

Multicurious: A Multidisciplinary Guide to Multiverse Analysis

**Cassie Ann Short¹, Nate Breznau², Maria Bruntsch^{3,4},
Micha Burkhardt¹, Niko A. Busch⁵, Elena Cesnaite⁵,
Maximilian Frank⁶, Carsten Gießing¹, Daniel Krähmer⁷,
Daniel Kristanto¹, Tina Lonsdorf^{3,4}, Claudia Neuendorf^{8,9},
Hung H. V. Nguyen², Manuel Rausch^{10,11,12}, Xenia Schmalz¹³,
Andreas Schneck⁷, Cem Tabakci¹¹, and Andrea Hildebrandt¹**

¹Department of Psychology, Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany; ²German Institute for Adult Education - Leibniz Center for Lifelong Learning, Bonn, Germany; ³Department of Psychology, University of Bielefeld, Bielefeld, Germany; ⁴Department of Systems Neuroscience, University Medical Center Hamburg-Eppendorf, Hamburg, Germany; ⁵Institute of Psychology, University of Münster, Münster, Germany; ⁶Department of Psychology, Ludwig-Maximilians-Universität München, Munich, Germany; ⁷Institute of Sociology, Ludwig-Maximilians-Universität München, Munich, Germany; ⁸Education Department, University of Potsdam, Potsdam, Germany; ⁹Hector Research Institute for Education Sciences and Psychology, University of Tübingen, Tübingen, Germany; ¹⁰Faculty Society and Economics, Rhine-Waal University of Applied Sciences, Cleves, Germany; ¹¹Fakultät für Philosophie und Pädagogik, Katholische Universität Eichstätt-Ingolstadt, Eichstätt, Germany; ¹²Institut für Psychologie, Alpe-Adria Universität Klagenfurt, Klagenfurt, Austria; and ¹³Department of Child and Adolescent Psychiatry, Ludwig-Maximilians-Universität München, Munich, Germany

Abstract

Multiverse analysis offers a comprehensive response to a core vulnerability in empirical research: the uncertainty of scientific conclusions arising from defensible yet flexible data-processing and -analysis decisions. By systematically mapping and computing all or a sample of all plausible data-processing pipelines, multiverse analysis reports the robustness of findings across analytical flexibility and increases transparency in the research process. As its adoption grows across disciplines, so too does the need for clarity on how to design, report, and interpret multiverse results responsibly. In this article, we provide interdisciplinary guidance on key procedural considerations, including defensibility and equivalence evaluations, preregistration, and computational demands. We aim to harmonize terminology, promote best practices, and foster conceptual cohesion across fields, supported by reference to domain-specific resources when appropriate. By doing so, we contribute to the broader movement toward more robust, reproducible, and transparent science, one that not only reports results but also interrogates the analytical pipelines that produce them.

Keywords

multiverse analysis, robustness

Received 5/23/25; revision accepted 3/1/26

Corresponding Author:

Cassie Ann Short, Department of Psychology, Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany
Email: cassie.ann.short@uni-oldenburg.de

Advances in Methods and
Practices in Psychological Science
April-June 2026, Vol. 9, No. 2,
pp. 1–21
© The Author(s) 2026
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/25152459261434881
www.psychologicalscience.org/AMPPS



Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 4.0 License (<https://creativecommons.org/licenses/by/4.0/>) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

Researchers make a series of discretionary choices at the stages of data collection, processing, and analysis. These choices are often methodologically defensible yet rarely grounded in strong theoretical rationale (Gelman & Loken, 2014). Defensible combinations of these decisions along the workflow collectively form a vast analytical landscape, a “garden of forking paths” (Gelman & Loken, 2013), in which each path or “pipeline” reflects a different yet plausible route to answering the same research question, creating substantial “researcher degrees of freedom” (Simmons et al., 2011). In various scientific disciplines, many-analyst studies have shown that teams working independently on the same data set select unique data-processing and -analysis pipelines, arriving at heterogeneous results, driven not by error but by defensible yet divergent analytical decisions in the garden of forking paths (e.g., in social psychology, Silberzahn et al., 2018; in functional-MRI neurocognitive psychology, Botvinik-Nezer et al., 2020; in sociology, Breznau et al., 2022; in electroencephalogram neurocognitive psychology, Trübutschek et al., 2024). If an effect is sensitive to defensible pipeline variations, replicability across studies is compromised (e.g., Bryan et al., 2019), and a theoretical multiverse of statistical models and outcomes goes unreported.

The single-pipeline approach introduces a multiple-testing problem in that each decision node in the garden of forking paths generates implicit tests that are not executed and hypothesis tests are rarely adjusted for comparisons that have not explicitly been conducted. This creates what Gelman and Loken (2014) termed an “invisible multiplicity,” a hidden source of inferential inflation, that increases the likelihood of a Type I error (Rubin, 2017). The consequence is not merely statistical noise but also potentially false positives and inflated effect sizes that threaten the credibility of empirical claims (Ioannidis, 2005; Simmons et al., 2011).

The analytical flexibility available to researchers can be systematically quantified. Using Figure 1 to demonstrate, we imagine three options to measure the independent variables (e.g., one scale score, sum or average of z values of multiple scales, scores from the first factor of a factor analysis of multiple scales), two options to process the data (e.g., outlier definition, handling missing data), three options to structure the statistical model (e.g., additive linear regression, generalized additive model, mixed-effects model), and two statistical methods to estimate model parameters (e.g., ordinary least squares, maximum likelihood estimator). Each option may be defensible in its own right in a traditional single-pipeline analysis. If all options at each decision node are compatible with all options at each other decision node, the number of potential defensible outcomes to answer the same research question from the same data

set would be the Cartesian product $(3 \times 2 \times 3 \times 2) = 36$. A traditional single-pipeline approach, computing and reporting the result based on only one pipeline, introduces critical sources of uncertainty across measurement, processing, modeling, and estimation (method uncertainty; Hoffmann et al., 2021), thereby limiting replicability and potentially misleading subsequent research or meta-analyses that are based on these results. In our example above, not all options across nodes may be defensibly combined. For example, applying an aggressive outlier-removal rule in a small sample analyzed with a complex mixed-effects model may not be defensible. This deliberately exemplifies the importance of evaluating defensibility at the pipeline level, not only at the level of individual options, because some combinations of individually defensible options may be incompatible and produce unstable results.

Multiverse Analysis as a Solution

Multiverse analysis is a comprehensive, transparent, and systematic approach to report the robustness of results to variation across the defensible garden of forking paths. It explicitly acknowledges methodological and analytical uncertainty, increases transparency, and minimizes the risk of overestimating the robustness and reproducibility of results (Steenen et al., 2016). This approach involves identifying the stages in the research workflow in which different defensible decisions can be made, which may include theoretical conceptualization, data processing, variable quantification, model selection, and statistical specifications (Hoffmann et al., 2021). Each defensible combination of options across decision nodes forms a unique pipeline, all pipelines are computed, and the results are reported as a distribution. Each result in the distribution is often visualized alongside the corresponding pipeline (e.g., Simonsohn et al., 2020), enabling visual assessment of robustness and the sources of uncertainty.

Although several approaches exist to mitigate methodological and analytical uncertainty, hidden multiplicity, and p -hacking, multiverse analysis offers a uniquely transparent and comprehensive solution, which has been proposed to contribute to an ideal test of robustness and reproducibility (Dreber & Johannesson, 2025). For example, preregistrations that include a preanalysis plan can reduce the likelihood of p -hacking and publication bias (Brodeur et al., 2024), but preregistered studies do not necessarily report the robustness of an effect to defensible methodological choices. Because invisible multiplicity remains, a preregistered single-pipeline result may still be vulnerable to inflated Type I error rates (Rubin, 2024). Furthermore, adjusting the prespecified α level¹ to compensate for plausible and

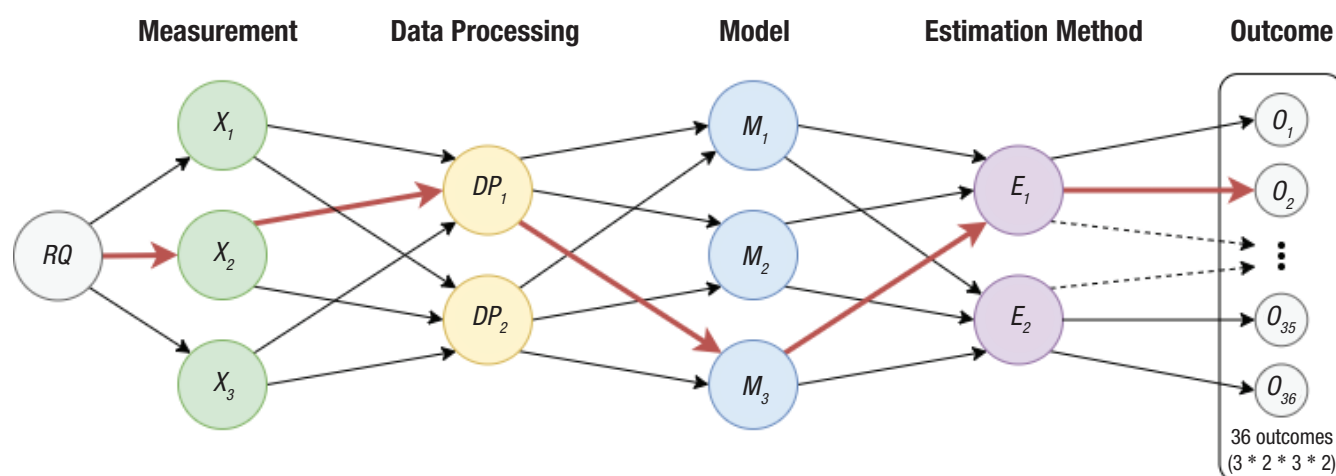


Fig. 1. A schematic example of researcher degrees of freedom at the levels of measurement, data-processing, model, and estimation procedures that produces a multiverse of statistical outcomes. The red line represents the single-pipeline approach that produces one statistical outcome, leaving uncertainty as to the robustness of the outcome to alternative defensible pipelines.

defensible pipelines that are not explicitly computed is a strategy that may mitigate the multiple-comparisons problem (Rubin, 2024). However, this requires prior identification of all defensible pipelines for the given hypothesis, and α correction alone does not explicitly report which decisions an effect may or may not be sensitive to. Although multiverse analysis requires more time and computational resources, it provides a systematic and transparent report of robustness across the full range of plausible pipelines for a given hypothesis and data set. Preregistration and α adjustment can be implemented together with multiverse analysis in a complementary manner (e.g., Beauducel et al., 2024; Paul et al., 2025; which are two of many preregistered multiverse analyses as part of the CoScience Project available at <https://osf.io/rdpze/>), but they do not provide the same degree of explicit robustness information when applied in isolation.

Interdisciplinary History of Multiverse Analysis and the Benefits of a Standardized Term

The practice of reporting the uncertainty associated with methodological flexibility by examining multiple data-processing or model choices is well established (e.g., Leamer, 1985; Ramsey, 1969); however, this approach has tended to develop independently across scientific disciplines. In engineering and physics, the focus on robustness is often to determine whether a device functions reliably under a range of plausible parameter settings and operating conditions (e.g., Kozak et al., 2023). In neuroscience and psychophysiology, complex data-processing routines are required to preserve signal while

removing noise from data recordings, with many processing options and combinations of options being often defensible, culminating in extensive researcher degrees of freedom at the level of data preprocessing (e.g., Paul et al., 2022). In sociology and education, robustness assessments tend to focus on model specification (e.g., Muñoz & Young, 2018). These distinct epistemic priorities have likely contributed to the development of varied robustness terminology across disciplines.

Commonly used terms describing similar, partial, or synonymous approaches include “sensitivity analysis” (e.g., Saltelli et al., 2019), “robustness checks” (e.g., Neumayer & Plümper, 2017), “vibration of effects” (Ioannidis, 2008), “specification-curve analysis” (Simonsohn et al., 2020), “manyverse analysis” (Kuhn et al., 2022), “many-analysts approach” (Silberzahn et al., 2018), “cooperative forking-paths analysis” (Wacker, 2017), “multimodel analysis” (Young & Holsteen, 2017), and “computational robustness analysis” (Muñoz & Young, 2018). The approaches described by these terms vary in scope and methodology, often reflecting the diversity of focus across scientific fields (Table 1). The definitions of these terms also vary in the literature; the definitions given in Table 1 are taken from the above references and the Framework for Open and Reproducible Research Training (FORRT) *Glossary* (FORRT, n.d.).

Each of the approaches in Table 1 can be conducted transparently and systematically, but this is not always the case. For example, a review of footnotes describing robustness checks in sociology found that pipeline selection as part of a robustness check was often not clearly justified, transparent, or systematic (Young & Holsteen, 2017). Establishing a unifying, standardized term to denote a transparent and systematic procedure applicable across disciplines would serve multiple purposes.

Table 1. A List of Related Analytical Approaches to Multiverse Analysis With Definitions

Term	Definition
Robustness analysis/ computational robustness	This is the broadest term, reflecting the examination of the stability of results under one or more experimental designs, data preprocessing pipelines, and/or alternative model specifications; see also FORRT (n.d.). Computational robustness is an approach to systematically test the sensitivity of results to varying modeling assumptions and computational choices. Computational robustness is a broad concept that refers to the property that research results are stable and consistent across a wide range of plausible analytical choices. Computational robustness is the desired outcome; the terms below refer to the tools used to achieve robustness.
Multiverse analysis	In multiverse analyses, a single hypothesis is tested across all possible data-processing and statistical pipelines considered justifiable; see also FORRT (n.d.).
Specification-curve analysis	Multiverse analyses have been visualized in many forms, including, for example, histograms, tile plots, and specification curves. A specification curve is one of these visualization forms and has placed more emphasis on inferential statistics to jointly interpret the multiverse of results; see also FORRT (n.d.).
Vibration of effects	According to Ioannidis (2008), vibration of effects conveys the extent to which an effect may change in alternative analytical approaches. The focus is mainly on comparing model fit and evaluating results across alternative model structures and less on data-processing decisions. In addition, the visualization of results differs from those used in more recent multiverse and specification-curve analyses.
Multimodel analysis	Multimodel analysis starts from the premise that there is often no single “correct” model. Researchers face uncertainty about which model specification is most appropriate for a given research question and data set, so multimodel analysis is used to test the robustness of results across model specifications.
Manyverse analysis	This approach refers to a small-scale, focused examination of the robustness of estimates to various data-processing or -quantification choices.
Cooperative forking- paths analysis	Represents a larger-scale robustness report that is implemented by multiple experts in a given research field to examine the robustness of results across varied analytical choices that were collaboratively decided on.
Many-analysts approach	In multianalyst studies, multiple researchers independently analyze the same data to address a common research question; see also FORRT (n.d.).
Sensitivity analysis	According to Saltelli et al. (2019), sensitivity analysis investigates how uncertainty in a model's output can be attributed to different sources of uncertainty in its inputs. It aims to quantify the influence of individual input variations on the overall uncertainty of the model's results. Note that this is not the only definition of sensitivity analysis. Rubin (2017), for example, used the term “sensitivity analysis” as a synonym for multiverse or “forking paths” analysis.

It would communicate that the approach applied meets recommended standards, thereby increasing confidence in the credibility of reported results; it would encourage methodological rigor; and it would promote cohesion across the currently fragmented terminological landscape (e.g., Voracek et al., 2025) to provide a common framework for collaborative methodological development and discussion. We and others (Young & Cumberworth, 2025) propose that the relatively new term “multiverse analysis” can serve as this umbrella term for transparent and systematic procedures regardless of the specific field and stage of the research process being assessed for robustness.

How to Do It

In this section, we provide a step-by-step multidisciplinary guide to multiverse analysis that spans

conceptualization to interpretation of the results. A visual map of the steps is illustrated in Figure 2. Although in the present article we focus on quantitative research, the underlying principles of multiverse analysis apply to qualitative and mixed-methods approaches, in which defensible analytical variation may arise from multiple defensible coding schemes, thematic frameworks, or interpretive lenses.

Step 1: defining the focus of the multiverse

The first step in performing a multiverse analysis is to (a) specify the stages of the research process in which robustness will be examined and (b) select the metric of interest that will be assessed for robustness. Researcher degrees of freedom exists at multiple stages of the research process (Hoffmann et al., 2021), including measurement (in which conceptualization, measurement,

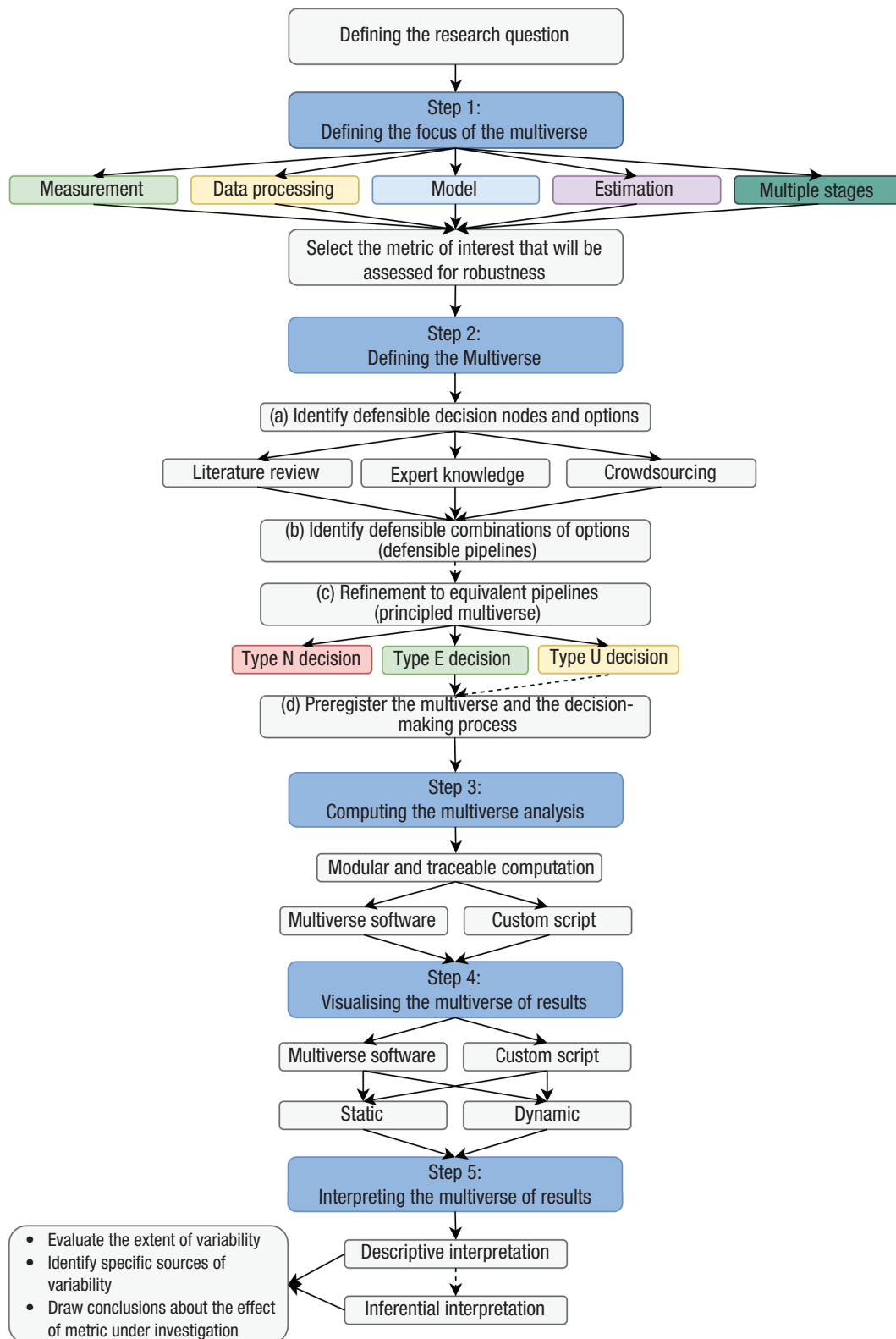


Fig. 2. A flowchart of the recommended procedural steps for defining, computing, and reporting a systematic multiverse analysis.

and collection are defined), data processing (in which the data are prepared for analysis, such as setting different exclusion criteria), modeling (in which the statistical model is specified, such as different types of regression models and choosing which independent variables to include), and estimation (in which the approach to parameter estimation is determined, such as the ordinary-least-squares method or a robust maximum-likelihood estimator). Researchers may choose to examine robustness in a single stage or across multiple stages; the latter enables exploration of potential interactions between options across stages, revealing whether certain combinations of defensible choices influence results. After identifying the relevant stage(s), it is essential to define the metric(s) of interest that will be assessed for robustness. Multiverse analyses have traditionally reported distributions of p values (e.g., Steegen et al., 2016), effect sizes, and confidence intervals (e.g., Beauducel et al., 2024); common effect measures across models (e.g., Lonsdorf et al., 2022); and overall variability in multiverse effect sizes (Olsson-Collentine et al., 2025). Alternatively, psychometrics, such as reliability estimates, can also be estimated across the multiverse to assess, in part, the psychometric quality of estimates across the garden of forking paths.

Three interpretive goals of multiverse analysis include to evaluate the extent of variability in the metric of interest across all pipelines, to identify specific sources of this variability, and to draw conclusions about the effect or metric under investigation (Sarma et al., 2024). Clarifying the goal in the early stages of multiverse design can inform subsequent decisions. For example, if the primary aim is to report the extent of variability, a key focus may be to develop a structured approach to measure and report empirical similarity across data sets. If identifying the sources of variability is the primary aim, a key focus may be to develop a structured procedure for measuring and reporting pipeline similarity. Establishing the interpretive aims in advance can support preregistration and promote coherence between the analytical design and interpretation.

Step 2: defining the multiverse

Defining the multiverse of pipelines involves a systematic procedure of substeps, which we describe in detail.

Distinguishing between defensibility and equivalence. Before defining the specific decision nodes and options that constitute the multiverse, it is important to distinguish between defensibility and equivalence. Defensibility concerns whether a given analytical pipeline can be justified independently; in other words, whether it could be reasonably defended in the context of a peer review if it were the only pipeline applied to answer the

research question using the data set at hand. A pipeline may be considered defensible based on theoretical rationale, established conventions, statistical logic, or empirical validation. Approaches to evaluating defensibility are described in Steps 2a and 2b. The term “equivalence” was introduced as a classification framework to systematically evaluate whether “alternative specifications” (i.e., options or pipelines) can reasonably be treated as interchangeable (Del Giudice & Gangestad, 2021), building on the argument that results depend on analytical decisions being defensible, arbitrary, and motivated (Simonsohn et al., 2020). Del Giudice and Gangestad (2021) argued that such arbitrariness can be justified when alternatives are expected a priori to be equivalent with respect to a researcher-specified criterion relating to construct measurement, the effect of interest, or the power to/precision in estimating an effect, for example. Accordingly, their framework classifies candidate pipelines as equivalent, nonequivalent, or uncertain with respect to that criterion, providing a principled basis for determining when alternatives may reasonably be considered interchangeable for inclusion in a multiverse analysis.

Building on this framework, we formalize the conceptual and procedural distinction between defensibility and equivalence on two bases. First, many multiverse analyses have justified the inclusion of pipelines by applying alternative transparent evaluation frameworks, such as those used in the existing literature as determined by a systematic literature review (e.g., Kristanto et al., 2024). In such cases, pipelines are evaluated independently as defensible. However, the framework proposed by Del Giudice and Gangestad (2021) emphasized that defensibility, in the sense of being commonly applied, does not necessarily warrant treating them as interchangeable. For example, in a behavioral task, performance may be quantified as either the proportion of correct responses or as d' ($z[\text{hit rate}] - z[\text{false-alarm rate}]$). Both measures are widely used in the literature and defensible in isolation (e.g., Allard et al., 2023; Duschek et al., 2022; Pergher et al., 2018), yet they target different constructs (overall accuracy vs. bias-corrected sensitivity) and would therefore be nonequivalent under a criterion of construct validity. Although multiple pipelines may be defensible in isolation, some may be objectively preferable to others with respect to a relevant criterion, and treating such pipelines as interchangeable would therefore be inappropriate (Del Giudice & Gangestad, 2021). This highlights a distinction between a defensible multiverse, which is defined by independent justification, and a principled multiverse, which is constrained by equivalence based on relative evaluation.

Second, candidate pipelines must first be identified to evaluate equivalence. Therefore, we place equivalence as a subsequent step to the identification of defensible pipelines. This procedural order is reflected in the

Table 2. Overview of Theoretical and Empirical Approaches to Assessing Defensibility and Equivalence

	Theoretical approach	Empirical approach
Defensibility (individual pipelines)	Definition: conceptual and methodological justification of a pipeline in isolation. A pipeline is defensible if it could reasonably be justified as appropriate for the research question and data. Example methods: systematic literature review, expert panel, multilab expert consensus Benefits: aligns with disciplinary standards, facilitates comprehensive mapping of analytical flexibility Limitations: subject to publication bias, constrained by researcher expertise	Definition: empirical evaluation of individual pipelines to test whether they produce data or estimates that meet an adequacy threshold when applied to the data set. For example, does this pipeline meet a minimum standard on its own? Example methods: psychometric properties (e.g., reliability), data-quality checks (e.g., signal-to-noise ratio) Benefits: ensures each pipeline produces functionally adequate data or estimates Limitations: may exclude pipelines that are commonly used in the literature but perform poorly in the current data set, thereby limiting the ability to investigate whether such pipelines contribute to low replicability
Equivalence (pipelines relative to each other)	Definition: conceptual assessment of whether defensible pipelines can be expected a priori to yield practically equivalent answers with respect to a researcher-defined criterion (e.g., construct validity or estimation precision) and can therefore be treated as interchangeable (Del Giudice & Gangestad, 2021) Example methods: expert consensus, theoretical expectations, or evidence from prior methodological studies Benefits: clarifies interpretive comparability, provides principled justification for treating alternatives as arbitrary when appropriate. Limitations: not tested on the present data set, sensitive to theoretical assumptions	Definition: empirical comparison of defensible pipelines using the data set at hand to evaluate equivalence relative to a defined criterion and threshold. For example, given that these pipelines are defensible, are their results on a specified criterion sufficiently similar to be treated as interchangeable? Example methods: using an interval threshold for similarity of results on the specified criterion to establish equivalence, as in SMART (Short, Inceler, et al., 2025) Benefits: ensures interpretative comparability across pipelines Limitations: equivalence conclusions may be data-set-specific

Note: All approaches can be transparent when well documented and involve a degree of subjectivity in selecting evaluation criteria, adding to researcher degrees of freedom. SMART = Systematic Multiverse Analysis Registration Tool.

Systematic Multiverse Analysis Registration Tool (SMART; Short, Inceler, et al., 2025), a recently published application designed to guide users through the systematic and transparent identification of multiverse pipelines. In SMART, potentially defensible pipelines are first identified and evaluated individually, and justification strategies are defined by the researcher. An optional subsequent stage then evaluates the defensible pipelines relative to one another for equivalence based on a researcher-specified criterion. Consequently, defensibility can be justified independently without an equivalence assessment, whereas an equivalence assessment further structures interpretation by pruning the multiverse based on a relevant researcher-specified criterion to pipelines that are interchangeable. Identifying the analytical space precedes assessing equivalence within it, highlighting a distinction in the procedural order.

For these two reasons, we frame equivalence assessment as a subsequent step in the procedure. Assessing equivalence requires a set of individual defensible pipelines, whereas a defensible multiverse does not

necessarily need to be equivalent. The appropriate approach depends on the goal of the multiverse analysis. For example, when the goal is to assess robustness of an effect across analytical diversity in the literature, such as investigating potential sources of low replicability, defensibility without pruning for equivalence may be appropriate. On the other hand, if the goal is to report the robustness of a single underlying construct or theoretically postulated effect, then an equivalence evaluation may prevent defensible but nonarbitrary decisions from being treated as arbitrary (Del Giudice & Gangestad, 2021). Approaches to evaluating equivalence are discussed in Step 2c. A summary of the distinctions is presented in Table 2, and Figure 3 illustrates three practical examples based on behavioral, biophysiological, and survey data.

Step 2a: identification of decision nodes and options.

First, researchers should decide on the method for identifying defensible pipelines. Approaches in the literature include systematic literature reviews (e.g., Kristanto et al.,

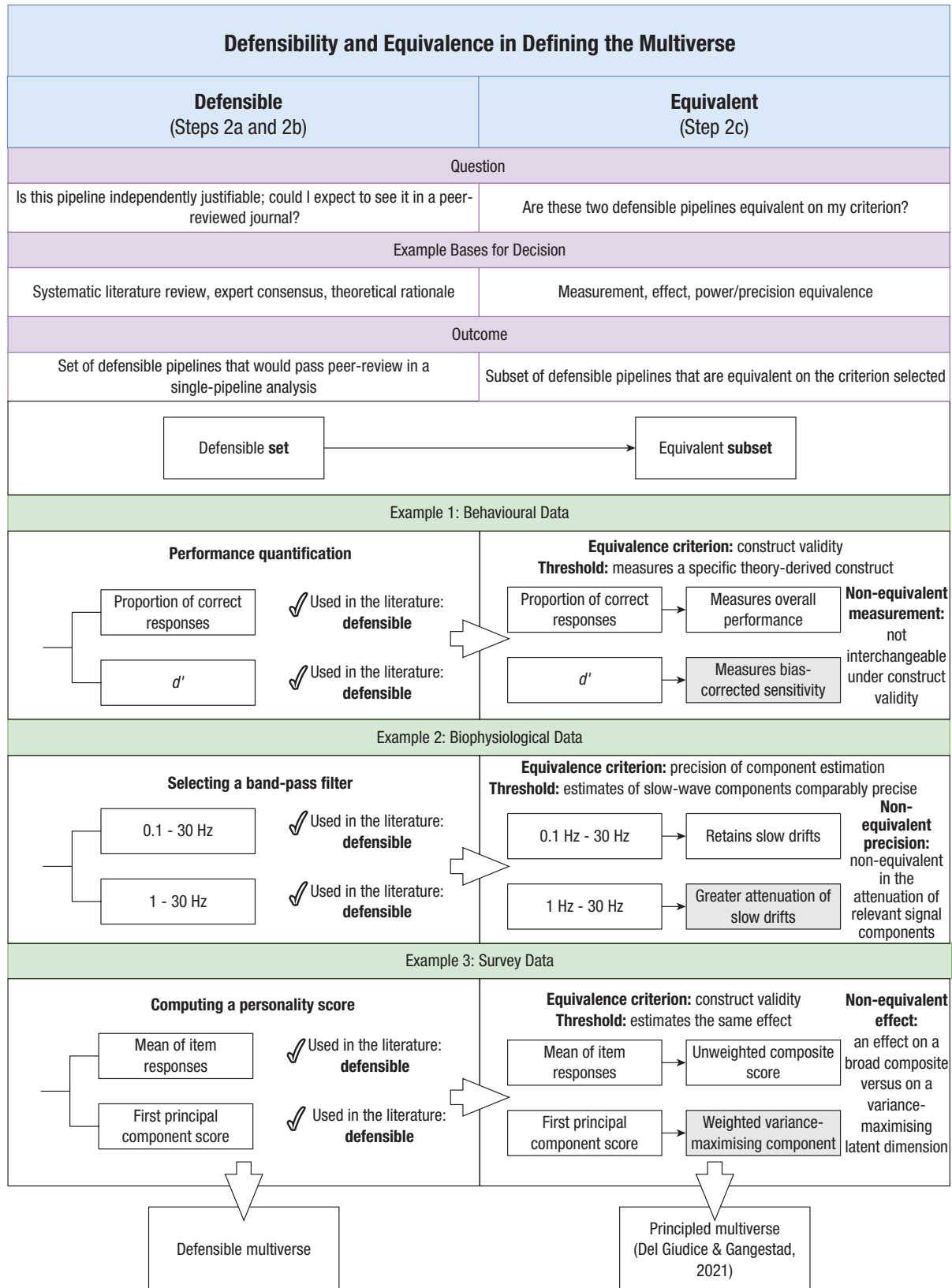


Fig. 3. Example distinctions between defensibility and equivalence in defining the multiverse. Options may be defensible because they are commonly used in the literature and could reasonably appear in a peer-reviewed analysis. However, once an equivalence criterion and threshold are specified for the effect or component of interest, some defensible options may be excluded from a principled multiverse because they are nonequivalent relative to that target. Examples illustrate equivalence assessed with respect to three equivalence modes proposed by Del Giudice and Gangestad (2021): measurement, effect, and power/precision equivalence. RT = response time; d' = sensitivity index (signal-detection theory); gray shading = removal on the basis of nonequivalence.

2024), collaborative expert consultation (e.g., Paul et al., 2022; Wacker, 2017), and many-analyst or crowdsourcing approaches in which independent experts, or teams of experts, each provide one preferred pipeline (e.g., Trübutschek et al., 2024). Regardless of the approach, a degree of subjectivity is inherent. For example, the inclusion criteria in systematic reviews are flexible, and expert consensus reflects collective but nonobjective judgment. When the expert or collaborative-expert approach is taken, researchers should first list each decision node along the pipeline that was identified and whether multiple options exist at each node. Each decision node that was identified and each potential option at each node should be documented. These will be used to build defensible pipelines in the following step. Varying the inclusion of covariates as a decision node warrants particular consideration, as discussed in the Ongoing Debates section.

Step 2b: identification of defensible combinations of defensible options (defensible pipelines). Once the available nodes and options have been identified, the next step is to determine whether all Cartesian combinations of options across nodes are defensible or if some options are incompatible with each other. Although this step is time-consuming, it is essential to define a multiverse that is free of indefensible pipelines, which would contaminate the final reported distribution with noise. This process should be repeated for each defensible order of options along the pipeline because the defensibility of combined options can depend on the sequence of operations. This defines the defensible multiverse, which is the set of pipelines that are individually plausible. All combinations deemed indefensible should also be transparently documented with clear justification because these justifications influence the multiverses' interpretive boundaries, enabling readers to appraise potential bias and supporting transparency (Short, Inceler, et al., 2025).

In the literature, defensibility, whether at the option or pipeline level, has tended to be assessed on a theoretical basis, such as relying on precedent in the literature, expert discussion, or multilab consensus, as outlined above. Although these strategies draw on accumulated disciplinary knowledge and align with prevailing methodological norms, they remain vulnerable to publication bias, are restricted to the search criteria or expertise of participating researchers, and risk overlooking specific data characteristics that might render a theoretically defensible choice unsuitable for a particular data set. An empirical approach to assessing defensibility on a given data set can involve testing each pipeline individually against a prespecified threshold of adequacy, such as a data-quality or psychometric threshold. Any pipelines that pass this threshold, regardless of the similarity of the results between pipelines, would be included within a defensible multiverse. Such an

approach could help ensure that each pipeline is statistically adequate. Both theoretical and empirical approaches to determining defensibility introduce additional researcher degrees of freedom in terms of selecting criteria and thresholds.

Step 2c: selection of equivalent pipelines for a principled multiverse analysis. A proposed further step in the procedure is to refine the defensible multiverse into a principled multiverse (Del Giudice & Gangestad, 2021). In a principled multiverse, all pipelines are judged to be equivalent on a relevant criterion, such as measuring the same construct, testing for the same effect, and estimating the effect with comparable power or precision (Del Giudice & Gangestad, 2021). Whereas defensibility assessment concerns the adequacy of each individual pipeline considered in isolation, equivalence assessment concerns interpretive comparability across pipelines, evaluating whether defensible pipelines are sufficiently similar with respect to a researcher-defined criterion and threshold to be meaningfully compared within the same multiverse.

The equivalence-criterion and -assessment approach may vary across multiverse analyses. For instance, one study may define equivalence theoretically, based on a priori expert expectations, and another study may define equivalence empirically, defining an acceptable deviation within which pipelines can fall to be considered equivalent. Theoretical assessments of equivalence rely on conceptual reasoning, prior literature, or domain expertise. Although these approaches enhance interpretive clarity, they may not reflect the actual data structure. In contrast, empirical approaches can test equivalence directly by computing the chosen criterion on the data set in question. Equivalence assessment may lend itself to empirical evaluation in some contexts, particularly in cases in which a priori knowledge is limited.

When comparing all defensible pipelines for equivalence, Del Giudice and Gangestad (2021) proposed three categorical outcomes: (a) If pipelines are deemed equivalent, they are classified as "Type E" and should be included in the principled multiverse analysis. (b) If pipelines are deemed nonequivalent with respect to the chosen criteria, those that are dissimilar and less justifiable are classified as "Type N" and should be excluded. (c) If the equivalence of a pipeline to the others is uncertain, it is classified as "Type U," and the researcher may decide whether to include or exclude it based on whether the multiverse analysis is exploratory. The concept of pipeline equivalence is discussed in the Ongoing Debates section.

Step 2d: preregister the multiverse analysis and the decision-making process. In addition to the content that is preregistered for single-pipeline studies, a multiverse-analysis preregistration should document the procedure,

Table 3. A List of Open-Source Packages and Toolboxes to Facilitate Computation and/or Visualization of Multiverse Analyses

Package/ toolbox	Functionality	Language	Authors	Link to documentation
Boba/Boba Visualizer	Compute and visualize	Agnostic (Boba is a metalanguage)	Liu et al. (2020)	https://github.com/uwdata/boba
Specr	Compute and visualize (specification curve)	R	Masur and Scharkow (2020)	https://masurp.github.io/specr/
Multifear	Compute and visualize (in the field of human fear conditioning)	R	Lonsdorf et al. (2022)	https://github.com/AngelosPsy/multifear
Multiverse/ mverse	Compute (mverse is an extension of multiverse with a student-friendly interface)	R	Sarma et al. (2023)	https://github.com/MUCollective/multiverse https://github.com/mverseanalysis/mverse
Milliways	Visualize	JavaScript (standalone) Code output: R	Sarma et al. (2024)	https://abhsarma.github.io/milliways/
Multitool	Compute and visualize	R	E. S. Young and Vermeent (2024)	https://ethan-young.github.io/multitool/
Multiverse Sampling Tool	Sample pipelines when the multiverse is too large to compute exhaustively	Python	Short, Hildebrandt, et al. (2025)	https://github.com/cassiesh/MultiverseSamplingTool
Comet	Compute and visualize (in the field of functional MRI dynamic functional connectivity)	Python	Burkhardt and Gießing (2026)	https://github.com/mibur1/comet

justification, and outcomes of Steps 1 to 2c and the planned descriptive and inferential analysis of results across the multiverse. Practical support to implement and document the full decision-making process has recently been developed (Short, Inceler, et al., 2025).

Step 3: computing the multiverse analysis

Several packages and toolboxes have been developed across different software programs to facilitate the computation of multiverse analyses. Using a tool that aligns with the specific needs of a given researcher can save time and improve reproducibility. For an overview of the tools currently available, see Table 3. If no existing tool is fully suited to the desired analysis, some design aspects may still be useful. In line with good open-science practices (Nosek et al., 2015), care should be taken to avoid creating a black box that is difficult for other researchers to understand and reproduce.

Step 4: visualizing of the multiverse of results

Clear visualization is essential for accurately interpreting the results. Several visualization approaches have been developed and applied throughout the literature. These include vibration of effects plots (Patel et al., 2015),

outcome matrices (Steege et al., 2016), histograms and density plots (Young & Holsteen, 2017), and specification curves (Simonsohn et al., 2020), which were reviewed by Hall et al. (2022). The specification curve, illustrated in Figure 4, offers a particularly comprehensive representation by plotting all statistical results in vertical alignment with the corresponding pipeline. However, one limitation of the specification curve is that it does not intuitively plot more complex structural variability between pipelines, such as variation in the order of decision nodes along the pipelines and varying pipeline lengths. Although additional visual features (e.g., icon size or transparency) could be used to encode some of this variability, subtle but potentially impactful distinctions may not be clearly communicated. Furthermore, as the number of pipelines increases into the thousands, displaying all results within one curve can become challenging because of spatial constraints. To address this issue, new formats such as multiverse plots (Krähmer & Young, 2026) are being developed to accommodate large-scale multiverses while maintaining high informational value about influential options and decision nodes.

Interactive visualization tools offer another promising solution, allowing users to explore multiverse results dynamically. Examples include interactive specification curves, created using applications such as ShinyApp (e.g., <https://coscience.psy.uni-hamburg.de/sample-apps/>

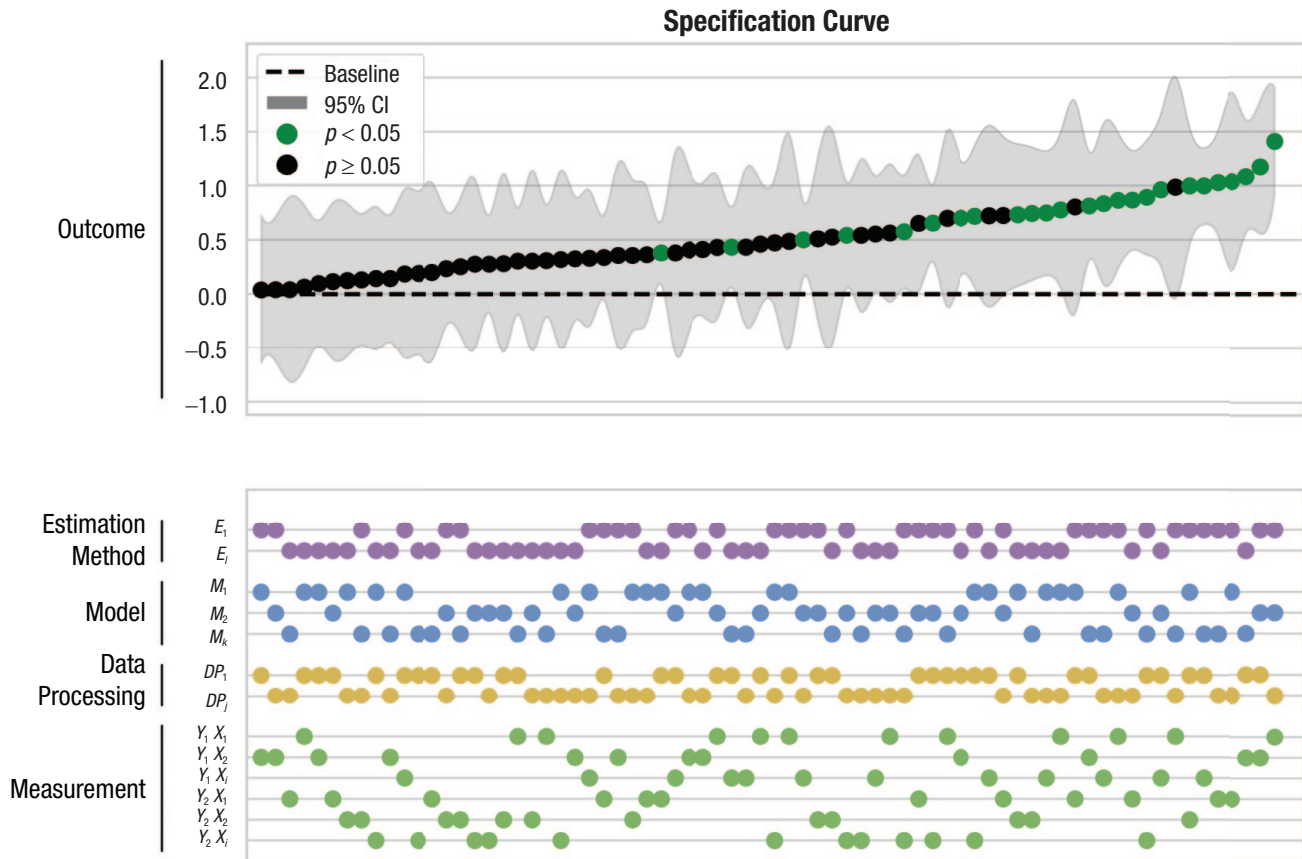


Fig. 4. An example specification-curve figure that visualizes the outcomes of a multiverse in vertical alignment with the respective pipelines. Created using the Comet Toolbox (Burkhardt & Gießing, 2026).

FAA_Experimenter), interactive documents (Dragicevic et al., 2019), and specialized toolboxes such as Boba for Python (Liu et al., 2020) and Milliways for R (Sarma et al., 2024). These dynamic visualizations may improve clarity for large-scale multiverse analyses, helping to identify patterns and influential decisions that may be obscured in static figures.

Although visualizations are central to communicating multiverse results, it is the responsibility of authors to guide interpretation rather than leave readers to independently infer meaning. Statistical literacy encompasses accurate analysis and transparent, ethical communication, which includes providing a concise narrative that summarizes the key findings ethically (e.g., Gigerenzer et al., 2007). Providing such guided interpretation ensures that visualizations are informative and accessible.

Step 5: interpreting of the multiverse of results

Interpretation of results should align with the broader goals of the multiverse analysis. In this context, Sarma et al. (2024) distinguished three interpretive objectives:

(a) to evaluate the extent of variability in the metric of interest across all pipelines, (b) to identify specific sources of this variability, and (c) to draw conclusions about the effect or metric under investigation. Below, we highlight which interpretation approaches address which of these objectives.

Descriptive interpretation. Descriptive interpretation remains the most common and accessible approach to interpreting the multiverse of results. This involves reporting statistical summaries of central tendency, such as the median effect size and the percentage of pipelines that reach statistical significance in each direction and those that do not (e.g., Beauducel et al., 2024; Paul et al., 2025). These approaches contribute toward evaluating the extent of variability in the metric of interest across all pipelines. Descriptive interpretations can also entail identifying options and combinations of options across nodes that are proportionally overrepresented in extreme results, as visualized by specification curves (e.g., Simonsohn et al., 2020). These approaches contribute toward identifying specific sources of this variability. Furthermore, researchers may use visualizations or descriptive statistics to

evaluate the overall robustness of the effect of interest, interpreting what the distribution implies about the effect. For example, if the majority of pipelines yield effects of similar magnitude in the same direction, the effect may be considered robust across various defensible analytical decisions. Conversely, widely distributed or conflicting effects may indicate that a more cautious interpretation of the effect is required. This approach contributes toward drawing conclusions about the effect or metric under investigation. By descriptively examining the variability in multiverse results, preliminary claims about robustness or the lack thereof can be made, and influential decision nodes or options can be identified. However, caution should be exercised when drawing conclusions because the degree of overlap between pipelines may vary, and specific options or combinations thereof may be unequally represented. Empirical similarity across data sets can produce the illusion of consistency rather than genuine robustness (for further discussion, see dataset similarity across pipelines).

Inferential interpretation. Inferential analysis of the multiverse is more complex and remains an area of active development. Such analyses require careful consideration of several statistical challenges, including the nonindependence of pipelines, varying degrees of independence between pipelines, unequal representation of some options in the computed multiverse (e.g., because of dependencies or pipeline sampling), and the multiple comparisons now made explicit. Several inferential approaches have emerged that aim to overcome these issues. For example, Step 3 of the specification-curve analysis of Simonsohn et al. (2020) tests whether the median, share, or average z value of all effects are more extreme than would be expected if all specifications had an effect of zero. Postselection inference in multiverse analysis (Girardi et al., 2024) uses a sign-flip, score-based, conditional-resampling procedure to enable postselection inference with family-wise error-rate control across pipelines. Some data meta-analysis approaches account for interpipeline dependence (Lefort-Besnard et al., 2025) and avoid overconfident inference under shared data (Bartoš et al., 2025). A statistical sensitivity-analysis framework estimates the dependence structure between pipeline-specific estimates, using it to estimate dependence-aware pooled effects and test cross-pipeline hypotheses (e.g., no effect across all pipelines or at least one null effect; Ozenne et al., 2025). Young and Cumberworth (2025) proposed a thought experiment that quantifies how strong unobserved confounding would need to be in aggregate to fully account for an observed multiverse-level effect.² Although inferential approaches to multiverse analysis are still developing, these methods offer promising solutions for dealing with the inherent complexity. However, inference tests that reduce the

variability of interest to a binary outcome (e.g., significant vs. nonsignificant) may risk oversimplifying the uncertainty that multiverse analysis aims to make transparent (Götz et al., 2024).

Probabilistic versus possibilistic interpretation.

When one interprets the results of a multiverse analysis, it is important to distinguish between probabilistic and possibilistic uncertainty (Sarma et al., 2024). Probabilistic uncertainty arises in each individual-analysis pipeline and reflects the inherent variability in estimating unknown quantities (e.g., regression coefficients) from a finite sample size. This type of uncertainty supports probabilistic statements to be made about the likelihood of certain values, conditional on the model and assumptions. On the other hand, possibilistic uncertainty refers to the range of possible results across the entire multiverse of analysis decisions. It arises from the multiple defensible decisions made during data processing, model selection, or other stages of the procedure, producing a set of defensible, possible results. These results should not be interpreted as being more or less likely to be true based on their frequency in the multiverse. For example, although a particular result may occur more frequently than others across different pipelines, this does not imply that it is more likely to be the “correct” result. This is because possibilistic uncertainty does not rely on probability; rather, it recognizes each result as a valid possibility arising from defensible analytical choices. Recognizing this distinction is essential for avoiding misinterpretation. Tools such as Milliways (Sarma et al., 2024) are designed to make the difference explicit, using visualizations such as consonance curves and probability boxes to distinguish between probabilistic uncertainty within a pipeline and possibilistic uncertainty across pipelines. Emerging inferential frameworks, such as those described above, offer promising approaches to inference without assuming probabilistic independence.

Dataset similarity across pipelines. The potential for empirical similarity between data sets produced by different defensible pipelines should be considered. If such similarity is prevalent but unreported, consistent results across the multiverse may be misinterpreted as robustness to analytical variation when they instead reflect underlying empirical similarity. However, empirical similarity is not inherently problematic. Rather, the degree of empirical overlap can provide valuable information about which analytical decisions exert minimal influence on the data and which are most consequential. This knowledge could guide future robustness assessments when resources are limited, inform interpretations of replication efforts, and be incorporated into methodological developments aimed at integrating multiverse results. Therefore, we recommend quantifying and reporting empirical similarity across

the multiverse. For example, pairwise similarity or distance metrics can be used for this purpose, such as cosine similarity (e.g., Dafflon et al., 2022) and Euclidean distance (e.g., Short, Hildebrandt, et al., 2025) or graph neural networks (e.g., Burkhardt et al., 2024). Reporting empirical similarity would therefore enhance transparency, reduce the risk of overstating robustness, and inform the interpretation of multiverse results.

Feature-Specific Procedures

Specific data features may bring particular procedural considerations. For example, although secondary data sets (e.g., Program for International Student Assessment, OECD, 2022; the British Household Survey, University of Essex, 2023) are resource-efficient, there is a lack of control over data-collection methods. This restricts the decision nodes available for multiverse analysis at the design stage, and theoretical hypotheses may not always align perfectly with the constructs and measures present in the data set. In addition, limited contextual knowledge about the sampling procedures and data-collection conditions may make accurate a priori equivalence ratings more challenging. We recommend reviewing the dataset documentation to understand the data-collection process and study design, inform the defensibility and equivalence assessments, and explicitly explain how each decision is consistent with the documented properties of the data set. Further feature-specific considerations arise in longitudinal designs, in which additional uncertainty is introduced by time-related factors and autoregressive structures, and in the presence of missing data, in which the nature of missingness (D. B. Rubin, 1976) and the strategy with which missingness is handled (Peugh & Enders, 2004) may influence whether bias is introduced to the multiverse-analysis results.

Large and complex multiverse analyses have arisen across multiple research domains, including the social sciences, observational studies, and biophysiological research (e.g., Botvinik-Nezer et al., 2020; Engzell & Mood, 2023; Muñoz & Young, 2018; Paul et al., 2022; Trübtschek et al., 2024). Biophysiological data provide a salient example of the challenges this can pose in defining the multiverse because of the complex preprocessing required to extract signal from noise across a large number of potential decision nodes, options, and sequences. Various approaches have been used in the literature to systematically identify defensible pipelines, such as multilaboratory expert consensus (e.g., Paul et al., 2022), systematic literature reviews (e.g., Kristanto et al., 2024), and crowdsourcing pipelines through a many-analysts approach (e.g., Trübtschek et al., 2024). However, the defensibility and equivalence of preprocessing pipelines for biophysiological data are context-dependent. What is defensible and equivalent may

depend on factors that shape variability in the raw data set, such as the population studied (Haller et al., 2014), data quality (Clayson et al., 2021), and the experimental task (Clayson, 2024). Consequently, defensibility and equivalence evaluation should be tailored to these contexts, particularly in cases in which the inclusion of options can contribute to a large number of pipelines being generated. When literature reviews are used to identify pipelines, the eligibility criteria should be specific enough to reflect contextual factors; otherwise, unnecessarily large or internally incoherent multiverses may be generated.

Open Challenges and Future Directions

Although multiverse analysis is a powerful tool for improving the transparency and rigor of empirical research, its implementation is not without challenges.

Transparently balancing broadness and specificity

A central challenge in multiverse analysis is defining the analytical scope and doing so transparently because the goals of multiverse analysis may be undermined by undisclosed bias in the selection of included pipelines. An overly restrictive scope risks inadvertently introducing bias, whereas an overly broad scope may dilute interpretability. Following a systematic procedure and a clear conceptual framework for pipeline selection, such as that proposed by Del Giudice and Gangestad (2021), and conducting a data-driven quality or psychometric assessment of pipelines (Short, Inceler, et al., 2025) when relevant to the goals can help researchers to appraise potential bias. Decisions regarding scope, options, and pipelines should be explicitly justified in relation to the stated goals of the multiverse analysis and documented in sufficient detail to allow readers to evaluate them.

Preregistration

Preregistration has become an important tool for promoting transparency and reducing bias in empirical research by specifying research questions, hypotheses, study design, and analysis plans before data collection and analysis begin. This practice is increasingly recognized across scientific disciplines for its contribution to research integrity. However, questions commonly arise about whether and how to preregister multiverse analyses.

A common misconception is that multiverse analysis negates the need for preregistration. Multiverse analysis involves exploring all defensible or equivalent analytical pipelines within a specified scope to report robustness, whereas preregistration commits researchers to a pre-specified analysis plan. Preregistration and the disclosure

of analytical flexibility have both been proposed as two of four approaches to the forking-paths problem (Rubin, 2017). However, nonpreregistered multiverse analyses may allow post hoc selection of pipelines whereby unfavorable results could be omitted. Furthermore, nontransparent and unsystematic selection of pipelines before computation can be biased toward those expected to produce a desired result. Combining preregistration with transparent and systematic pipeline selection offers a powerful safeguard against selective reporting by ensuring all defensible pipelines are disclosed even if it does not resolve all inferential challenges associated with forking paths (e.g., Rubin, 2017). Accordingly, prespecification of pipelines is encouraged (e.g., Ankel-Peters et al., 2024), and transparent, systematic documentation of the decision-making process can ease the cognitive and logistical load (e.g., Short, Inceler, et al., 2025).

Practical guidance for preregistering multiverse analyses remains limited. The availability of numerous preregistration templates for the single-pipeline approach (e.g., a variety of targeted templates are provided on the OSF) has likely contributed to the widespread adoption of preregistrations. However, relatively few templates exist for multiverse analyses, for which multiple pipelines and the decision-making procedure used to identify them should be documented. Although an initial preregistration template for multiverse analyses has been published (Flournoy et al., n.d.), it is customized for neuroimaging data. However, a newly developed documentation tool supports such preregistration across multiple scientific disciplines (Short, Inceler, et al., 2025). In addition, the number of preregistered multiverse analyses is increasing (e.g., Beauducel et al., 2025; Sabey, 2023), offering structural examples. Consequently, support for preregistering multiverse analyses is growing.

Computational challenges

Multiverse analysis can impose substantial computational demands, particularly when the multiverse of pipelines is large and complex combinations must be navigated. This is further exacerbated when multiple software packages or programming languages are required, which can introduce compatibility issues, fragmented workflows, and pipeline-specific errors that can be relatively difficult to identify and resolve compared with global errors. Consequently, computational feasibility becomes a methodological concern and a technical one with implications for transparency, completeness, and interpretability.

Several strategies have been proposed to address these challenges. For example, parallel processing can substantially reduce run time by distributing pipelines across cores or computing nodes, and a modular

workflow design can improve efficiency, reproducibility, and error tracing. Tools such as Boba (Liu et al., 2020) and Multiverse (Sarma et al., 2023) implement this structure by breaking down pipelines into modular components. When the exhaustive computation of all pipelines becomes infeasible, machine-learning-based sampling approaches can infer the outcomes of unsampled pipelines based on the sampled ones (e.g., Dafflon et al., 2022; Kristanto et al., 2023; Short, Hildebrandt, et al., 2025). Transparent reporting of computational constraints, optimization strategies, and approximation methods and careful validation of these methods provide computational support and enable readers to evaluate the scope, limitations, and interpretability of multiverse results.

Defensibility and equivalence assessments

The assessment of defensibility and equivalence remains a methodological challenge in multiverse analysis. Theoretical approaches, such as those based on literature precedent, expert consensus, or many-analyst designs, are currently the predominant means of establishing defensibility (e.g., Breznau et al., 2022; Jacobsen et al., 2025; Kristanto et al., 2023; Paul et al., 2022; Trübutschek et al., 2024; Wacker, 2017). Although these approaches draw on accumulated disciplinary knowledge, they necessarily rely on researcher judgment and may reflect publication bias or prevailing conventions in a field. Empirical approaches to defensibility and equivalence assessments are increasingly discussed but remain comparatively underdeveloped. Empirical defensibility assessments evaluate whether individual pipelines produce data or estimates that meet minimum adequacy criteria (e.g., acceptable signal-to-noise ratio or psychometric properties). In contrast, empirical equivalence assessments involve relative comparisons across defensible pipelines and evaluate whether alternative pipelines yield sufficiently similar results with respect to a researcher-defined inferential target (e.g., measurement, effect, or power/precision equivalence; Del Giudice & Gangestad, 2021) given an explicitly stated threshold. Although both approaches involve criteria and thresholds, their objectives differ: Defensibility concerns whether a pipeline is acceptable in isolation, whereas equivalence concerns whether pipelines are sufficiently comparable to be treated as interchangeable for a given analytical purpose. Standardized implementation guidelines to assess empirical defensibility and empirical equivalence remain underdeveloped. This represents a critical direction for future work that could improve transparency and comparability and strengthen robustness claims in multiverse analyses.

Statistically handling the similarity of pipelines

Pipelines that share common options, such as the same normalization technique, may exhibit greater covariance than pipelines that share fewer options, particularly if the shared options heavily influence the result. This can lead to pseudoreplication, whereby the dependency structure of the variance is unequal across pipelines, inflating the Type I error rate. Essentially, the same variation in the data is being tested repeatedly across pipelines that are not fully independent. Such overlap may also create a false impression of robustness because pipelines sharing influential analytical options can yield empirically similar data sets, as discussed in Dataset Similarity Across Pipelines.

To mitigate this challenge, researchers should avoid treating pipelines as independent samples (e.g., applying a traditional ordinary-least-squares-based analysis across all results) and should avoid treating the results based on a theory of repeated sampling. As described above, nuanced statistical methods are being developed (e.g., Girardi, 2024). When the goal is to identify key moderating variables and qualitatively compare underlying models and their assumptions, organizing the results into smaller components may be useful. This can be achieved at the hypothesis-test level by identifying that different estimands are being tested and running separate multiverses for each (e.g., Auspurg & Brüderl 2021) and at the model level by analyzing the variance or using set theory to identify model components that lead to certain results and seeking to understand these differences using theory or causal-inference logic (e.g., Breznau et al., 2022). Because variation in multiverse results is systematically generated by analytical decisions, it can, in principle, be explained by those decision parameters. Machine-learning models, such as random forests, gradient boosting, and permutation-based feature-importance methods, may offer promising strategies for identifying and ranking the most influential decision nodes or combinations of options given that decision nodes can interact in nonadditive ways. We view this as a promising avenue for future methodological innovation. However, it is important to note the potential for reduced interpretability in some instances, such as when the model is nonlinear, contains hundreds of features, or relies on complex architectures, particularly when post hoc explanations are required (Rudin, 2019; Rudin et al., 2022). These approaches highlight the need for careful evaluation of multiverse-analysis results. A standardized approach to interpretation is yet to be developed.

Ongoing Debates

While preparing this article, it became clear that researchers from different disciplines may hold different opinions

on some aspects of multiverse analysis. These varying opinions highlight the importance of bringing cohesion across disciplines in terminology and standards to facilitate clear and effective debate and interdisciplinary innovation. Here, we summarize the collaborative positions we reached, in line with the existing literature.

Using multiverse analysis to identify the “best” pipeline

Although there appears to be a broad consensus that the primary goals of multiverse analysis are to report uncertainty and assess robustness, there is an ongoing debate about whether and how multiverse analyses can or should be used to identify an optimal data-processing pipeline. It has been proposed that empirically evaluating pipelines against predefined quality criteria, such as reducing measurement noise or improving psychometric reliability, can prune the multiverse of pipelines within a specific analytical context (e.g., Clayson et al., 2021), thereby supporting empirical assessment of defensibility or equivalence. Importantly, such evaluations do not aim to identify a universally superior pipeline or recover a ground truth but, rather, to refine the space of defensible pipelines for a particular population, component, and study design. However, caution is warranted when selecting the criteria for such assessments. In particular, effect size is not an appropriate criterion for ranking pipelines because larger or more consistent effects can arise from measurement bias rather than by improved validity or accuracy when the true effect is unknown (Feuerriegel & Bode, 2022). Accordingly, although empirical assessments of pipeline quality can improve context-specific pruning of the multiverse when appropriate criteria are used, they do not justify general claims about the “best” pipeline without additional theoretical justification, simulation evidence, or cross-validation.

Inclusion of covariates as a decision node

The inclusion of covariates as a decision node in a multiverse analysis is a key point of debate. One perspective treats the inclusion of covariates analogously to other analytical decisions, arguing that varying covariates across pipelines can illuminate the sensitivity of results to plausible modeling choices, particularly in observational research, in which confounding is a primary concern. However, Del Giudice and Gangestad (2021) cautioned that alternative covariate sets often do not represent different ways of answering the same question; rather, they correspond to substantively different estimands. From this perspective, pipelines that differ in covariate structure alter the underlying model structure and no longer estimate the same causal effect or association. As a result, pipelines are rendered effect nonequivalent and should

not be interpreted as interchangeable within a single multiverse (Del Giudice & Gangestad, 2021). Recent work grounded in causal-inference principles, although not addressing multiverse analysis specifically, supports this view, demonstrating that conditioning on different variables may open or close causal paths, introduce bias, and fundamentally alter the causal quantity being estimated (Chen et al., 2024). Consequently, varying covariates across pipelines may shift the focus from assessing the robustness of a single model to comparing substantively different models, complicating interpretation if causal equivalence is not explicitly addressed. Ultimately, the decision of whether to include covariates as a decision node in a multiverse analysis depends on the research aims. If the primary aim is to explore how different, theoretically defensible modeling choices correspond to different estimands, including covariates in the multiverse analysis may be appropriate, provided the results are interpreted in a causal framework. However, if the aim is to assess the robustness of results from a single model, covariate variation should be handled with caution, potentially through separate multiverses based on effect equivalence.

Existence and measurement of equivalence between options or pipelines

There is ongoing debate surrounding the existence and measurement of equivalence between pipelines in a multiverse analysis. Some argue that analytically distinct pipelines can reasonably be treated as equivalent with respect to a clearly defined criterion, such as testing the same effect with comparable validity or precision (Del Giudice & Gangestad, 2021), despite functional differences in how the data are processed. However, others caution that any functional variation necessarily alters the data in nonidentical ways and therefore conflicts with a strict interpretation of “equivalence.” This concern is particularly pertinent when equivalence is operationalized using point or interval thresholds. Under such approaches, pipelines that yield nonidentical results may nonetheless be classified as equivalent if they fall within a prespecified bound. Critics argue that this usage risks semantic confusion because the term “equivalence” may be interpreted as implying identical values rather than practical interchangeability relative to a defined criterion. Relatedly, there is debate about how equivalence should be represented. Del Giudice and Gangestad (2021) proposed a discrete classification framework comprising Types E, N, and U as a pragmatic tool for decision-making, whereas others have suggested that equivalence is better conceptualized as a continuum. Regardless of whether equivalence is treated categorically or continuously, it remains essential that

researchers clearly document the rationale, assumptions, and thresholds underlying equivalence judgments to avoid misinterpretation.

Conclusion

Multiverse analysis is a powerful tool for assessing robustness and articulating uncertainty in empirical research, but its application requires transparency, systematicity, and careful consideration of the data and design. In this article, we aimed to promote interdisciplinary cohesion by providing comprehensive guidance for conducting multiverse analysis appropriate for a variety of scientific disciplines. As multiverse analysis continues to evolve, several key challenges and opportunities arise for future attention. First, further development is required in the area of empirical assessments of defensibility and equivalence, and interdisciplinary collaborations in this respect could provide well-informed, multiperspective standards that can be applied across disciplines. Second, advancing statistical methods that account for variable nonindependence between pipelines is a promising area for methodological innovation. This could involve the integration of causal-inference logic, dependence-adjusted aggregation techniques, and machine-learning models to detect influential decision nodes. Third, improving transparency in how and why defensible options are included or excluded and developing empirically grounded metrics for measuring both dataset and pipeline similarity will facilitate valid interpretation. Finally, achieving consensus on best practices will require continued interdisciplinary dialogue, shared open-source tools, and a collective commitment to transparent reporting. Together, these developments could lead to a more rigorous, standardized, and collaboratively developed framework for multiverse analysis.

Glossary

Multiverse analysis: A method that systematically explores and evaluates the impact of multiple defensible analytical decisions on research outcomes, such as those relating to data processing or model specification, to assess robustness and report on methodological and analytical uncertainty.

*Decision node*³: A specific step in the research workflow (pipeline) in which the researcher must decide to either include or exclude the step or a step in which multiple alternative options are available for selection (e.g., outlier detection).

*Option*³: One of the alternative choices available at a given decision node (e.g., a method for outlier detection, such as ± 3.29 SD from the group mean).

Pipeline: One sequential combination of options from the first to the last decision node, representing one complete workflow alternative.

Defensible: An option or pipeline can be considered defensible if it is rationally justifiable within the context of a single-pipeline analysis. Although this can be partly subjective, the evaluation can be guided by domain knowledge, accepted methodological approaches, theory, application in recent peer-reviewed literature, or empirical assessment.

Equivalent: Options or pipelines are considered equivalent when relative to a researcher-specified criterion, they can be reasonably treated as interchangeable for the purpose of a given multiverse analysis. Equivalence is assessed relatively and can rely on, for example, theoretical expectations, expert consensus, or empirical assessment.

Pipeline similarity/analytical similarity: The degree to which analytical pipelines share overlapping decision nodes, decision-node sequences, and options; leading to structural or methodological dependence between pipelines.

Dataset similarity/empirical similarity: The degree of overlap or correlation between data sets produced by different analytical pipelines, reflecting how similar the resulting data are in their empirical properties.

Transparency

Action Editor: Karoline Huth

Editor: David A. Sbarra

Author Contributions

Cassie Ann Short: Conceptualization; Methodology; Project administration; Supervision; Visualization; Writing – original draft; Writing – review & editing.

Nate Breznau: Writing – original draft; Writing – review & editing.

Maria Brunsch: Visualization; Writing – original draft; Writing – review & editing

Micha Burkhardt: Visualization; Writing – original draft; Writing – review & editing.

Niko A. Busch: Writing – original draft; Writing – review & editing.

Elena Cesnaite: Writing – original draft; Writing – review & editing.

Maximilian Frank: Writing – original draft; Writing – review & editing.

Carsten Gießing: Writing – original draft; Writing – review & editing.

Daniel Krähmer: Writing – original draft; Writing – review & editing.

Daniel Kristanto: Writing – original draft; Writing – review & editing.

Tina Lonsdorf: Writing – original draft; Writing – review & editing.

Claudia Neuendorf: Writing – original draft; Writing – review & editing.

Hung H. V. Nguyen: Writing – original draft; Writing – review & editing.

Manuel Rausch: Writing – original draft; Writing – review & editing.

Xenia Schmalz: Writing – original draft; Writing – review & editing.

Andreas Schneck: Conceptualization; Writing – original draft; Writing – review & editing.

Cem Tabakci: Writing – original draft; Writing – review & editing.

Andrea Hildebrandt: Conceptualization; Visualization; Writing – original draft; Writing – review & editing.

Declaration of Conflicting Interests

The authors declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This work was made possible by the Deutsche Forschungsgemeinschaft (DFG)-funded priority program “META-REP” (SPP 2317). C. A. Short and D. Kristanto were funded by the projects HI 1780/7-1 and GI 682/5-1 awarded to coauthors C. Gießing and A. Hildebrandt within the SPP 2317. N. Breznau and H. H. V. Nguyen were supported by DFG Grant No. 464546557. X. Schmalz was supported by DFG Grant Nos. 464350745 and 456356582. C. Tabakci was supported by DFG Grant Nos. RA 2988/4-1 and ZE 887/9-1. D. Krähmer was supported by DFG Grant Nos. AU 394/5-1 and BR 1898/4-2. M. Rausch was supported by DFG Grant No. RA 2988/4-1. C. Neuendorf was supported by the Alexander von Humboldt Professorship endowed by the German Federal Ministry of Education and Research to Kou Murayama and by the German Federal Ministry of Education and Research (Grant No. 01JG2106A). M. Brunsch and T. Lonsdorf were supported by DFG Grant Nos. LO/1980/4-1 and LO/1980/10-1 awarded to T. Lonsdorf. N. A. Busch and E. Cesnaite were supported by the DFG Grant BU 2400/11-1.

ORCID iDs

Cassie Ann Short  <https://orcid.org/0000-0001-7799-7783>
 Claudia Neuendorf  <https://orcid.org/0000-0002-3024-0000>
 Xenia Schmalz  <https://orcid.org/0000-0002-3365-257X>

Acknowledgments

A preprint version of this article was uploaded to MetaArXiv: https://osf.io/preprints/metaarxiv/4yzeh_v2.

Notes

1. “Alpha level” is the threshold probability (commonly set at .05 or 5% or less) used to determine whether to reject the null hypothesis. It represents the maximum acceptable risk of making a Type I error—incorrectly rejecting a true null hypothesis.
2. The formula to assess the threat of a confounding variable is $\text{expected.corr}(x,y) = \text{corr}(x,y) - (\text{corr}(x,cv) \times \text{corr}(y,cv))$.
3. We prefer the terms “decision node” and “option” to the corresponding terms “parameter” and “parameter value” that were proposed by Hall et al. (2022) because the term “parameter” is usually used to indicate the population estimate within a statistical model.

References

- Allard, T. L., Canada, K. L., Botdorf, M., & Riggins, T. (2023). Longitudinal exploration of binding ability across early childhood: The differential contribution of hits and false alarms. *The Journal of Genetic Psychology*, 184(6), 385–398. <https://doi.org/10.1080/00221325.2023.2213268>
- Ankel-Peters, J., Brodeur, A., Dreber, A., Johannesson, M., Neubauer, F., & Rose, J. (2024). *A protocol for structured robustness reproductions and replicability assessments* (No. 143). Institute for Replication. <https://www.econstor.eu/handle/10419/301917>
- Auspurg, K., & Brüderl, J. (2021). Has the credibility of the social sciences been credibly destroyed? Reanalyzing the “Many Analysts, One Data Set” Project. *Socius*, 7. <https://doi.org/10.1177/23780231211024421>
- Bartoš, F., Hoogeveen, S., Sarafoglou, A., & Pawel, S. (2025). *Single-dataset meta-analysis for many-analysts and multiverse studies*. arXiv. <https://doi.org/10.48550/arXiv.2511.17064>
- Beauducel, A., Hennig, J., Hewig, J., Hildebrandt, A., Härpfer, K., Kührt, C., Lange, L., Mueller, E., Munk, A., Osinsky, R., Paul, K., Carsten, H. P., Porth, E., Riesel, A., Rodrigues, J., Rommerskirchen, L., Scheffel, C., Short, C., Stahl, J., . . . Wacker, J. (2025). *Preregistrations*. OSF. osf.io/rdpze
- Beauducel, A., Scheuble-Cabrera, V., Hennig, J., Hewig, J., Hildebrandt, A., Kührt, C., Lange, L., Mueller, E. M., Osinsky, R., Paul, K., Porth, E., Riesel, A., Rodrigues, J., Scheffel, C., Short, C., Stahl, J., Strobel, A., & Wacker, J. (2024). The association of dispositional anxiety with the NoGo N2 under relaxation instruction vs. speed/accuracy instruction. *Biological Psychology*, 192, Article 108850. <https://doi.org/10.1016/j.biopsycho.2024.108850>
- 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., . . . Rieck, J. R. (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., 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., . . . Van Assche, J. (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), Article e2203150119. <https://doi.org/10.1073/pnas.2203150119>
- Brodeur, A., Cook, N., Hartley, J., & Heyes, A. (2024). *Do pre-registration and pre-analysis plans reduce p-hacking and publication bias? Evidence from 15,992 test statistics and suggestions for improvement*. *Journal of Political Economy Microeconomics*, 2(3), 527–561. <https://doi.org/10.1086/730455>
- Bryan, C. J., Yeager, D. S., & O'Brien, J. M. (2019). Replicator degrees of freedom allow publication of misleading failures to replicate. *Proceedings of the National Academy of Sciences*, 116(51), 25535–25545. <https://doi.org/10.1073/pnas.1910951116>
- Burkhardt, M., & Gießing, C. (2026). The comet toolbox: Improving robustness in network neuroscience through multiverse analysis. *Imaging Neuroscience*, 4, Article IMAG.a.1122. <https://doi.org/10.1162/IMAG.a.1122>
- Burkhardt, M., Hildebrandt, A., Gießing, C., & Kristanto, D. (2024). Quantifying similarity between graph-theoretic resting-state fMRI data processing pipelines for efficient multiverse analysis. *Brainiacs Journal of Brain Imaging and Computing Sciences*, 5(2), Article Edoc XEE8F298E. <https://doi.org/10.48085/XEE8F298E>
- Chen, G., Cai, Z., & Taylor, P. A. (2024). Through the lens of causal inference: Decisions and pitfalls of covariate selection. *Aperture Neuro*, 4. <https://doi.org/10.52294/001c.124817>
- Clayson, P. E. (2024). Beyond single paradigms, pipelines, and outcomes: Embracing multiverse analyses in psychophysiology. *International Journal of Psychophysiology*, 197, Article 112311. <https://doi.org/10.1016/j.ijpsycho.2024.112311>
- Clayson, P. E., Baldwin, S. A., Rocha, H. A., & Larson, M. J. (2021). The data-processing multiverse of event-related potentials (ERPs): A roadmap for the optimization and standardization of ERP processing and reduction pipelines. *NeuroImage*, 245, Article 118712. <https://doi.org/10.1016/j.neuroimage.2021.118712>
- Dafflon, J., Da Costa, P. F., Váša, F., Monti, R. P., Bzdok, D., Hellyer, P. J., Turkheimer, F., Smallwood, J., Jones, E., & Leech, R. (2022). A guided multiverse study of neuroimaging analyses. *Nature Communications*, 13(1), Article 3758. <https://doi.org/10.1038/s41467-022-31347-8>
- 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). <https://doi.org/10.1177/2515245920954925>
- Dragicevic, P., Jansen, Y., Sarma, A., Kay, M., & Chevalier, F. (2019). Increasing the transparency of research papers with explorable multiverse analyses. In *CHI '19: Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1–15). Association for Computing Machinery. <https://doi.org/10.1145/3290605.3300295>
- Dreber, A., & Johannesson, M. (2025). A framework for evaluating reproducibility and replicability in economics. *Economic Inquiry*, 63(2), 338–356. <https://doi.org/10.1111/ecin.13244>
- Duschek, S., de Guevara, C. M. L., Serrano, M. J. F., Montoro, C. I., López, S. P., & Reyes del Paso, G. A. (2022). Variability of reaction time as a marker of executive function impairments in fibromyalgia. *Behavioural Neurology*, 2022(1), Article 1821684. <https://doi.org/10.1155/2022/1821684>
- Engzell, P., & Mood, C. (2023). Understanding patterns and trends in income mobility through multiverse analysis. *American Sociological Review*, 88(4), 600–626. <https://doi.org/10.1177/00031224231180607>
- Feuerriegel, D., & Bode, S. (2022). Bring a map when exploring the ERP data processing multiverse: A commentary on Clayson et al. 2021. *NeuroImage*, 259, Article 11944. <https://doi.org/10.1016/j.neuroimage.2022.119443>

- Flournoy, J., Barendse, M., Byrne, M., Cosme, D., & Pfeifer, J. (n.d.). *Pre-registration of multiverse analyses*. <https://johnflournoy.science/multiverse-preregistration/>
- Framework for Open and Reproducible Research Training. (n.d.). *Glossary*. <https://fortr.org/glossary/english/>
- 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*. Department of Statistics, Columbia University. <https://sites.stat.columbia.edu/gelman/research/unpublished/forking.pdf>
- Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, 102(6), 460–465. <https://www.americanscientist.org/article/the-statistical-crisis-in-science>
- Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, 8(2), 53–96. <https://doi.org/10.1111/j.1539-6053.2008.00033.x>
- Girardi, P., Vesely, A., Lakens, D., Altoè, G., Pastore, M., Calcagni, 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>
- Götz, M., Sarma, A., & O'Boyle, E. H. (2024). The multiverse of universes: A tutorial to plan, execute and interpret multiverses analyses using the R package multiverse. *International Journal of Psychology*, 59(6), 1003–1014. <https://doi.org/10.1002/ijop.13229>
- 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>
- Haller, S., Monsch, A. U., Richiardi, J., Barkhof, F., Kressig, R. W., & Radue, E. W. (2014). Head motion parameters in fMRI differ between patients with mild cognitive impairment and Alzheimer disease versus elderly control subjects. *Brain Topography*, 27(6), 801–807. <https://doi.org/10.1007/s10548-014-0358-6>
- Hoffmann, S., Schönbrodt, F., Elsas, R., Wilson, R., Strasser, U., & Boulesteix, A. L. (2021). The multiplicity of analysis strategies jeopardizes replicability: Lessons learned across disciplines. *Royal Society Open Science*, 8(4), Article 201925. <https://doi.org/10.1098/rsos.201925>
- Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), Article e124. <https://doi.org/10.1371/journal.pmed.0020124>
- Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology*, 19(5), 640–648. <https://doi.org/10.1097/EDE.0b013e31818131e7>
- Jacobsen, N. S., Kristanto, D., Welp, S., Inceler, Y. C., & Debener, S. (2025). Preprocessing choices for P3 analyses with mobile EEG: A systematic literature review and interactive exploration. *Psychophysiology*, 62(1), e14743. <https://doi.org/10.1111/psyp.14743>
- Kozak, J. P., Zhang, R., Porter, M., Song, Q., Liu, J., Wang, B., Wang, R., Saito, W., & Zhang, Y. (2023). Stability, reliability, and robustness of GaN power devices: A review. *IEEE Transactions on Power Electronics*, 38(7), 8442–8471. <https://doi.org/10.1109/TPEL.2023.3266365>
- Krähmer, D., & Young, C. (2026). Visualizing vastness: Graphical methods for multiverse analysis. *PLOS ONE*, 21(2), Article e0339452. <https://doi.org/10.1371/journal.pone.0339452>
- Kristanto, D., Burkhardt, M., Thiel, C. M., Debener, S., Gießing, C., & Hildebrandt, A. (2024). The multiverse of data pre-processing and analysis in graph-based fMRI: A systematic literature review of analytical choices fed into a decision support tool for informed analysis. *Neuroscience & Biobehavioral Reviews*, 165, Article 105846. <https://doi.org/10.1016/j.neubiorev.2024.105846>
- Kristanto, D., Gießing, C., Marek, M., Zhou, C., Debener, S., Thiel, C., & Hildebrandt, A. (2023, October). An extended active learning approach to multiverse analysis: Predictions of latent variables from graph theory measures of the human connectome and their direct replication. In *2023 Guardians Workshop (Guardians)* (pp. 1–13). IEEE. <https://doi.org/10.48085/J962E0F53>
- Kuhn, M., Gerlicher, A. M., & Lonsdorf, T. B. (2022). Navigating the manyverse of skin conductance response quantification approaches—A direct comparison of trough-to-peak, baseline correction, and model-based approaches in Ledalab and PsPM. *Psychophysiology*, 59(9), Article e14058. <https://doi.org/10.1111/psyp.14058>
- Leamer, E. E. (1985). Sensitivity analyses would help. *The American Economic Review*, 75(3), 308–313. <https://www.jstor.org/stable/1814801>
- Lefort-Besnard, J., Nichols, T. E., & Maumet, C. (2025). Statistical inference for same data meta-analysis in neuroimaging multiverse analyzes. *Imaging Neuroscience*, 3, Article imag_a_00513. https://doi.org/10.1162/imag_a_00513
- Liu, Y., Kale, A., Althoff, T., & Heer, J. (2020). 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>
- Lonsdorf, T. B., Gerlicher, A., Klingelhöfer-Jens, M., & Kryptos, A. M. (2022). Multiverse analyses in fear conditioning research. *Behaviour Research and Therapy*, 153, Article 104072. <https://doi.org/10.1016/j.brat.2022.104072>
- Masur, P. K., & Scharrow, M. (2020). *spectr: Conducting and visualizing specification curve analyses* [Software]. <https://cran.r-project.org/package=spectr>
- Muñoz, J., & Young, C. (2018). We ran 9 billion regressions: Eliminating false positives through computational model robustness. *Sociological Methodology*, 48(1), 1–33. <https://doi.org/10.1177/0081175018777988>
- Neumayer, E., & Plümper, T. (2017). *Robustness tests for quantitative research*. Cambridge University Press.
- Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., . . . Yarkoni, T. (2015). Promoting an open research culture. *Science*, 348(6242), 1422–1425. <https://doi.org/10.1126/science.aab2374>
- OECD. (2022). *PISA Programme for International student assessment data*. <https://datahub.aalto.fi/en/data-sources/oecd-pisa-programme-for-international-student-assessment-data>

- Olsson-Collentine, A., van Aert, R. C. M., Bakker, M., & Wicherts, J. (2025). Meta-analyzing the multiverse: A peek under the hood of selective reporting. *Psychological Methods*, 30(3), 441–461. <https://doi.org/10.1037/met0000559>
- Ozenne, B., Nørgaard, M., Pernet, C., & Ganz, M. (2025). A sensitivity analysis of preprocessing pipelines: Toward a solution for multiverse analyses. *Imaging Neuroscience*, 3, Article imag_a_00523. https://doi.org/10.1162/imag_a_00523
- Patel, C. J., Burford, B., & Ioannidis, J. P. (2015). Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology*, 68(9), 1046–1058. <https://doi.org/10.1016/j.jclinepi.2015.05.029>
- Paul, K., Beauducel, A., Hennig, J., Hewig, J., Hildebrandt, A., Kührt, C., Lange, L., Mueller, E. M., Osinsky, R., Porth, E., Riesel, A., Rodrigues, J., Scheffel, C., Short, C. A., Stahl, J., Strobel, A., & Wacker, J. (2025). Frontal alpha asymmetry as a marker of approach motivation? Insights from a cooperative forking path analysis. *Journal of Personality and Social Psychology*, 128(1), 196–210. <https://doi.org/10.1037/pspp0000503>
- Paul, K., Short, C. A., Beauducel, A., Carsten, H. P., Härpfer, K., Hennig, J., Hewig, J., Hildebrandt, A., Kührt, C., Mueller, E. M., Munk, A., Osinsky, R., Porth, E., Riesel, A., Rodrigues, J., Scheffel, C., Stahl, J., Strobel, A., & Wacker, J. (2022). The methodology and dataset of the coscience EEG-personality project—A large-scale, multi-laboratory project grounded in cooperative forking paths analysis. *Personality Science*, 3(1), Article e7177. <https://doi.org/10.5964/ps.7177>
- Pergher, V., Wittevrongel, B., Tournoy, J., Schoenmakers, B., & Van Hulle, M. M. (2018). N-back training and transfer effects revealed by behavioral responses and EEG. *Brain and Behavior*, 8(11), Article e01136. <https://doi.org/10.1002/brb3.1136>
- Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, 74(4), 525–556. <https://doi.org/10.3102/00346543074004525>
- Ramsey, J. B. (1969). Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 31(2), 350–371. <https://doi.org/10.1111/j.2517-6161.1969.tb00796.x>
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592. <https://doi.org/10.1093/biomet/63.3.581>
- Rubin, M. (2017). An evaluation of four solutions to the forking paths problem: Adjusted alpha, preregistration, sensitivity analyses, and abandoning the Neyman-Pearson approach. *Review of General Psychology*, 21, 321–329. <https://doi.org/10.1037/gpr0000135>
- Rubin, M. (2024). *Preregistration does not improve the transparent evaluation of severity in Popper's philosophy of science or when deviations are allowed*. arXiv. <https://doi.org/10.48550/arXiv.2408.12347>
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
- Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistical Surveys*, 16, 1–85. <https://doi.org/10.1214/21-SS133>
- Sabey, H. (2023). *Pre-registration: A multiverse analysis on Wu et al. (2020): The influence of male faces on stereotype activation among women in STEM*. OSF. <https://doi.org/10.17605/OSF.IO/BEVHG>
- Saltelli, A., Aleksankina, K., Becker, W., Fennell, P., Ferretti, F., Holst, N., Li, S., & Wu, Q. (2019). Why so many published sensitivity analyses are false: A systematic review of sensitivity analysis practices. *Environmental Modelling & Software*, 114, 29–39. <https://doi.org/10.1016/j.envsoft.2019.01.012>
- Sarma, A., Hwang, K., Hullman, J., & Kay, M. (2024). Milliways: Taming multiverses through principled evaluation of data analysis paths. In *CHI '24: Proceedings of the 2024 Conference on Human Factors in Computing Systems* (pp. 1–15). Association for Computing Machinery. <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. In *CHI '23: Proceedings of the 2023 Conference on Human Factors in Computing Systems* (pp. 1–15). Association for Computing Machinery. <https://doi.org/10.1145/3544548.3580726>
- Short, C. A., Hildebrandt, A., Bosse, R., Debener, S., Özyağcılar, M., Paul, K., Wacker, J., & Kristanto, D. (2025). Lost in a large EEG multiverse? Comparing sampling approaches for representative pipeline selection. *Journal of Neuroscience Methods*, 424, Article 110564. <https://doi.org/10.1016/j.jneumeth.2025.110564>
- Short, C. A., Inceler, Y. C., Frank, M., & Hildebrandt, A. (2025). The Systematic Multiverse Analysis Registration Tool (SMART) for defining multiverse analyses. *Royal Society Open Science*, 12(1), Article 250800. <https://doi.org/10.1098/rsos.250800>
- Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtry, 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., Cervantes, I. F., . . . 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, 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>
- Steege, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse anal-

- ysis. *Perspectives on Psychological Science*, 11(5), 702–712. <https://doi.org/10.1177/1745691616658637>
- Trübutschek, D., Yang, Y. F., Gianelli, C., Cesnaite, E., Fischer, N. L., Vinding, M. C., Marshall, T. R., Algermissen, J., Pascarella, A., Puoliväli, T., Vitale, A., Busch, N. A., & Nilsson, G. (2024). EEGManyPipelines: A large-scale, grassroots multi-analyst study of electroencephalography analysis practices in the wild. *Journal of Cognitive Neuroscience*, 36(2), 217–224. https://doi.org/10.1162/jocn_a_02087
- University of Essex. (2023). *British Household Survey, 2023 dataset*. UK Data Service. <https://www.ukdataservice.ac.uk>
- Voracek, M., Tran, U. S., & Kern, J. (2025). *Mapping the landscape of multiverse-style data-analytic methods: A scoping review*. PsyArXiv. https://doi.org/10.31234/osf.io/g4zck_v1
- Wacker, J. (2017). Increasing the reproducibility of science through close cooperation and forking path analysis. *Frontiers in Psychology*, 8, Article 1332. <https://doi.org/10.3389/fpsyg.2017.01332>
- Young, C., & Cumberworth, E. (2025). *Multiverse analysis: Computational methods for robust results*. Cambridge University Press.
- 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>
- Young, E. S., & Vermeent, S. (2024). *Multitool: Run multiverse style analyses*. <https://doi.org/10.32614/CRAN.package.multitool>