declared, the code declared within a multiverse object is not executed immediately. Instead, the *multiverse* R package captures the declared code and calculates the multiverse of analysis paths by combining each alternative at every step of the analysis. *Multiverse* employs delayed execution to ensure faster computation, as multiverse analyses can quickly grow exponentially, executing the code at every step could end up being computationally expensive.

Executing Table 1

First, one must load the required data and packages, such as the *multiverse* package itself (Sarma et al., 2023). Preparatory data wrangling steps, such as merging datasets, renaming variables, and checking descriptive statistics, can easily be accomplished in base R or by using the *tidyverse* package (Wickham et al., 2023). Note that we provided both, data and R code, online to follow our tutorial in detail: https://osf.io/p5gtj/.

Next, users must create a multiverse object that holds all the specifications of the multiverse before its actual execution; here, we simply called this multiverse object *M* (Figure 1a). Using the *inside()*-function, *M* will subsequently be manipulated to eventually hold all specifications that comprise the multiverse (Figure 1b). First, we need to define the dataset as a variable in *M*. Users can then turn to subsequently creating actual parameters and options to define the specifications in their multiverse; here, the *branch()*–function is our workhorse. For instance, two alternatives for outlier treatment—all continuous variables in either winsorised, or raw form—can be declared using *branch()* (Figure 1c). To do so, we manipulate *M* to hold our original dataset (i.e., *data.formatted* that we simply renamed *df*), and then declare the specifications of outlier handling using the *branch()*-function and its arguments: The first argument to *branch()* indicates a parameter (i.e., decision point) that can hold two or more options, which are denoted as subsequent arguments. Hereby, the logic is always the same: (a) Take the indicated dataset, (b) create and name a parameter (here, outlier treatment is named *outlier_exclusion*) using the *branch()*-function and (c) consider and name the following options using the respective functions from the plethora of R packages available (here, winsorise all continuous variables at 1/99% and name the option *winsorise*, or leave them in their raw form and name the option *no_exclusion*). A parameter specified in such a way is then stored in *M*, which can be manipulated further until it holds all specifications (i.e., decisions and options) that collectively establish one's multiverse.

Applying this logic to formulate a parameter also allows to branch the standardisation of the respective variables and note it in the multiverse object *M*. Turning to the analytic approaches as well as the consideration of covariates simultaneously, users can now create their model branches, as they would do when performing a single universe analysis in R. Thus, we created an analysis branch (Figure 1d) spanning the OLS linear regression, using base R, as well as the alternative ordinal regression, using the R package *ordinal* (Christensen, 2023). Again, a parameter can hold more than two options; therefore, including additional analytic approaches is easily achievable. Next, the branches to toggle the inclusion of covariates in the respective models are created (Figure 1e). Upon programming all specifications, *M* must be extended by imputing a *summary()*-function to extract the results from the specified models. Notably, users can use the *expand()*-function at each step of creating parameters and corresponding options to inspect the combinations of specifications—in other words, universes—created thus far.

Having ensured that the number of universes within the multiverse in R matches the number of universes manually calculated from Table 1, the multiverse can be executed using the *execute_multiverse()*-function. For testing and debugging purposes, subsets of the multiverse (e.g., the first 10 universes) can also be exe-

[3] The GitHub repository for the *multiverse* package provides recent developments, bug fixes, and detailed tutorials: https://github.com/MUCollective/multiverse. The repository contains additional examples of how to create and run multiverse analyses for studies by Durante et al. (2013) (see also Steegen et al., 2016) and Jung et al. (2014), thereby also covering more advanced topics, such as interactions, and trimming a multiverse by exclusion of nonsensical universes.

[4] In programming, an expression is a code (in any language) that can be evaluated to determine a value.

a. The first step is to declare the multiverse object, a dedicated execution environment for multiverse analysis

b. Code for multiverse analysis should be declared using the `inside()` function. This specifies that the code should be executed in the dedicated multiverse environment.

c. The `branch()` function is used to declare decisions. First argument is the name of the parameter. Alternate analysis are passed as arguments in the form `options_name ~ option_value`

d. `branch(es)` can be declared flexibly. Implementing a linear and ordinal regression requires different functions (`lm` vs `clm`) as well as "types" of the dependent variable (`numeric` vs `factor`). As they represent the same conceptual decision, they have the same parameter name (`estimator`).

Note: this is not the only way to implement this decision using the `multiverse` library. The supplementary materials depict alternative ways to implement this decision.

e. Decisions on whether to include or exclude a covariates to a model `branch(es)` requires a workaround using NULL as shown here.

Note: uncertainty whether a covariate should be included or excluded represents insufficient theoretical understanding of the data-generating process. While it is reasonable to include such exploratory decisions in a multiverse, if the outcome is indeed sensitive to such decisions, it is imperative to determine why it is so (see Del Giudice & Gangestad 2021).

```r
M = multiverse

inside(M, {
  df = data.formatted |>
    mutate_if(is.numeric, ~ branch(outlier_exclusion,
      "winsorize" ~ datawizard::winsorize(., threshold = 0.01),
      "no_exclusion" ~ identity(.)
  ))
})

...

inside(M, {
  model_fun = branch(estimator, "linear" ~ lm, "ordinal" ~ clm)
  dep_variable = branch(estimator,
    "linear" ~ df$Health_T2,
    "ordinal" ~ as.factor(df$Health_T2))

  model.fit = model_fun(dep_variable ~ Purpose_RCI
                        + branch(purpose_ctrl,
                                 "include" ~ Purpose_T1,
                                 "exclude" ~ NULL)
                        + branch(health_ctrl,
                                 "include" ~ Health_T1,
                                 "exclude" ~ NULL)
                        + branch(chronic_ctrl,
                                 "include" ~ Multimorbidity_T1,
                                 "exclude" ~ NULL)
                        + branch(sex_ctrl,
                                 "include" ~ Sex,
                                 "exclude" ~ NULL)
                        + branch(age_ctrl,
                                 "include" ~ Age,
                                 "exclude" ~ NULL)
                        + branch(education_ctrl,
                                 "include" ~ Education,
                                 "exclude" ~ NULL),
                        data = df)
})
```

Figure 1. Annotated exemplary code to use multiverse.

cuted using the *execute_universe()*-function. Depending on the size of the multiverse, the available computational resources and the complexity of the analytic strategies, this step can take an extensive amount of time; in our case, computation of the outlined multiverse of 512 universes took less than 1 minute on a laptop. Nonetheless, we would like to caution users that a larger and more complex multiverse can quickly result in a significant growth in demand for computation resources and/or execution times (e.g., Bayesian estimation, complex hierarchical data structures and non-normal response functions).

Upon execution of the multiverse, one can extract the actual results from *M* for further processing. In our example, change in the sense of purpose regarding general health was the most relevant estimate to the research question (i.e., a substantive criterion in Willroth et al.'s [2021] work). Thus, we extracted these specific estimates and split them by (a) standardisation and (b) analytic procedure to allow for their respective comparability (see Figure 2, Table 2). Specifically, regarding the analytic procedure, the coefficients of the linear model describe linear changes, whereas the coefficients of the ordinal model describe changes in cumulative log odds.

THE INTERPRETATION

Researchers are well-trained to interpret a single statistical test associated with a single model specification, but multiverse will inundate them with hundreds, thousands or even millions of individual results that require data aggregation and interpretation, and drawing conclusions for future research and practice. Depending on the researchers' goals, three broad approaches exist to make sense of multiverse results (Sarma et al., 2024): (a) Assessing the overall variability of the parameters of interest across the multiverse; (b) identifying the sources of variability, if any; and (c) drawing conclusions regarding the overall effect being studied.

First, we advocate that researchers use multiverse as a tool to map variability in results (see also Hall et al., 2022; Sarma et al., 2023). Here, histograms may be used as an initial visualisation to depict the descriptive statistics of the coefficients of interest (Figure 2). If the histograms show a wide range of estimates, this suggests that the effect is susceptible to choices in the data analysis process. The specification curve plot (SCP), as described by Simonsohn et al. (2020), helps identify the specific decisions that affect study outcomes. In simple terms, a SCP allows researchers to visually inspect the specifications that appear most interesting in producing the

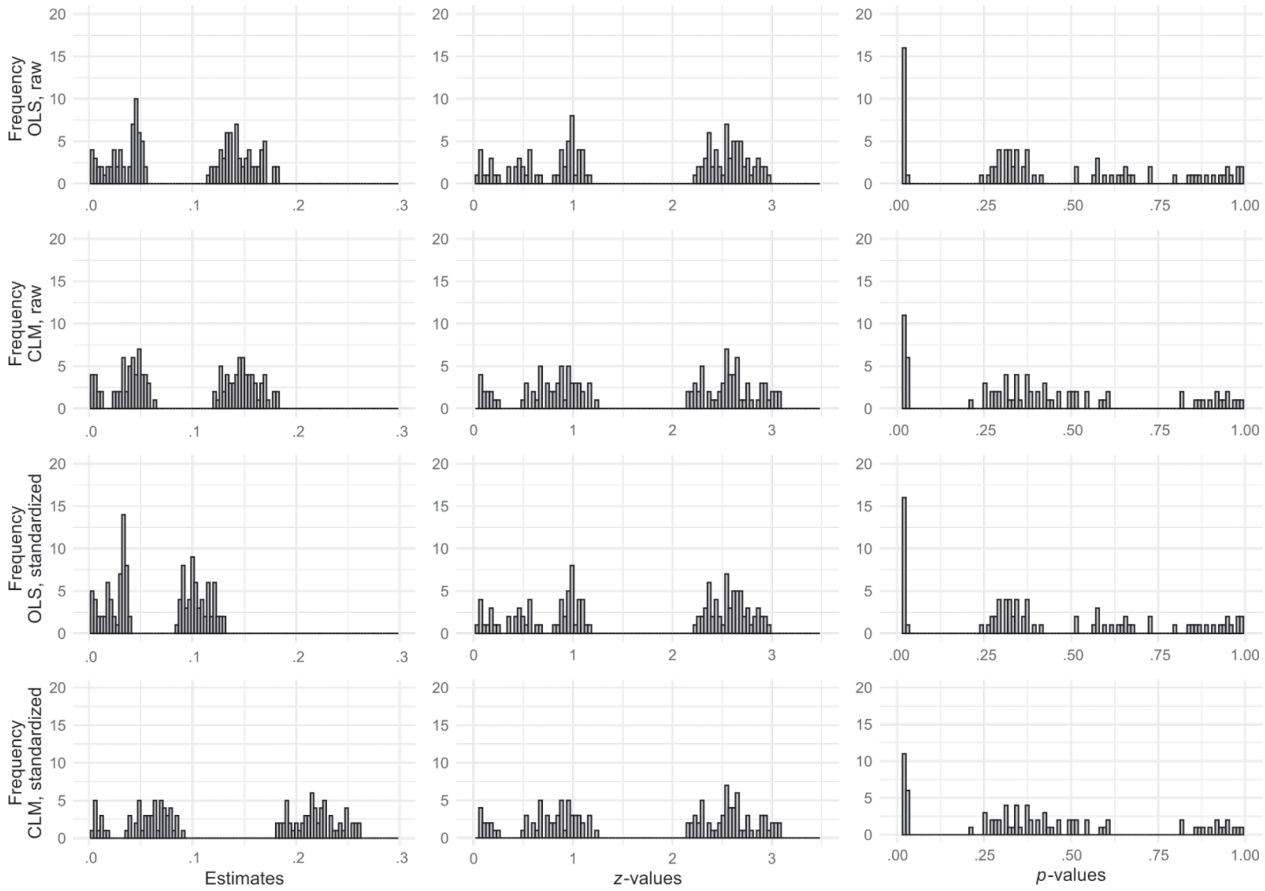

Figure 2. Histograms for the focal coefficient of interest. *Note:* Histograms of the descriptive statistics of the regression coefficients for change in the sense of purpose regarding general health (T2) across 512 universes, split by standardisation and analytic approach; each row represents the respective set of 128 universes. CLM = cumulative link model; OLS = ordinary least squares.

TABLE 2
Descriptive statistics for the focal coefficient of interest

				95% CI	
	Estimate	*SE*	p	*Lower*	*Upper*
OLS, unstandardised					
M	0.09	0.05	.28	−0.02	0.19
SD	0.06	0.00	.32	0.05	0.07
Range	0.18	0.02	.99	0.17	0.20
OLS, standardised					
M	0.06	0.04	.28	−0.01	0.14
SD	0.04	0.00	.32	0.04	0.05
Range	0.13	0.01	.99	0.12	0.14
CLM, unstandardised					
M	0.09	0.05	.27	−0.01	0.20
SD	0.06	0.00	.31	0.05	0.07
Range	0.18	0.01	1.00	0.16	0.20
CLM, standardised					
M	0.13	0.08	.27	−0.02	0.29
SD	0.09	0.01	.31	0.08	0.10
Range	0.26	0.01	1.00	0.24	0.29

Note: Descriptive statistics for the estimate of change in sense of purpose regarding general health (T2) across all 512 universes in the entire multiverse, split by standardisation and estimation. CLM = cumulative link model; OLS = ordinary least squares.

results, thereby representing two approaches to make sense of a multiverse simultaneously. The top panel displays the coefficient of interest, ordered by magnitude and its colour coding represents its statistical significance. Here, we chose the common $p < .05$-level, but any other threshold might be used (see also Benjamin et al., 2017; Benjamini & Hochberg, 1995). The lower panel shows the corresponding specifications regarding sources of variability (i.e., the specific set of decisions that created that specific universe and coefficient of interest).

The SCP associated with the regression coefficient of change in sense of purpose regarding general health (T2) in our multiverse analysis is presented in Figure 3. We find that regardless of the specifications of the individual universes, this coefficient is positive but only statistically significant in about half of the universes. Closer inspection reveals that when the change in sense of purpose regarding the general health coefficient is only statistically significant in those universes where one specific variable (i.e., sense of purpose) is included as a predictor in the regression model. Put differently, the inclusion or exclusion of the sense of purpose at T1 singularly determines whether the actual coefficient of interest and presumably associated hypothesis is supported.

Second, for multiverses spanning hundreds, or more, universes, static visualisations, such as the SCP, may no longer be feasible. Interactive visualisation systems, such as *Milliways* (Sarma et al., 2024) and *Boba Vizualizer* (Liu et al., 2021, 2023), allow users to explore the results of such large multiverse analyses. Both systems allow users to validate the construction of a multiverse. Boba (Liu et al., 2021, 2023) proposed a metric-based approach (e.g., RMSE) to evaluate the quality of individual universes. In contrast, Milliways (Sarma et al., 2024) emphasised the need to evaluate universes based on domain knowledge and statistical expertise—"*why is the outcome sensitive to the choices of a decision, and are these choices equally, theoretically, justifiable?*" In other words, such an evaluation of a multiverse may even highlight gaps in the existing theoretical understanding of the phenomena. Alternatively, additional analytic approaches may be used to obtain a better understanding of the sources of the observed variability, such as the post-selection inference approach (Girardi et al., 2024), regression analyses and relative importance approaches (e.g., Grömping, 2006), functional analyses of variance (e.g., Centofanti et al., 2023; Górecki & Smaga, 2019) and cross-validation approaches (e.g., De Rooij & Weeda, 2020; Yarkoni & Westfall, 2017). However, we stress that researchers should exercise caution in turning to these approaches as they eventually reduce the just embraced variability to overly simplistic decisions (e.g., binary informed by statistical significance; cf. Wasserstein & Lazar, 2016).

Finally, these steps collectively allow researchers to draw conclusions regarding the overall effect being studied. A first take away is whether there is substantial variability with regards to a focal association of interest owing due to (potentially) arbitrary RDF. Once a reasonably constructed multiverse illustrates the potential variability in the results, and its potential sources are identified, researchers must consider, on theoretical and methodological grounds, whether results from specific universes, or sets thereof, can answer the original research question satisfactorily. Thus, an explanation of why the coefficients of change in sense of purpose regarding general health are statistically significant only, if one includes sense of purpose (T1), would have to draw on domain expertise to eventually identify the *correct* universe.

DISCUSSION

Variability in the results of scientific studies is inherent because of RDF (e.g., Botvinik-Nezer et al., 2020; Ebersole et al., 2020; Landy et al., 2020). Multiverse is a valuable, efficient and powerful means to illustrate and contextualise RDF in order to partition a given result (universe) into that which is attributable to what is being researched (i.e., phenomenon) from who is doing the research. We presented a brief tutorial on constructing, running and interpreting a multiverse.[5] We now close with a discussion of the critical considerations when embarking on one's own multiverse, as well as future applications and developments for multiverse analyses.

Theory and methodology first, multiverse second!

Multiverse analysis is often positioned as an exploratory data analysis tool (Sarma et al., 2023), but theoretical considerations must guide its construction to avoid losing the needle in the haystack (Del Giudice & Gangestad, 2021). In other words, the inclusion of flawed parameters and/or options leads to an exponentially large and uninterpretable hodgepodge of viable universes, alongside confounded ones. Alternatively, and equally problematic, weak theory or a poor grasp of the literature can lead to exceptionally large multiverses that are nevertheless anaemic because they have omitted critical options at key parameters. If contaminated by non-viable universes and/or deficient due to the lack of viable ones, a multiverse is just as likely to mislead and confuse as it is to inform and illuminate.

[5] Again, we would like to point interested readers to the vast online resources for the *multiverse* package provided on GitHub: https://github.com/MUCollective/multiverse.

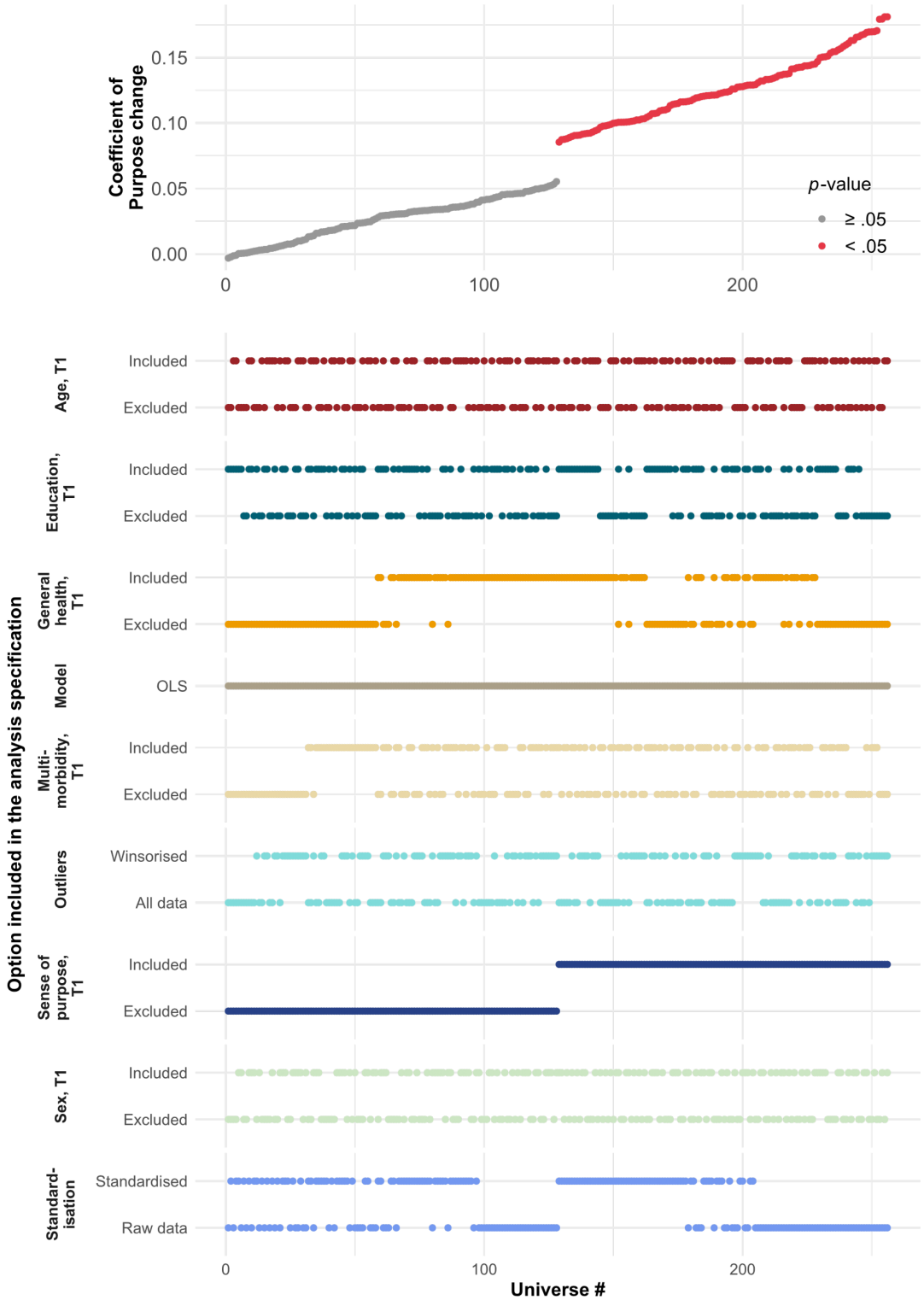

Figure 3. A specification curve plot of the exemplary multiverse. *Note:* Results from 256 universes in the multiverse where ordinary least squares (OLS) regression was used. Ideally, this plot should be inspected column by column. We provide a larger figure, as well as additional ones for the cumulative link models (CLM) online: https://osf.io/p5gtj/.

We also note that multiverse can be considered a tool that can help with the crafting (and pruning) of theory (see also Leavitt et al., 2010). As variability is inherent to the research process, multiverse is a formidable tool to embrace it and eventually arrive at a shared mental model for research in a certain topical space (Gelman & Loken, 2013; Steegen et al., 2016). For example, scientific debates on the existence (and importance) of construct proliferation, (mis)usage of control variables and correct choices of analytic procedures (see also Götz & O'Boyle, 2023) could be advanced by spanning respective multiverses over research questions to help identify meaningful chunks of variability, inquire them and resolve them for future research.

A related point comes in the form of the oft-cited mantra, "It is better to design around than analyze through." That is, the nature of the design dictates the strength of the inference. For example, while a multitude of RDF (e.g., decisions on operationalisations of constructs, transformation of variables and their inclusion/exclusion) can induce variability in the results of scientific endeavours, no statistical analysis—multiverse-assisted or otherwise—can directly speak to causality if the underlying study design does not allow for it. Relatedly, multiverse is also no panacea to RDF lurking around the data-collection stage of a study, such as considering or omitting competing operationalisations of a construct of interest (see also Harder, 2020).

CONCLUSION

Variability in the conduct of research is inherent; at every step of the way, researchers must make decisions in the form of choosing data sources, construct operationalisations, variable transformations, etc. This variability is not only an expression of creativity in tackling research questions, but also an important source of variability in the results of statistical analyses. This variability must be explicitly acknowledged, modelled and investigated to allow for a transparent, holistic and overall informative science. In providing our tutorial on how to plan, execute and interpret a multiverse analysis, we hope to have supplied fellow researchers with an understanding of, and a pragmatic how-to on appreciating variability via the multiverse, and taking lessons from it.

ACKNOWLEDGEMENT

Open access funding provided by Universitat Zurich.

Manuscript received January 2024
Revised manuscript accepted June 2024

REFERENCES

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., … Johnson, V. E. (2017). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, *57*(1), 289–300. https://doi.org/10.2307/2346101

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., Avesani, P., Baczkowski, B. M., Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., … Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, *582*(7810), 84–88. https://doi.org/10.1038/s41586-020-2314-9

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., … Żółtak, T. (2022). Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proceedings of the National Academy of Sciences*, *119*(44), e2203150119. https://doi.org/10.1073/pnas.2203150119

Centofanti, F., Colosimo, B. M., Grasso, M. L., Menafoglio, A., Palumbo, B., & Vantini, S. (2023). Robust functional ANOVA with application to additive manufacturing. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *72*(5), 1210–1234. https://doi.org/10.1093/jrsssc/qlad074

Christensen, R. H. B. (2023). *Package 'ordinal'* (2023.12-4) [R]. https://cran.r-project.org/web/packages/ordinal/ordinal.pdf

Daus, C. S., & Ashkanasy, N. M. (2005). The case for the ability-based model of emotional intelligence in organizational behavior. *Journal of Organizational Behavior*, *26*(4), 453–466. https://doi.org/10.1002/job.321

De Rooij, M., & Weeda, W. (2020). Cross-validation: A method every psychologist should know. *Advances in Methods and Practices in Psychological Science*, *3*(2), 248–263. https://doi.org/10.1177/2515245919898466

Del Giudice, M., & Gangestad, S. W. (2021). A traveler's guide to the multiverse: Promises, pitfalls, and a framework for the evaluation of analytic decisions. *Advances in Methods and Practices in Psychological Science*, *4*(1), 251524592095492. https://doi.org/10.1177/2515245920954925

Durante, K. M., Rae, A., & Griskevicius, V. (2013). The fluctuating female vote: Politics, religion, and the ovulatory cycle. *Psychological Science*, *24*(6), 1007–1016. https://doi.org/10.1177/0956797612466416

Ebersole, C. R., Mathur, M. B., Baranski, E. N., Bart-Plange, D.-J., Buttrick, N. R., Chartier, C. R., Corker, K. S., Corley, M., Hartshorne, J. K., IJzerman, H., Lazarević, L. B., Rabagliati, H., Ropovik, I., Aczel, B., Aeschbach, L. F., Andrighetto, L., Arnal, J. D., Arrow, H., Babincak, P., … Nosek, B. A. (2020). Many labs 5: Testing pre-data-collection peer review as an intervention to increase

replicability. *Advances in Methods and Practices in Psychological Science*, *3*(3), 309–331. https://doi.org/10.1177/2515245920958687

Gelman, A., Hill, J., & Vehtari, A. (2020). *Regression and other stories*. Cambridge University Press.

Gelman, A., & Loken, E. (2013). *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* (pp. 1–17). Department of Statistics, Columbia University. http://stat.columbia.edu/~gelman/research/unpublished/forking.pdf

Girardi, P., Vesely, A., Lakens, D., Altoè, G., Pastore, M., Calcagnì, A., & Finos, L. (2024). Post-selection inference in multiverse analysis (PIMA): An inferential framework based on the sign flipping score test. *Psychometrika*, *89*(2), 542–568. https://doi.org/10.1007/s11336-024-09973-6

Górecki, T., & Smaga, Ł. (2019). fdANOVA: An R software package for analysis of variance for univariate and multivariate functional data. *Computational Statistics*, *34*(2), 571–597. https://doi.org/10.1007/s00180-018-0842-7

Götz, M., & O'Boyle, E. H. (2023). Cobblers, let's stick to our lasts! A song of sorrow (and of hope) about the state of personnel and human resource management science. In M. R. Buckley, A. R. Wheeler, J. E. Baur, & J. R. B. Halbesleben (Eds.), *Research in personnel and human resources management* (Vol. *41*, pp. 7–92). Emerald. https://doi.org/10.1108/S0742-730120230000041004

Grömping, U. (2006). Relative importance for linear regression in *R*: The package **relaimpo**. *Journal of Statistical Software*, *17*(1), 1–27. https://doi.org/10.18637/jss.v017.i01

Hall, B. D., Liu, Y., Jansen, Y., Dragicevic, P., Chevalier, F., & Kay, M. (2022). A survey of tasks and visualizations in multiverse analysis reports. *Computer Graphics Forum*, *41*(1), 402–426. https://doi.org/10.1111/cgf.14443

Hanel, P. H. P., & Zarzeczna, N. (2023). From multiverse analysis to multiverse operationalisations: 262,143 ways of measuring well-being. *Religion, Brain & Behavior*, *13*(3), 309–313. https://doi.org/10.1080/2153599X.2022.2070259

Harder, J. A. (2020). The multiverse of methods: Extending the multiverse analysis to address data-collection decisions. *Perspectives on Psychological Science*, *15*(5), 1158–1177. https://doi.org/10.1177/1745691620917678

Hünermund, P., & Louw, B. (2023). On the nuisance of control variables in causal regression analysis. *Organizational Research Methods*, 10944281231219274. https://doi.org/10.1177/10944281231219274

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*(5), 640–648. https://doi.org/10.1097/EDE.0b013e31818131e7

Jung, K., Shavitt, S., Viswanathan, M., & Hilbe, J. M. (2014). Female hurricanes are deadlier than male hurricanes. *Proceedings of the National Academy of Sciences*, *111*(24), 8782–8787. https://doi.org/10.1073/pnas.1402786111

Klau, S., Schönbrodt, F. D., Patel, C. J., Ioannidis, J. P. A., Boulesteix, A.-L., & Hoffmann, S. (2023). Comparing the vibration of effects due to model, data pre-processing and sampling uncertainty on a large data set in personality psychology. *Meta-Psychology*, *7*, 1–18. https://doi.org/10.15626/MP.2020.2556

Landy, J. F., Jia, M. L., Ding, I. L., Viganola, D., Tierney, W., Dreber, A., Johannesson, M., Pfeiffer, T., Ebersole, C. R., Gronau, Q. F., Ly, A., van den Bergh, D., Marsman, M., Derks, K., Wagenmakers, E.-J., Proctor, A., Bartels, D. M., Bauman, C. W., Brady, W. J., … Uhlmann, E. L. (2020). Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychological Bulletin*, *146*(5), 451–479. https://doi.org/10.1037/bul0000220

Leavitt, K., Mitchell, T. R., & Peterson, J. (2010). Theory pruning: Strategies to reduce our dense theoretical landscape. *Organizational Research Methods*, *13*(4), 644–667. https://doi.org/10.1177/1094428109345156

Liu, Y., Althoff, T., & Heer, J. (2023). *Approximation and progressive display of multiverse analyses* (Version 1). https://doi.org/10.48550/ARXIV.2305.08323

Liu, Y., Kale, A., Althoff, T., & Heer, J. (2021). Boba: Authoring and visualizing multiverse analyses. *IEEE Transactions on Visualization and Computer Graphics*, *27*(2), 1753–1763. https://doi.org/10.1109/TVCG.2020.3028985

Olsson-Collentine, A., Van Aert, R. C. M., Bakker, M., & Wicherts, J. (2023). Meta-analyzing the multiverse: A peek under the hood of selective reporting. *Psychological Methods*. https://doi.org/10.1037/met0000559

R Development Core Team. (2024). *R: A language and environment for statistical computing* (4.3.3) [Computer software]. https://cran.r-project.org/

Ryff, C. D., Kitayam, S., Karasawa, M., Markus, H., Kawakami, N., & Coe, C. (2011). *Survey of midlife in Japan (MIDJA), April-September 2008: Version 3* (Version v3) [dataset]. ICPSR—Interuniversity Consortium for Political and Social Research. https://doi.org/10.3886/ICPSR30822.V3

Ryff, C. D., Kitayama, S., Karasawa, M., Markus, H., Kawakami, N., & Coe, C. (2016). *Survey of midlife in Japan (MIDJA 2), May-October 2012: Version 3* (Version v3) [dataset]. ICPSR—Interuniversity Consortium for Political and Social Research. https://doi.org/10.3886/ICPSR36427.V3

Sarma, A., Hwang, K., Hullman, J., & Kay, M. J. S. (2024). Milliways: Taming multiverses through principled evaluation of data analysis paths. *Proceedings of the 2024 CHI conference on human factors in computing systems*, 1–15. https://doi.org/10.1145/3613904.3642375

Sarma, A., Kale, A., Moon, M. J., Taback, N., Chevalier, F., Hullman, J., & Kay, M. (2023). multiverse: Multiplexing alternative data analyses in R notebooks. *Proceedings of the 2023 CHI conference on human factors in computing systems*, 1–15. https://doi.org/10.1145/3544548.3580726

Schweinsberg, M., Feldman, M., Staub, N., van den Akker, O. R., van Aert, R. C. M., van Assen, M. A. L. M., Liu, Y., Althoff, T., Heer, J., Kale, A., Mohamed, Z., Amireh, H., Venkatesh Prasad, V., Bernstein, A., Robinson, E., Snellman, K., Amy Sommer, S., Otner, S. M. G., Robinson, D., … Luis Uhlmann, E. (2021). Same data, different conclusions: Radical dispersion in empirical results when independent analysts operationalize and test the same hypothesis. *Organizational Behavior and Human Decision Processes*, *165*, 228–249. https://doi.org/10.1016/j.obhdp.2021.02.003

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., … Nosek, B. A. (2018). Many analysts, one data

set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. https://doi.org/10.1177/2515245917747646

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

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*(11), 1208–1214. https://doi.org/10.1038/s41562-020-0912-z

Singmann, H., Heck, D. W., Barth, M., Erdfelder, E., Arnold, N. R., Aust, F., Calanchini, J., Gümüsdagli, F. E., Horn, S. S., Kellen, D., Klauer, K. C., Matzke, D., Meissner, F., Michalkiewicz, M., Schaper, M. L., Stahl, C., Kuhlmann, B. G., & Groß, J. (2024). Evaluating the robustness of parameter estimates in cognitive models: A meta-analytic review of multinomial processing tree models across the multiverse of estimation methods. *Psychological Bulletin*. https://doi.org/10.1037/bul0000434

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Villanova, D. (2023). The appropriateness of outlier exclusion approaches depends on the expected contamination: Commentary on André (2022). *Advances in Methods and Practices in Psychological Science*, *6*(3), 25152459231186577. https://doi.org/10.1177/25152459231186577

Wagenmakers, E.-J., Sarafoglou, A., & Aczel, B. (2022). One statistical analysis must not rule them all. *Nature*, *605*, 423–425. https://doi.org/10.1038/d41586-022-01332-8

Waldrop, M. M. (1992). *Complexity: The emerging science at the edge of order and chaos*. Simon & Schuster.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on *p*-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (2023). *R for data science: Import, tidy, transform, visualize, and model data* (2nd ed.). O'Reilly Media.

Willroth, E. C., Mroczek, D. K., & Hill, P. L. (2021). Maintaining sense of purpose in midlife predicts better physical health. *Journal of Psychosomatic Research*, *145*, 110485. https://doi.org/10.1016/j.jpsychores.2021.110485

Xie, Y., Dervieux, C., & Riederer, E. (2020). *R markdown cookbook*. CRC Press.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Young, C., & Holsteen, K. (2017). Model uncertainty and robustness: A computational framework for multimodel analysis. *Sociological Methods & Research*, *46*(1), 3–40. https://doi.org/10.1177/0049124115610347