Meta-analyzing the multiverse

Olsson-Collentine, Anton; van Aert, Robbie C.M.; Bakker, Marjan; Wicherts, Jelte

Published in:
Psychological Methods

DOI:
10.1037/met0000559

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

Link to publication in Tilburg University Research Portal

Citation for published version (APA):
Olsson-Collentine, A., van Aert, R. C. M., Bakker, M., & Wicherts, J. (in press). Meta-analyzing the multiverse: A peek under the hood of selective reporting. *Psychological Methods*. https://doi.org/10.1037/met0000559

Psychological Methods

Meta-Analyzing the Multiverse: A Peek Under the Hood of Selective Reporting

Anton Olsson-Collentine, Robbie C. M. van Aert, Marjan Bakker, and Jelte Wicherts

Online First Publication, May 11, 2023. https://dx.doi.org/10.1037/met0000559

CITATION

Olsson-Collentine, A., van Aert, R. C. M., Bakker, M., & Wicherts, J. (2023, May 11). Meta-Analyzing the Multiverse: A Peek Under the Hood of Selective Reporting. *Psychological Methods*. Advance online publication. https://dx.doi.org/10.1037/met0000559

AMERICAN
PSYCHOLOGICAL
ASSOCIATION

Psychological Methods

ISSN: 1082-989X

https://doi.org/10.1037/met0000559

Meta-Analyzing the Multiverse: A Peek Under the Hood of Selective Reporting

Anton Olsson-Collentine, Robbie C. M. van Aert, Marjan Bakker, and Jelte Wicherts
Department of Methodology and Statistics, Tilburg School of Social and Behavioral Sciences, Tilburg University

OPEN DATA

OPEN MATERIALS

Abstract

Researcher degrees of freedom refer to arbitrary decisions in the execution and reporting of hypothesis-testing research that allow for many possible outcomes from a single study. Selective reporting of results (*p*-hacking) from this "multiverse" of outcomes can inflate effect size estimates and false positive rates. We studied the effects of researcher degrees of freedom and selective reporting using empirical data from extensive multistudy projects in psychology (Registered Replication Reports) featuring 211 samples and 14 dependent variables. We used a counterfactual design to examine what biases could have emerged if the studies (and ensuing meta-analyses) had not been preregistered and could have been subjected to selective reporting based on the significance of the outcomes in the primary studies. Our results show the substantial variability in effect sizes that researcher degrees of freedom can create in relatively standard psychological studies, and how selective reporting of outcomes can alter conclusions and introduce bias in meta-analysis. Despite the typically thousands of outcomes appearing in the multiverses of the 294 included studies, only in about 30% of studies did significant effect sizes in the hypothesized direction emerge. We also observed that the effect of a particular researcher degree of freedom was inconsistent across replication studies using the same protocol, meaning multiverse analyses often fail to replicate across samples. We recommend hypothesis-testing researchers to preregister their preferred analysis and openly report multiverse analysis. We propose a descriptive index (underlying multiverse variability) that quantifies the robustness of results across alternative ways to analyze the data.

Translational Abstract

Researcher degrees of freedom refer to arbitrary decisions in the execution and reporting of research that could create many possible outcomes in a single study, sometimes called a multiverse of outcomes. If researchers compute several outcomes but selectively choose to report only some based on statistical significance, this can lead to bias in reported outcomes as they appear in publications and later meta-analyses that summarize results across many studies on the same topic. We used data from strictly controlled large-scale studies in social and cognitive psychology to examine what biases might have emerged had they not been strictly controlled, which is the case with most research in the literature. In the 211 samples we studied, we find large variations in potential outcomes depending on analytic decisions and demonstrate that selective reporting among outcomes could severely bias meta-analytic summaries, despite meta-analysis being considered a "gold-standard" of evidence. Hence, our results call for a need for meta-analyses to evaluate primary studies for risk of bias due to selective reporting and for original studies to lower the risk of bias by careful registration of analytic choices before studies are conducted.

Supplemental materials: https://doi.org/10.1037/met0000559.supp

Anton Olsson-Collentine https://orcid.org/0000-0002-4948-0178

This research was supported by a Consolidator Grant 726361 (IMPROVE) from the European Research Council (ERC, https://erc.europa.eu), awarded to J.M. Wicherts. The results of this study were presented as an online talk at the IOPS Winter conference (December 2020) and at the Metascience 2021 conference (September 2021), and have been made available as a preprint (https://psyarxiv.com/43yae/). All code and data for this project are available on the OSF at http://osf.io/j8yg2/ and permanently archived at Zenodo (https://zenodo.org/record/7341292#.ZEs8P3ZBxdg). Preregistrations of code and coding protocol is available at: http://osf.io/h397y/.

The authors have no financial or nonfinancial competing interests.

Anton Olsson-Collentine developed the study concept. All authors contributed to the study design. Anton Olsson-Collentine wrote the code and analyzed the data. All the authors interpreted the results and contributed to the writing of the manuscript, with Anton Olsson-Collentine writing the first draft and Anton Olsson-Collentine, Jelte Wicherts and Marjan Bakker writing the revision. All the authors approved the final manuscript for submission.

The data are available at https://doi.org/10.5281/zenodo.7341292.

The experiment materials are available at https://doi.org/10.5281/zenodo.7341292

Correspondence concerning this article should be addressed to Anton Olsson-Collentine, Department of Methodology and Statistics, Tilburg School of Social and Behavioral Sciences, Tilburg University, Warandelaan 2, 5037 AB Tilburg, The Netherlands. Email: anton@olssoncollentine.com

Researcher Degrees of Freedom (DFs; Simmons et al., 2011) refer to the many arbitrary decisions that need to be made in designing, collecting, analyzing, and reporting research. In the analysis of hypothesis-testing research, the focus of this article, researcher DFs involve decisions such as choosing between different approaches for dealing with missing observations, excluding participants from the analysis depending on different criteria, and a range of other data processing and modeling decisions (for more examples, see Wicherts et al., 2016). Researcher DFs allow for many possible outcomes in a single study where the reported result depends on the specific combination of decisions made. This was illustrated recently by Silberzahn et al. (2018): when 29 independent teams examined the same data with the same research question, the teams' estimated effect sizes (measured as odds ratios) varied from 0.89 to 2.93, with 20 teams finding a statistically significant effect in the expected direction. That different independent teams of researchers reached different estimates shows that there often is no clearly preferable analysis in hypothesis-testing research (see also Botvinik-Nezer et al., 2020; Breznau et al., 2021; Huntington-Klein et al., 2021; Wagenmakers et al., 2022).

The many possible statistical results that are enabled by researcher DFs have been referred to as a "multiverse" of statistical results (Steegen et al., 2016), "vibration of effects" (Patel et al., 2015), or a "specification curve analysis" (Simonsohn et al., 2020). These multiverse style analyses entail (sensitivity) analyses of the robustness of results to researcher DFs and offer insights into potential biases that might emerge if researchers selectively report outcomes from them to present more convincing evidence in favor of a hypothesized effect. The proponents of multiverse style analyses are careful to define "reasonable" or "arbitrary" decisions in light of substantive, statistical, methodological, and psychometric grounds (Del Giudice & Gangestad, 2021; Patel et al., 2015; Simonsohn et al., 2020; Steegen et al., 2016). In the current study, we use multiverse analysis based on arbitrary choices to demonstrate the potential impact of selective reporting on study-level effects and subsequent meta-analyses of resulting effect sizes. To ensure arbitrariness in our researcher DFs, we consider the effect size computations to be given. That is, we apply researcher DFs that we consider to not change the independent or dependent variables, and we do not add covariates or change the statistical model or constructs of interest.

From the perspective of the broader literature, the principal concern with researcher DFs is not that they allow multiple statistical results to be computed, but rather that they allow for selective reporting of possibly desirable outcomes. Throughout this article, we use "selective reporting" to refer to cases where multiple statistical results are examined in a study, but some go unreported (Page et al., 2020). We do not include in this definition the special case where no study results are reported and hence do not focus on publication bias of entire studies. Selective reporting is often focused on the significance of outcomes and can be intentional ("*p*-hacking") or happen unintentionally due to hindsight and confirmation biases (Nickerson, 1998; Roese & Vohs, 2012). Selective reporting from the multiverse of statistical results is problematic as it can allow researchers to present statistical evidence even for incredible phenomena (Simmons et al., 2011). Numerous formal approaches and simulation studies have been used to show that selective reporting leads to an overrepresentation of false positive findings (Ioannidis, 2005) and inflated effect size estimates (Ioannidis, 2008) in the literature.

Unfortunately, selective reporting appears common among researchers. In psychology, about 50% to 60% of researchers admit to not reporting all dependent measures in a study (Agnoli et al., 2017; John et al., 2012), and in a study registry comparison 70% of studies did not report all outcome variables (Franco et al., 2016). Moreover, there is extensive literature on selective reporting in the fields of biomedicine, with evidence from, for example, neurology (Fusar-Poli et al., 2014), hematology (Wayant et al., 2017), pediatrics (Rosati et al., 2016), orthopedics (Rongen & Hannink, 2016), obesity (Rankin et al., 2017), and cancer research (Kyzas et al., 2005). A recent study examining results in 67 trials published between October and November 2015 in five top journals from general medicine found that 42% of prespecified outcomes went unreported (Goldacre et al., 2019). Further evidence from the fields of education (Pigott et al., 2013) and studies on partner violence (Madden et al., 2019) suggests the problem of selective reporting is widespread indeed.

The biases created by selective reporting in primary studies are inherited by meta-analyses that seek to quantitatively review effects or associations across many studies. Each of the studies included in a meta-analysis has its own multiverse. Since the results used for meta-analysis are subsets from these multiverses, meta-analytic result(s) also represent a subset from the multiverse of possible meta-analyses. To avoid that this subset is biased, meta-analytic reporting guidelines such as PRISMA (Moher et al., 2009) and MARS (Appelbaum et al., 2018) recommend meta-analysts to evaluate primary studies for selective reporting. We do not consider arbitrary choices made in the context of meta-analyses themselves (i.e., multiverse meta-analysis: Palpacuer et al., 2019; Voracek et al., 2019), but rather vary the analyses in the primary studies while keeping the meta-analytic inclusion criteria and analysis constant (i.e., we meta-analyze multiverses) to study the biasing effects of selective reporting based on researcher DFs in primary studies on meta-analytic outcomes.

Such biasing effects have been studied in simulated data for meta-analysis (e.g., Botella et al., 2021; Carter et al., 2019; Friese & Frankenbach, 2020), but simulated data from known distributions may not be representative of actual psychological data that feature unknown (distributional) complexities. Also, the effects of researcher DFs have been studied in observed data of individual studies (Botvinik-Nezer et al., 2020; Breznau et al., 2021; Huntington-Klein et al., 2021; e.g., Silberzahn et al., 2018), but not in meta-analytic context to inform how they might affect cumulative knowledge. We combine these streams of research and study the effects of researcher DFs and selective reporting in observed meta-analytic data, taking advantage of the unique opportunity offered by the open data of 10 recent multilab direct replication projects in psychology (Registered Replication Reports [RRRs]) that featured a total of 211 samples studying 14 different outcome variables.

RRRs each consist of a set of studies (labs) that collected data on an effect in psychology using the same prespecified research design, decision plan, and materials, collectively known as a "preregistration." Each RRR can be seen as making up one (or more) meta-analysis of direct (also called "exact") replications, where the only difference between included studies is where they collected their data. Even though the preregistrations used in the actual RRRs limited the effect of researcher DFs in the original analyses, the open data from these extensive studies enable us to use a counterfactual design to see what biases could have emerged if the studies (and ensuing meta-analyses) had not been preregistered and could have been subjected to selective reporting based on the significance of the outcomes in the primary studies. In doing so, we demonstrate the variability in results that may arise in meta-analytic data in the absence of preregistration, the limitations of multiverse analysis when applied to a single study, and illustrate the entire process of selective reporting, from the researcher DFs in primary studies that enable the practice to the consequences for meta-analysis.

Method

Figure 1 summarizes the design of this study. We identify decision points in each RRR where reasonable alternative decisions could have been made (absent any preregistration) and compute all resulting outcomes (create a multiverse) for each included lab. We then combine effect sizes from the lab multiverses in meta-analysis within each RRR. This design allows us to explore the effects of researcher DFs on research output by (a) examining the underlying multiverse variability (UMV) in effect size estimates at the primary study level, (b) examining the resulting multiverse variability at the meta-analytic level, and (c) examining different mechanisms for selecting effect sizes from primary study multiverses for inclusion in the meta-analysis. We refer to the variability due to researcher DFs as the UMV(statistically defined in "The Multiverses" section).

Transparency and Openness

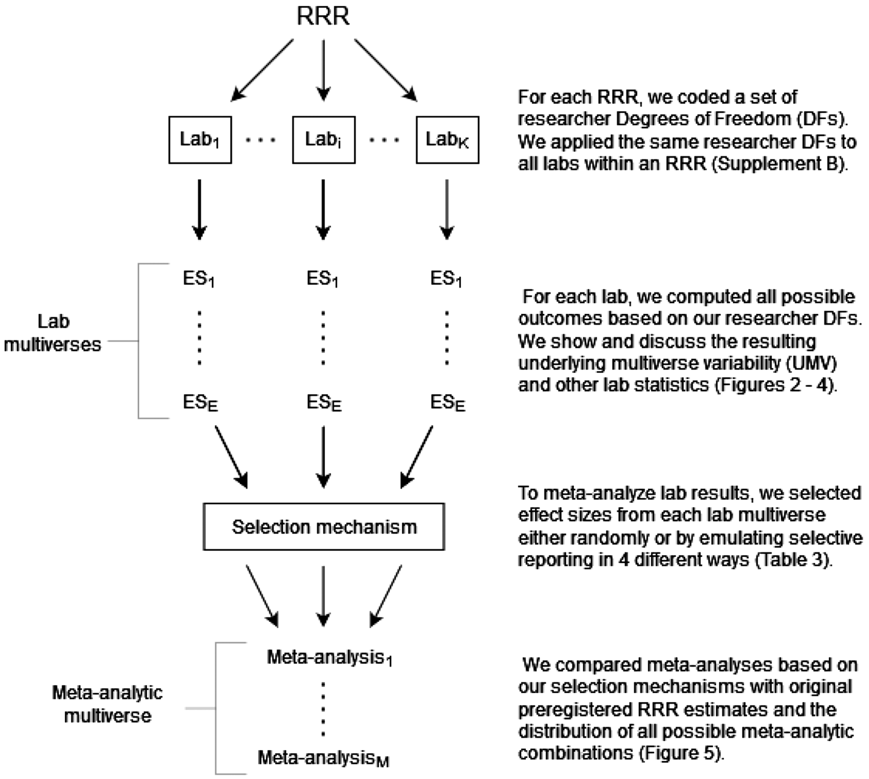
All our code and data for this project are available on the Open Science Framework (OSF) at http://osf.io/j8yg2/ (Olsson-Collentine et al., 2019), and permanently archived at Zenodo at doi.org/10.5281/zenodo.7341292 (Olsson-Collentine et al., 2022). We refer directly to relevant files on the OSF using brackets and links in the sections below. We registered the data cleaning code and the researcher DFs available in each RRR before proceeding to analysis (http://osf.io/h397y/). We only made minor code corrections and clarifications of researcher DFs after registration, fully detailed in Supplement A (in the online supplemental material; http://osf.io/xem2y/). We handled all data in R version 4.0.2 (R Core Team, 2020), and cite used packages in the reference list (Henry & Wickham, 2020; Revelle, 2020; van Aert, 2020; Wickham, 2016; Wickham & Bryan, 2019; Wickham et al., 2020; Wickham & Miller, 2020; Zhu, 2019).

Data Collection

We included all 10 RRRs available at the time of data collection (i.e., published up until May 15, 2019) available in the journals "*Perspectives on Psychological Science*" and "*Advances in Methods and Practices in Psychological Science*" (see also Olsson-Collentine et al., 2020). Three RRRs (RRR3, RRR5, and RRR9) had multiple primary outcome variables (as explicitly identified in the accompanying publications). In total, we included 10 projects containing 14 primary

Figure 1

Summary of the Study Design



For each RRR, we coded a set of researcher Degrees of Freedom (DFs). We applied the same researcher DFs to all labs within an RRR (Supplement B).

For each lab, we computed all possible outcomes based on our researcher DFs. We show and discuss the resulting underlying multiverse variability (UMV) and other lab statistics (Figures 2 - 4).

To meta-analyze lab results, we selected effect sizes from each lab multiverse either randomly or by emulating selective reporting in 4 different ways (Table 3).

We compared meta-analyses based on our selection mechanisms with original preregistered RRR estimates and the distribution of all possible meta-analytic combinations (Figure 5).

Note. RRR = Registered Replication Report; ES = effect size; DF = degree of freedom. For each lab in an RRR, multiverses were computed, analyzed, and used for meta-analysis. Each RRR consists of K labs, lab $i = 1, 2, 3, \ldots K$. Each lab has E effect size estimates in its multiverse. There are M possible combinations of the E effect sizes across labs, resulting in a meta-analytic multiverse of size M. We approximate the meta-analytic multiverse by randomly sampling 10^5 meta-analyses from the meta-analytic multiverse. For details on how we selected researcher DFs, see Methods under the header "Selection and justification of researcher DFs." This figure was created using the website draw.io.

outcome variables that could be meta-analyzed, consisting of 211 unique lab samples and 34,357 participants (Table 1). These values correspond to the sum of the labs and participants of RRR1—RRR9, as RRR9 and RRR10 used the same sample.

We use the RRRs because the meta-analyses they offer allow us to consider the effects of (selective reporting from) multiverse analyses in relation to a benchmark based on meticulously collected data from multiple labs (and different samples) using the same protocol. We selected the RRRs for our study over other multilab replication initiatives for two reasons. First, we wished to examine researcher DFs within a meta-analytic structure, which is what the RRRs nicely provide. The RRRs have the additional advantage that most of them report average results not significantly different from zero, allowing us to examine the bias from selective reporting under the most problematic circumstances (i.e., when there is no genuine effect) and the percentage of significant outcomes appearing across multiverses

Table 1
Preregistered Multilab Replication Projects

RP	Paper	Countries	Labs	Effects	*N*	Sample and settings	Description of effects
RRR1	Alogna et al. (2014)	10	31	1	4,832	31 out of 32 samples were undergraduate students aged 18–25, 1 general population which was also the only online sample.	Verbal overshadowing 1; independent two-group experiment. Participants either described a robber after watching a video or listed countries/capitals and after a filler task attempted to identify the robber in a lineup.
RRR2	Alogna et al. (2014)	8	26	1	2,932	22 out of 23 samples were undergraduate students aged 18–25, 1 general population which was also the only online sample.	Verbal overshadowing 2; different from 1 only in that the filler task took place before the descriptive task instead of after.
RRR3	Eerland et al. (2016)	2	10	3	1,210	11 out of 12 samples were undergraduate students mostly aged 18–25, one of which was online. One sample was a broader online sample.	Grammar's effect on interpretation; independent two-group vignette experiment with three outcome variables. Participants read about actions either described in the imperfect or perfect tense and then ratedthe protagonist's intentions (intentionality/intention attribution/detailed processing).
RRR4	Hagger et al. (2016)	10	24	1	3,127	All samples consisted of in-lab undergraduate students	Ego depletion; independent two-group experiment. Participants were either assigned to a cognitively demanding or a neutral task, and performance was then measured in a subsequent cognitive task.
RRR5	Cheung et al. (2016)	5	16	2	2,279	All samples consisted of in-lab undergraduate students aged 18–25	Commitment on neglect/exit; independent two-group experiment with two outcome variables. Participants were either primed to think about commitment to or independence from their partner.
RRR6	Wagenmakers et al. (2016)	8	17	1	2,542	All but one sample explicitly consisted of students and all took place in-lab. The last sample was recruited on university grounds.	Facial feedback hypothesis; independent two-group experiment. Participants were either induced to "smile" or "pout" by holding a pen in their mouth differently and simultaneously rated the funniness of cartoons.
RRR7	Bouwmeester et al. (2017)	12	21	1	3,669	All samples consisted of in-lab undergraduate students aged 18–34.	Intuitive cooperation; independent two-group experiment. Economic game with money contributed to a common pool either under time pressure or time delay.
RRR8	O'Donnell et al. (2018)	13	40	1	7,041	All samples consisted of in-lab undergraduate students aged 18–25	Professor priming; independent two-group experiment. Participants were primed with either a "professor" or "hooligan" stimuli. The outcome was percentage correct trivia answers.
RRR9	McCarthy et al. (2018)	13	26	2	6,720	All samples consisted of in-lab students aged 18–25	Hostility priming; independent two-group experiment with two outcome variables. Participants descrambled sentences, either 20% or 80% were hostile, then rated an individual and a list of ambiguous behaviors on perceived hostility.
RRR10	Verschuere et al. (2018)	12	25	1	3,245	All samples consisted of in-lab students aged 18–25	Moral reminder; independent two-group experiment. Participants either recalled the 10 Commandments or books they had read. The outcome was a degree of cheating when reporting results.

Note. All RRRs published up until May 15, 2019, in the journals *Perspectives on Psychological Science* and *Advances in Methods and Practices in Psychological Science*.' RP = Replication Project; Countries = number of lab country locations; Effects = number of primary effects studied; *N* = participants before exclusions; RRR = Registered Replication Report. Adapted from "Heterogeneity in Direct Replications in Psychology and its Association With Effect Size," by A. Olsson-Collentine, J. M. Wicherts, and M. A. L. M. van Assen, 2020, *Psychological Bulletin*, *146*(10), 922–940 (https://doi.org/10.1037/bul0000294). Copyright 2020 by the American Psychological Association. Code to reproduce table available at: https://osf.io/jehpy/.

(Type I error rate). Second, we wished to allow for researcher DFs to depend on the study design. The Many Labs series of replication projects (which we have worked with previously in Olsson-Collentine et al., 2020) consists of many effects studied at the same time in the same samples, meaning (almost) all researcher DFs will be identical across all studied effects. Hence the RRRs allow us to delve deep into the generalizability of the multiverse variability across labs and effects.

We downloaded individual-level data for all RRRs in Table 1. Summary data of all RRRs were available on the OSF. When the raw lab data were not publicly available via the OSF we contacted authors by email to request them. Only for one lab in RRR1 and RRR2 and two labs in RRR3 were we unable to acquire their individual-level raw data.

For each RRR, we standardized data formatting across labs, fixed minor mistakes (e.g., mislabeled columns in RRR8), and prepared the datasets for multiverse analysis (http://osf.io/cf86y/). We prepared the data in the same way as the original RRRs to the extent possible. However, we largely had to write our own code, because the alternative decisions needed to create our multiverses (e.g., exclusion criteria) could not be taken in the code by the original authors. In preparing the datasets for analysis, we only excluded participants due to reported experimenter error or when participants were reported to not have followed instructions or completed the experiment. Note that exclusions based on "not following instructions" are usually *ad hoc*, and hence are distinct from formalized exclusions based on attention checks.

The Multiverses

Although multiverse type analyses have been suggested by multiple authors under somewhat different names (Patel et al., 2015; Simonsohn et al., 2020; Steegen et al., 2016) all multiverse analyses consist of identifying points in the research process where multiple reasonable decisions could have been made, identifying what these decisions might be, and examining the impact of these decisions on the study results. A core assumption of multiverse type analysis is that the alternative decisions are all (approximately) equally reasonable (Simonsohn et al., 2020; Steegen et al., 2016; see also Del Giudice & Gangestad, 2021).

It is important that researcher DFs are defined such that these choices are indeed "reasonable" or "arbitrary" on a priori substantive and methodological grounds (Patel et al., 2015; Simonsohn et al., 2020; Steegen et al., 2016; ?) rightly pointed out that many decisions implemented in published multiverse analyses were not truly arbitrary (are "nonequivalent") because they can a priori be expected to result in different (a) measurement reliability/validity, (b) studied psychological effects (e.g., when including a covariate that changes the prediction), or (c) power/precision. Only decisions for which this does not hold (i.e., we are either confident they are equivalent or uncertain) should according to Del Giudice & Gangestad (2021) be included in a multiverse analysis.

We agree with Del Giudice & Gangestad (2021) that such substantive and methodological issues should be considered when performing multiverse analysis as a sensitivity analysis. Hence, we carefully selected our researcher DFs to reflect choices for which we saw no prior substantive or methodological grounds to expect them to affect the true effect sizes tapped by the different labs. However, we could imagine that others might object to some of those choices and hence we offer a range of supplementary results to assess how alternative choices in designing the multiverse affect our results. We also make our data and code available for reanalysis using alternative multiverse setups.

Selection and Justification of Researcher DFs

We selected our researcher DFs to correspond to normative researcher behavior that is at risk of selective reporting in the fields of the RRRs that make up our data. These RRRs belong to the fields of social and cognitive psychology (Olsson-Collentine et al., 2020). Although there are many researcher DFs before analyzing the data (Wicherts et al., 2016), due to using already collected data we were only able to vary post-data collection decisions. Moreover, because our focus was on researcher DFs in primary studies and their consequences for downstream meta-analysis, we only varied decisions in data processing (the data multiverse, Steegen et al., 2016) and not the statistical models used in data analysis (the model multiverse). Consequently, several chosen researcher DFs concerned with using different exclusion criteria (which we prepend with "E"), although we also varied how the composite score was computed from multiple indicators (researcher DFs prepended with "S").

When creating our list of researcher DFs, we proceeded in two steps: we (a) set up a list of "common" researcher DFs, and then (b) set up a list of researcher DFs unique to each RRR. These were then combined to create our final list of researcher DFs for each RRR, which we registered before analyzing any data (Supplement B in the online supplemental material; http://osf.io/wj38n/). Because all labs in an RRR used the same design, it was only necessary to identify decisions and create associated options once for each RRR and not for each lab/study separately. Our coded researcher DFs each consisted of a decision that needs to be made and several associated potential options for that decision. When defining the researcher DFs, we explored the data in the sense of examining whether potential DFs could be applied (i.e., whether the variables existed and how they were defined) but did not examine what effect applying them would have.

We created our list of common researcher DFs based on recommendations in statistical textbooks, common decisions by applied researchers as reported in the research literature, data analytic decisions made by the included RRRs, and our own experience of decisions encountered in the literature. Table 2 provides an overview of all common researcher DFs. We considered option (a) across decisions to be the default option, corresponding to no scale adjustments or participant exclusions (although for some researcher DFs, an active decision must be made; S2, E1 Table 2). In Supplement C in the online supplemental material, we detail how we selected each researcher DF and its options. We acknowledge that many additional multiverses could be run in these and other studies, but we consider our setup typical of researcher DFs that could be used in practice across a range of psychological studies, and as such useful to study the influence of selective reporting. Researchers instead interested in using multiverse analysis as a sensitivity analysis for a particular effect should carefully consider the advice of Del Giudice & Gangestad (2021) on equivalent pathways before applying any of the researcher DFs in Table 2.

In addition to the list of common researcher DFs, which we applied to all RRRs, each RRR had several unique researcher DFs. These arise from the uniqueness of each research topic and

Table 2
List of Common Researcher Degrees of Freedom Applied to Registered Replication Reports

Decision	Options	Short explanation
S1. Post hoc scale length	(a) No adjustment	It is unclear how common it is to post hoc drop items "that don't work" from a scale, but dropping more than a few seem unlikely. In the research, we are looking at (experimental) there are rarely long scales. Excluding 1-item scales, the average scale length in a large sample of psychological research in 2014 was 6.87 ($SD = 7.18$; Flake et al., 2017)
	(b) Drop the item with the lowest item-rest correlation	
	(c) Drop the two items with the lowest item-rest correlations	
S2. Composite score	(a) Unweighted average score	For Likert-type scales with multiple items. Other DVs, for example, reaction time variability (RRR3), dichotomous correct/incorrect (RRR1&2), continuous measures (RRR7), and single item DVs (RRR10) may need more unique choice options. We chose Varimax rotation to maximize the variance between outcomes.
	(b) Sum score	
	(c) PCA score: Varimax rotation, force two components and pick the first, requires at least three items.	
E1. Missingness DV	(a) Any missing items → list-wise deletion	List-wise deletion appears to be by far the most common approach to missing data. In van Ginkel et al.'s (2010) review of personality psychology 97% used list-wise deletion for missing data and several reviews in medicine have also found it to be an extremely common method (Burton & Altman, 2004; Rombach et al., 2016; ?). Nonetheless, we see that, for example, RRR6 used pair-wise deletion (option b) which may seem reasonable to some researchers, in particular with a longer scale.
	(b) If ≤25% items missing then pair-wise deletion of missing items. Otherwise list-wise.	
E2. Missingness E3–E4 variables	(a) No exclusion	RRRs that excluded data based on an E3–E4 variable (e.g., age) did list-wise deletion when data was missing. For other non-DV variables, we make no exclusions based on missingness, unless this was explicitly done by the project (e.g., "task completion" RRR10).
	(b) Missing on any variable → list-wise deletion	
E3. Age	(a) No exclusion	Used by 9 out of 10 RRRs for exclusions. Across 25 cohorts of Dutch bachelor psychology students 96.7% of students were below 25, 92.7% below 24, 86.6% below 23, and 77.8% below 22 (Wicherts, 2012). The oldest students in this dataset were 25. We choose a set of age ranges based on these data that we believed might go unremarked if used as exclusion criteria in the psychological literature with an accompanying motivation such as "we only included young adults."
	(b) Not 18–24	
	(c) Not 18–23	
	(d) Not 18–22	
	(e) Not 18–21	
E4. Language/student/ethnicity	(a) No exclusion	Used by 3 out of 10 RRRs. Demographic variables are sometimes used for exclusions. Language includes variables such as "native speaker," which may have a yes/no response. Ethnicity includes similar variables such as "country of birth" or "race." If multiple of these demographic variablesis available they are treated as separate exclusion criteria.
	(b) Exclude participants not belonging to the dominant category	
E5. Attention check	(a) No exclusion	Attention checks are common in psychology, as evidenced by the more than 1,500 citations of Oppenheimer et al. (2009) who introduced "instructional manipulation checks." Curran (2016) suggests "conservative" exclusions based on 50% failed attention checks when multiple items are used. This category does not include manipulation checks which vary more in format.
	(b) Exclude if failed > 50% of attention check items (i.e., with two items, must fail both, e.g., RRR7)	
	(c) Exclude if failed any attention check item	

E6. Univariate outliers	(a) No exclusion (b) DV score > 2 *SD* from mean (c) > 3 *SD* from mean (d) > 1.5 times the interquartile range	Used by 1/10 RRRs. Commonly recommended cutoffs (Bakker & Wicherts, 2014). Test for outliers across groups.
E7. Multivariate outliers	(a) No exclusion (b) Mahalanobi's distance with $p < .001$	If the outcome variable is a correlation. Tabachnick et al. (2007) recommend using Mahalanobi's distance with a cutoff of $p < .001$ for detecting multivariate outliers. Outliers tested within groups as recommended by Tabachnik et al.

Note. *S.* = degree of freedom (DF) affecting scale composition, *E.* = exclusion DF, DV = dependent variable, *SD* = standard deviation, PCA = principal component analysis, RRR = Registered Replication Report. Code to reproduce table: http://osf.io/jehpy/.

design and consist of different exclusion criteria. We coded between 2 (RRR3) and 10 (RRR7) unique researcher DFs for each RRR, each decision with 2 to 6 associated options. Due to the large number of unique researcher DFs, we do not describe them all in detail here, but provide only a broad overview and refer interested readers to Supplement B in the online supplemental material.

We can separate between two types of "unique" researcher DFs: either (a) the RRRs excluded participants based on some variable that was not defined in our list of common decisions or (b) an RRR measured variables (not in our list of common decisions) that they could have used for exclusions. As an example of the first case, in RRR4 (ego-depletion) participants with < 80% correct on the main task were excluded. However, 80% is a largely arbitrary number, and someone might also consider values such as 75%, 85%, 90%, or many others, in addition to no exclusions. In cases like these, when there are an infinite number of possible values to choose from, we have elected only a maximum of six possible values that we believe an applied researcher would reasonably pick.

As an example of where an RRR measured variables they could have used for exclusions: RRR5 (commitment to romantic partner), amongst other things, asked participants whether they lived within 60 miles of their partner (yes/no) but did not use this variable in their analysis. However, another researcher might have found it relevant to only consider participants (not) living close to each other and use this variable for exclusions. Collecting data on a variable with no clear purpose thus adds researcher DFs and increases the risk of selective reporting, which we in this case used to create our multiverses.

Applying Researcher DFs to the RRRs

After registering the coding protocol for "common" researcher DFs, we coded the applicability of each common researcher DFs to each RRR, which differed depending on, for example, how the outcome variable was measured (binary vs. continuous, one item vs. a scale) and how projects coded their data. Because some labs within RRRs prescreened their participants for the original RRR exclusion criteria, it was not always possible to apply all exclusion criteria to all labs in an RRR. Nonetheless, we still included such labs, prioritizing the inclusion of more labs over the possibility of less multiverse variation. The coded common and unique researcher DFs for all RRRs are available in Supplement B (in the online supplemental material; http://osf.io/wj38n/).

We computed resulting effect sizes from all possible combinations of decisions for each lab in an RRR (http://osf.io/zhdrx/). Incompatible decision combinations were not applied. For example, if we wished to drop two items from a scale (Table 2; S1c) but required at least three items in the scale for principal component analysis (Table 2; S2c), this decision combination was inapplicable to scales with fewer than 5 items. We standardized mean differences (Cohen's *d*, Borenstein, 2009, p. 226) and computed log odds ratios for RRR1 and RRR2. Effect sizes were originally analyzed unstandardized in all RRRs except for RRR4, and if certain researcher DF lowers the within-sample variance, as is highly likely, then standardized effect sizes will appear larger. However, because certain of our researcher DFs change the dependent variable, and we wanted to draw conclusions across RRRs, it was necessary to standardize effect sizes. As most meta-analyses use standardized effect sizes and we are interested here in the biasing effects of selective reporting on typical unregistered meta-analyses, we do not consider standardization of effect sizes a major concern for our analysis.

To prevent including lab multiverses with an unrealistically small number of participants, we only included lab multiverses with at least 24 participants per experimental group, the median sample size in psychology (Bakker et al., 2012), in our primary analyses. Three labs in RRR2 (L09, L17, and L26) and one lab in RRR8 (L24) had smaller sample sizes than required in all conditions and were excluded from these analyses. We present the results of our analyses also without this sample size restriction in Supplement D in the online supplemental material.

Analysis

A consequence of assuming that the alternative decisions in the multiverse are equally reasonable is that under the null hypothesis that no researcher DF has a systematic effect (i.e., is an actual moderator of the effect) we can consider the distribution of effect sizes in the multiverse as random variability around a true score. We refer to the variability underlying a given set of researcher DFs as the UMV and define it as the standard deviation (*SD*) in effect size estimates that are in the multiverse of the same study. A different set of researcher DFs will reveal different UMVs. Other researchers have focused on the distribution of *p*-values (Simonsohn et al., 2020; Steegen et al., 2016) or on the range of effect sizes in the multiverse (e.g., Patel et al., 2015), but we consider it more useful to treat multiverse variability in terms of the standard deviation of effect size estimates, in line with how sampling error is defined. The UMV should be seen as a descriptive tool that highlights some degree of variability that might have a relation with bias due to selective reporting over and beyond sampling error, rather than a well-defined statistic.

To demonstrate the effects of researcher DFs on research output, we (a) examined the variance in effect size estimates at the lab level (lab multiverses) and (b) compared meta-analytic average effect size estimates based on how lab outcomes were selected from their multiverses. To examine the variance in effect size estimates at the lab level, we created funnel plots, computed UMV, and standard deviations in effect size resulting from variation across the options within a single researcher DF. For the funnel plots, we plotted all effect size estimates at the lab level using either the standard error (for log odds ratios; RRR1/RRR2) or sample size as the *y*-axis (for standardized mean difference effect sizes; RRR3–RRR10). We used total sample size (*N*) on the *y*-axis for all standardized mean difference (SMD) effect types since most of our coded researcher DFs affected sample size.

To examine how large the effects of applying a single researcher DF can be and the relative impact of our different researcher DFs, we computed the standard deviation in a lab's estimated effect size across the options associated with each decision. For each researcher DF, we computed the standard deviation in effect size when all other researcher DFs were set to their default value (corresponding to option "a" for each researcher DF, see Table 2 and Supplement B in the online supplemental material). In addition to examining standard deviations for labs within RRRs, we also disaggregated these lab estimates across RRRs and then aggregated them across common and unique researcher DF categories. In doing so, we treated all unique researcher DFs as one category.

Due to computational limitations, and because it is often the case that some researcher DF must be applied before another (e.g., outliers cannot be removed before the composite score has been computed), we only applied the researcher DFs in a single fixed order. That is, if we have three researcher DFs (1, 2, and 3) then we always applied them in the order 1, 2, and 3 regardless of the chosen option, rather than also varying the order (e.g., 2, 1, and 3). This fixed order may affect results when removing items with the lowest item-rest correlation from a scale or excluding participants based on outlier criteria, although we see no reason to expect a systematic interaction between these two and any other researcher DFs. The fixed order also makes it impossible to compute the impact of a single researcher DF across all possible researcher DF combinations, although it remains possible to compute its impact when not applying any other researcher DFs (see previous paragraph).

When comparing meta-analytic average estimates, we compared (a) the original (preregistered) RRR estimates, with (b) an estimate of the distribution of all possible meta-analytic combinations, (c) randomly selected lab effect sizes, and (d) lab effect sizes selected by one of four biased selection mechanisms (see below). We ran all meta-analyses as random-effects models with the restricted maximum likelihood estimator for estimating the between-study variance using the R-package "metafor" (Viechtbauer, 2010).

The huge number of possible effect size combinations across labs for each RRR, the smallest consisting of 697×10^{33} possible meta-analyses, made it impossible to compute the full distributions of possible meta-analytic outcomes. Instead, we drew large random samples to approximate the distributions. For each RRR (or outcome variable when an RRR contained multiple primary outcomes), we proceeded as follows: we drew one random effect size from all possible effect sizes from Lab 1, one random effect size from all possible effect sizes from Lab 2, one from Lab 3, one from Lab 4, and so on until we had drawn one effect size from all labs in an RRR. The drawn effect sizes across labs were then combined using a meta-analysis. We repeated this procedure, sampling with replacement from each lab's multiverse of effect sizes, until we had sampled 100,000 effect sizes from each lab, and consequently computed 100,000 meta-analyses for each RRR. These samples of meta-analyses constituted our approximation of the distribution of possible meta-analyses for each RRR (or outcome variable when an RRR contained multiple primary outcomes). The means of these distributions (and the means of the estimated lower/upper 95% confidence intervals [CIs]) constituted our random sample of estimates.

When selectively reporting results, researchers may exhibit different behavior. We included four types of biased selection mechanisms (Table 3: "Most significant," "Below α," "Random significant," and "Bounded significant") with different motivations. All selection mechanisms focused on statistical significance, and we used a two-tailed test with $\alpha = .05$ for hypothesis testing. First, we selected the effect size with the lowest *p*-value in each lab. This allowed us to examine the most extreme selection of results possible due to *p*-hacking ("most significant"). We included this scenario as a worst-case scenario. Second, selective reporting may sometimes result in a "bump" just below $p = .05$ when aggregating *p*-values across selectively reported studies ("below α"). This is most likely in the case of incremental *p*-hacking approaches such as optional stopping (e.g., Hartgerink, 2017). To compare what a meta-analysis of such data might look like, in the "below α" condition, we selected, for each lab in an RRR, the outcome with a *p*-value closest below .05

Table 3
Summary of Outcome Selection Mechanisms

Selection mechanism	Hypothesized direction filter	Single outcome	Description
Pre-registered	No	Yes	The original RRR meta-analytic average effect size with preregistered decisions.
Random draw	No	No	The average point estimate and upper/lower 95% CI from 10^5 meta-analyses randomly sampled from all possible meta-analyses.
Most significant	Yes	Yes	Select the effect size in the multiverse with the smallest p-value.
Below α	Yes	Yes	Select the effect size in the multiverse with a p-value closest below $p = 0.05$. If no p value is below the cutoff, pick the smallest.
Random significant	Yes	No	Identical to the random draw, but with effect sizes first limited to only significant effect sizes.
Bounded significant	Yes	No	We drew 100 effect sizes from a lab's multiverse, and selected the effect size with the lowest p-value. This was repeated 10^5 times, resulting in 10^5 values per lab. These were then meta-analyzed and summarized as above for the random draw.

Note. Description of different implemented selection mechanisms for selecting effect sizes at the lab level to meta-analyze. "Hypothesized Direction Filter" = exclude effect sizes not in the predicted direction (yes/no); "Single outcome" = selection mechanism resulting in a single meta-analytic result (yes/no). Code to reproduce the table is available at: http://osf.io/jehpy/.

(or, if there were no p-values below .05, the lowest value). These two approaches ("most significant" and "below α") attempted to select a single result from the multiverse, but it may be that several effect sizes have equivalent p-values due to being based on exactly the same sample. If so, we picked the effect size with the fewest researcher DFs deviating from their default option a.

Third, we represent a p-hacker who is satisfied with any significant effect size they encounter (in the expected direction), by picking a random effect size out of those that were statistically significant ("random significant"). If no effect sizes were significant, the effect size with the lowest p-value was picked. Fourth and finally, when a p-hacking researcher tries multiple analyses, they might choose to report the analysis that resulted in the smallest p-value. However, selecting the result with the smallest p-value across the full multiverse suggests that the p-hacker systematically explored the full multiverse to find the strongest possible effect, whereas reality probably consists of a more ad hoc and limited search. Hence, we represent a "bounded" search by (a) randomly drawing 100 possible outcomes and (b) out of these 100 outcomes selecting the one with the smallest p-value.

With all our biased selection mechanisms (i.e., excluding the random draw and original meta-analytic results, see Table 3), we applied a "hypothesized direction filter." That is, when selecting an effect size at the lab level, we excluded all effect sizes that were in the opposite direction of the originally predicted effect (http://osf.io/r2dum). If there were no effect sizes in the predicted direction, we excluded all significant effect sizes in the "wrong" direction and selected outcomes from the remainder. We added this filter because we believe researchers who apply selective reporting, in reality, are unlikely to be agnostic about the direction of their effect, and our focus in this study is on selective reporting, not Hypothesizing After Results are Known (HARKing, Kerr, 1998)

Publication bias, the complete suppression of a study being published, and selective reporting (selection of reported results amongst multiple possibilities) are closely related, and it is intuitively appealing to believe correcting for publication bias may be sufficient for generally removing biases in the meta-analytic data (e.g., Kvarven et al., 2019). We applied three publication bias correction methods to examine their applicability to selective reporting in the absence of publication bias. These were the "precision-effect test and precision-effect estimate with standard errors" (PET-PEESE; Stanley & Doucouliagos, 2014), p-uniform* (van Aert & van Assen, 2020), and Vevea and Hedge's 3-Parameter Selection Model (3PSM; Vevea & Hedges, 1995) implemented using the R-package "weightr" (Coburn and Vevea, 2019).

Results

After excluding conditions that resulted in fewer than 24 participants per experimental group, 8 out of 14 RRR multiverses decreased in size, as can be seen in Table 4. The absolute decrease was largest for the largest multiverses (RRR05 and RRR07), with RRR07 showing the largest absolute decrease and decreasing from 2,621,440 to 525,680 (an 80% decrease) potential outcomes. However, the proportionally largest decrease was seen in RRR08, which decreased from 115,200 to 19,200 (83% decrease) potential outcomes. More importantly, as evidenced by the median number of multiverses per lab, even in the RRRs with relatively few researcher DFs, the researcher DFs jointly created thousands of alternative outcomes per lab. Nonetheless, many labs found zero significant (at $p = .05$) effect sizes within their multiverses. Across all studies (counting labs with multiple dependent variables (DVs) as separate studies), 205 out of 294 (70%) encountered no significant effect sizes in the hypothesized direction in their multiverses and 134 out of 294 (46%) found no significant effect sizes in any direction. The size of the multiverse was strongly correlated with the number of studies that encountered significant effect sizes in the hypothesized direction within their multiverses (Pearson's $r =$ 0.63, 95% CI [0.14, 0.87]).

Lab Multiverses

There can be substantial variation in effect sizes within labs due to researcher DFs. Figure 2 shows effect sizes across the multiverses for 16 out of 24 labs in RRR04. Similar plots for all RRRs (or outcome variables when an RRR contains multiple), including for all labs in RRR04, can be found in Supplement E (in the online supplemental material; http://osf.io/2htc6/).

Table 4
Multiverse Sizes Before and After Filtering Out Outcomes with <24 Participants Per Experimental Group

Meta-analysis	Common DFs	Unique DFs	*N* [lq, uq]	Multiverse size before exclusion	Multiverse size after exclusion	Labs after exclusion	Labs with any sig.	Labs with hyp. sig.
RRR01	5	5	116 [107, 125]	3,840	3,840 (100%)	31	4 (13%)	4 (13%)
RRR02	5	5	88 [84, 98]	3,840	3,840 (100%)	23	10 (43%)	10 (43%)
RRR03 Attribution	6	2	84 [82, 84]	3,840	3,816 (99%)	10	5 (50%)	3 (30%)
RRR03 Intention	6	2	84 [82, 84]	3,840	3,840 (100%)	10	3 (30%)	1 (10%)
RRR03 Process	7	2	84 [83, 85]	7,680	7,680 (100%)	10	2 (20%)	0 (0%)
RRR04	5	3	76 [68, 90]	23,040	20,160 (88%)	24	14 (58%)	7 (29%)
RRR05 Exit	6	9	82 [70, 94]	2,488,320	1,503,904 (60%)	16	12 (75%)	7 (44%)
RRR05 Neglect	6	9	82 [70, 94]	2,488,320	1,540,176 (62%)	16	15 (94%)	10 (62%)
RRR06	7	5	96 [77, 111]	122,880	61,440 (50%)	17	9 (53%)	5 (29%)
RRR07	3	10	74 [63, 95]	2,621,440	525,680 (20%)	21	18 (86%)	10 (48%)
RRR08	8	4	79 [65, 102]	115,200	19,204 (17%)	39	31 (79%)	16 (41%)
RRR09 Behavior	8	4	169 [114, 218]	46,080	46,080 (100%)	26	16 (62%)	5 (19%)
RRR09 Hostility	8	4	168 [114, 218]	46,080	46,080 (100%)	26	13 (50%)	10 (38%)
RRR10	5	5	90 [77, 107]	11,520	3,808 (33%)	25	8 (32%)	1 (4%)

Note. Meta-analytic distributions and estimates after excluding analytic choices that resulted in <24 participants per experimental group at the study level. Three labs in RRR2 (L09, L17, and L26) and one lab in RRR8 (L24) always had fewer than 24 participants and were excluded. "Labs with any sig." = number of labs (%) with any significant (at $p = .05$) effect size in their multiverse; "Labs with hyp. sig." = number of labs (%) with any significant (at $p = .05$) effect size in the hypothesized direction in their multiverse; DFs = degrees of freedom; "Common DFs" = DF from a common list of potential DFs; "Unique DFs" = study-unique DFs; M = median study multiverse size; N [lower quartile, upper quartile] = median study sample sizes across their multiverses. Code to reproduce table available at: http://osf.io/jehpy/.

Overall in Figure 2, statistically significant observations (indicated by observations falling outside the funnel lines) were rare (median = 0.87%, interquartile range = 0%–3%). There were labs with a higher proportion of significant outcomes (L14 = 25%, L04 = 24%, L16 = 17%), but in only one case were these in the hypothesized direction (L04). The median UMV across labs in Figure 2 was 0.1 *SD*, interquartile range (IQR) = 0.09–0.15. Effect sizes could change by as much as $d = 0.97$ (L08). Pearson's correlation based on the 16 labs in Figure 2 between UMV and sample size before applying researcher DFs was $r = -0.51$.

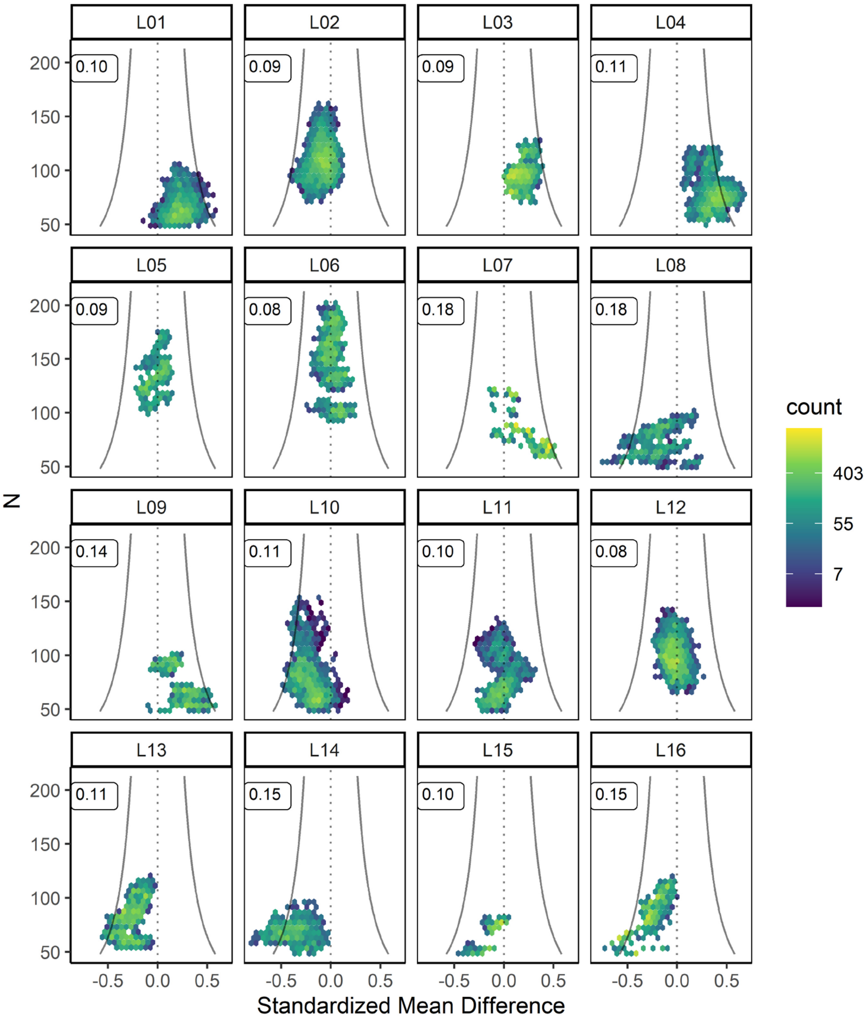
Across RRRs, the median lab UMV was 0.11 *SD* (IQR = .08–0.14) for SMD effect sizes and 0.07 *SD* (IQR = 0.04–0.12) for log odds ratio (OR), but researcher DFs could change effect sizes in a lab by as much as $d = 1.27$ (RRR05 Neglect, L10) and log OR = 1.31 (RRR01, L05). We expected the lab UMV, just as the standard error, to be generally negatively correlated with the (original) sample size. However, the median correlation between lab UMV and sample size (before applying researcher DFs) within RRRs was $r = 0.09$ (IQR = − 0.11 to 0.37). Hence, a large sample size does not ensure a small UMV.

There can be substantial variation between labs also in which researcher DF leads to variability in effect sizes (Figure 3). Figure 3 shows the standard deviation (*SD*) in effect size within the labs from Figure 2 when applying a single researcher DF. Despite identical study designs across labs and the same researcher DF being applied, no two bar plots look identical and labs differ in which researcher DF creates the most variation. For example, in Lab 7 (L07) excluding participants based on different accuracy criteria for the main DV (U1) resulted in the largest *SD*, whereas in Lab 6 (L06) using different criteria for defining and excluding outliers (E6) led to the most variation in estimated effect size. Figures 2 & 3 together demonstrate that multiverse analyses should not necessarily be expected to be replicable in new data, because across labs the same researcher DFs can yield different degrees of effect size variance (Figure 2) and this variance can be primarily caused by different researcher DFs (Figure 3). Note that the effect sizes in Figure 2 arose from all possible combinations of the researcher DFs and not by applying them separately as in Figure 3 (see "Analysis" in the Methods section), which explains the larger range of effect sizes in Figure 2.

Some researcher DFs generally contribute more to the UMV than others and thus constitute larger risk factors when considering selective reporting. Figure 4 shows the *SD*s in estimated effect sizes in labs resulting from applying each researcher DF individually, disaggregated across RRRs, and then aggregated into their respective categories. How the composite score was computed from a scale with multiple items had the largest median effect amongst SMD effects ($d = 0.04$, lower panel top row, Figure 4), and in Supplement F in the online supplemental material, we demonstrate how removing (the source of variation in) this researcher DF (S2.c; computing the composite score based on principal component analysis [PCA]) decreases overall UMV, showing how the removal of sources of variation effectively decreases the risk of selective reporting. Figure 4 also shows that excluding participants on age had a relatively strong effect (median upper panel, log OR = 0.05, lower panel, $d = 0.03$. In Supplement G in the online supplemental material, we show that this effect is driven by the large degree of exclusions across the options of the age researcher DF (E3), by comparing it with a version with more broad inclusion criteria. That is, when

Figure 2

Arbitrary Decisions in Research Cause Underlying Multiverse Variability (UMV) in Effect Size Estimates



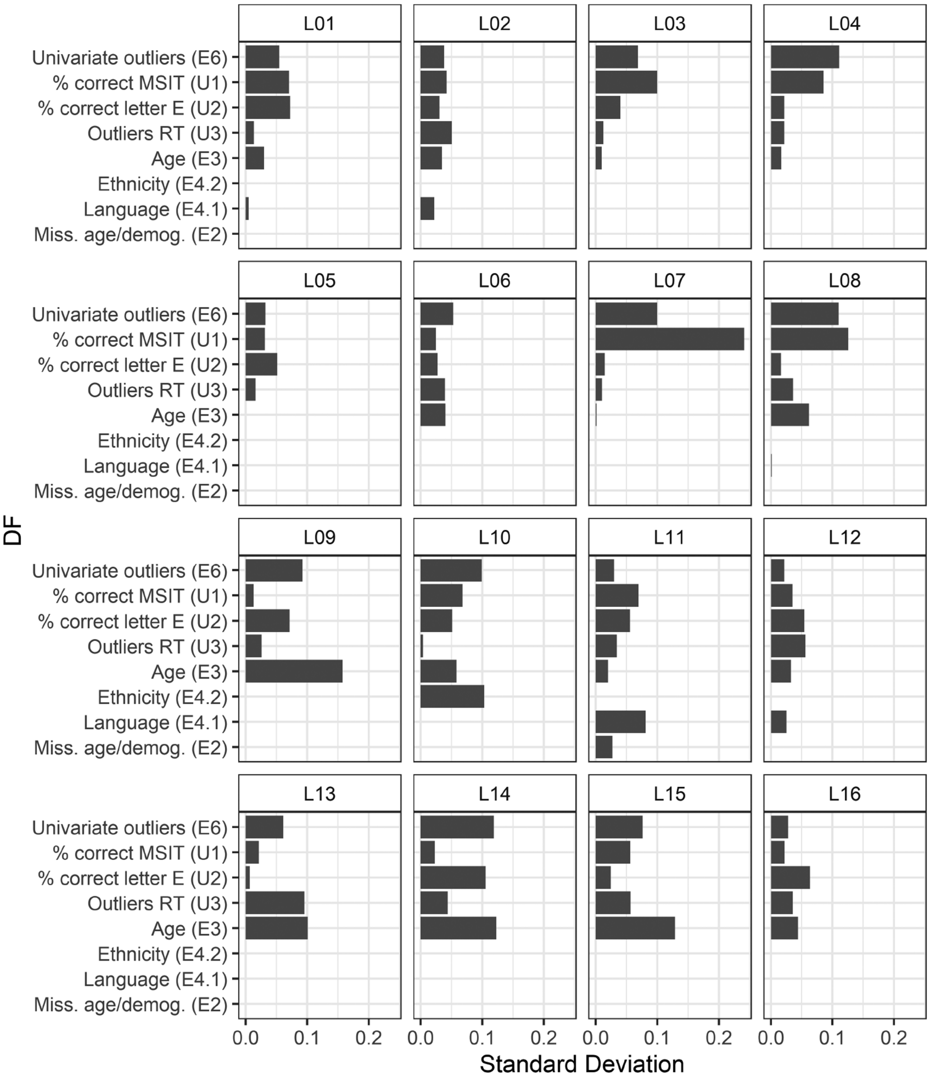
Note. Funnel plots showing the effect sizes based on the multiverses in 16 labs for RRR04, after removing cases where $n < 24$ in either experimental group. Values in the upper left corner of each facet are UMV for each lab. For legibility, 16 out of 24 RRR04 labs are shown; the figure including all labs is available in Supplement E (in the online supplemental material; http://osf.io/2htc6/). L01–L16 are lab indicators. Solid lines are funnel lines based on the *t* distribution. Effect sizes falling outside the funnel lines are statistically significant at $\alpha = .05$ using a two-tailed test. Dotted lines indicate zero effect size. Colors in the funnel plots indicate the frequency of occurrence of an effect size. Brighter colors indicate that an effect size occurred more often. N = total sample size. Code to reproduce figure available at: http://osf.io/thuyk/. See the online article for the color version of this figure.

researcher DFs result in datasets with less overlap (i.e., that are less correlated), the UMV increases. Hence, we can predict that certain researcher DFs are of more serious concern under selective reporting, although the observed effect in any given sample will depend on random fluctuations in the sample.

Unique researcher DFs only had the fifth highest median *SD* (d = 0.02) for SMD effects (lower panel Figure 4), likely due to many unique researcher DFs having little effect. However, they also show the largest range in possible outcomes. For example, within RRR07, choosing to exclude participants based on whether they complied with the set time limit or not (U2) resulted in the largest median effect size *SD* of all researcher DFs for that RRR (*SD* = 0.21, see also Supplement E in the online supplemental material). This same researcher DF in RRR07 resulted in 4 out of the 5 highest effect size *SD*s in Figure 4. The remaining observation (third from the right) belonging to RRR09 Hostility, Lab 1, and resulted from choosing whether to exclude participants based on their study major. Unique researcher DFs show less impact in the log odds ratio effects (upper panel Figure 4), which may be due to fewer labs/researcher DFs, and most unique researcher DFs in RRR01/

Figure 3

The Same Arbitrary Decisions have a Different Effect in Comparable Studies



Note. Standard deviation (*SD*) in effect size estimates in 16 labs in RRR04 resulting from applying different researcher DFs individually, after removing cases where $n < 24$ in either experimental group. For legibility, 16 out of 24 RRR04 labs are shown; the figure including all labs is available in Supplement E (in the online supplemental material; http://osf.io/2htc6/). L01–L16 are lab indicators. Indicators in parentheses on the *Y*-axis (E2, E3, E4_1, E4_2, E6, U1, U2, U3) refer to DF codes for each coded DF (Supplement B in the online supplemental material: http://osf.io/wj38n/). The *Y*-axis is ordered by median *SD* across labs. Code to reproduce figure is available at: http://osf.io/thuyk/.

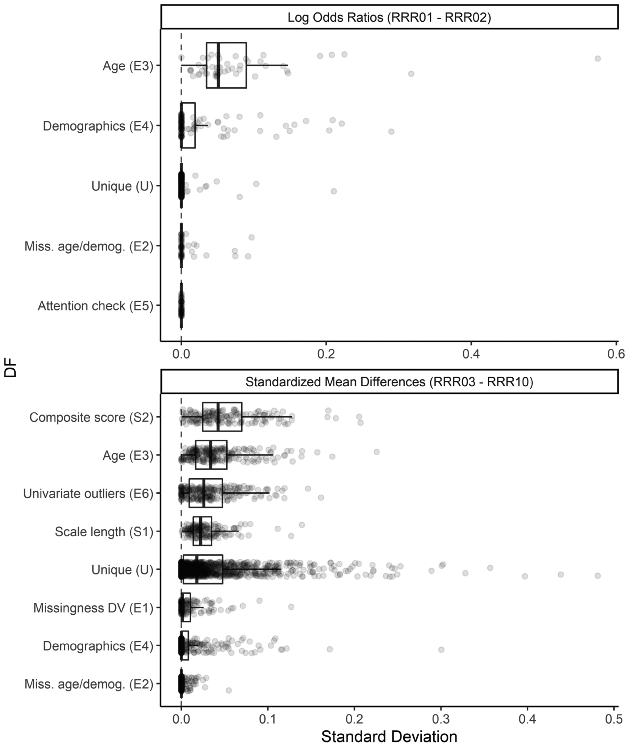
02 only being applicable to a few of the constituent labs (see Supplement B in the online supplemental material). For example, only three labs included a comprehension check (U1), and only three (different) labs coded “familiarity with effect” (U2).

Meta-Analytic Multiverses

Variability in effect sizes within labs due to researcher DFs implies that many different meta-analytic outcomes are possible. How and which effect sizes were selected in labs will change meta-analytic results. Figure 5 shows multiple meta-analytic average effect size estimates for all outcome variables, depending on how effect sizes were selected in the constituent labs. The grey density curves indicate the empirical distributions of meta-analytic point estimates across multiverses for each outcome. UMV in point estimates ranged from 0.02 *SD* (RRR09 Hostility) to 0.04 *SD* (RRR03 Intention) for standardized mean differences and for log odds ratios rounded to 0.02 *SD* for both RRR01 and RRR02. When outcomes were selected through a preregistered decision procedure (purple squares, Figure 5), meta-analytic mean estimates were generally close to the mean of the estimated multiverse distributions and matched the random draw estimates well (pink crosses,

Figure 4

Some Arbitrary Decisions Tend to Create More Effect Size Variability Than Others



Note. SD = standard deviation; DF = degree of freedom; RRR = Registered Replication Report. SD in effect size estimates in labs resulting from different researcher DFs applied individually. The top panel shows results for RRRs with an outcome measured as log odds ratio, and the lower panel for RRRs measured as standardized mean differences. The Y-axis is ordered by median effect size SD. Data are after removing cases where $n < 24$ in either experimental group, disaggregating DFs across RRRs, and aggregating into categories. Indicators in parentheses on the Y-axis (S1, S2, E1, E2, E3, E4, E5, E6) refer to DF codes in Table 2 or (U) to DFs coded as unique for each research project. The Unique (U) category was aggregated across all distinct unique DF. Code to reproduce figure available at: http://osf.io/thuyk/.

Figure 5). The point estimate of the random draws is by definition identical to the mean of the multiverse distribution, whereas the lower/upper bounds are the average 95% CIs of these draws.

When researcher DFs were combined with a biased selection of effect sizes in labs (p-hacking), meta-analytic mean estimates were also more extreme in the predicted direction (Figure 5). As expected, selecting the most significant effect size in each lab (yellow stars) and then meta-analyzing resulted in the most extreme mean estimates. Other p-hacking approaches (in Figure 5: turquoise triangles, red circles, and green squares) resulted in similar estimates. This similarity in outcome between biased selection mechanisms can be mostly attributed to the low number of significant results across labs and our biased selection procedures resulting in the same results if there were no significant outcomes in a lab (205 out of 294 studies across all RRRs, counting labs with multiple DVs as separate studies).

There is a tendency for projects with larger multiverses (e.g. RRR05, RRR06, and RRR07, as can be seen in Figure 5) to have more extreme estimated effect sizes when p-hacked. The difference between the average random draw (pink crosses) and the estimates based on the most significant effect sizes (yellow stars) ranged from 0.1 to 0.48 for SMDs and was about 0.1 for log OR. The correlation between effect size inflation and multiverse size was $r = 0.77$ for SMD effect sizes. The most extreme case corresponded to RRR07 (Figure 5), where the difference in meta-analytic average effect size estimate between the average random draw (pink cross, $d = -0.03$, 95% CI [-0.13, 0.07]) and the estimate based on the most significant effect sizes (yellow star, $d = 0.45$, 95% CI [0.34, 0.56]) was an increase of almost 0.5SD in the predicted direction. Applying publication bias correction methods (PET-PEESE, 3PSM and p-uniform*) did not lead to improvements in estimated average effect size estimates (Supplement H in the online supplemental material), in line with other research that has shown publication bias correction methods as unlikely to be useful in correcting for selective reporting (Carter et al., 2019; van Aert et al., 2016).

Discussion

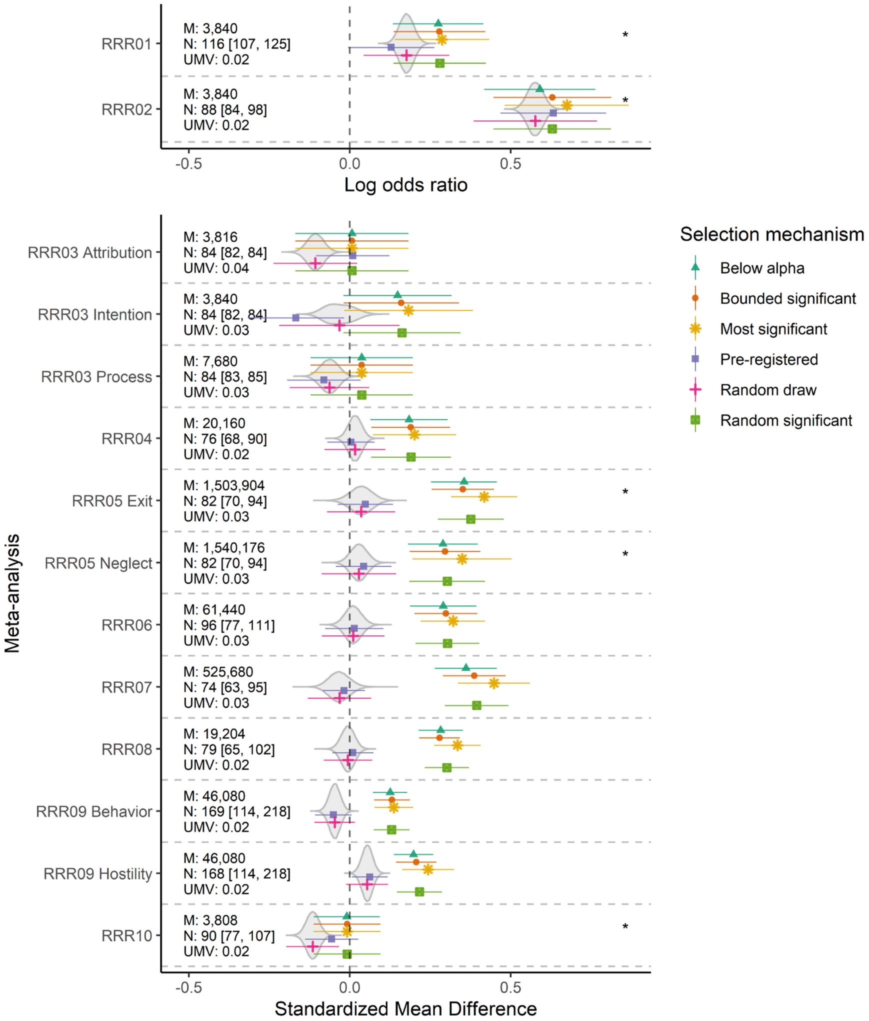
In this article, we performed multiverse analysis across multiple direct replication studies, using empirical data from 10 RRRs. Even though the preregistrations used in the actual RRRs limited the effect of researcher DFs in the original analyses, the open data from these extensive studies enabled us to use a counter-factual design to see what biases could have emerged if the studies (and ensuing meta-analyses that included them) had not been preregistered and could have been subjected to selective reporting based on the significance of the outcomes in the primary studies (p-hacking). We identified researcher DFs based on common decisions in the associated literature for each outcome variable, computed all possible outcomes across direct replications, and examined the variance in these so-called multiverses. We then combined effect sizes from the multiverses of each direct replication in meta-analysis and examined the consequences of different mechanisms for selecting effect sizes for inclusion. Our analyses highlight that multiverse analyses typically yielded thousands of different outcomes within single studies, that multiverse patterns of variation differed across labs using the same protocol, and that selective reporting of outcomes in primary studies could bias meta-analytic results, despite their status as a “gold-standard” of evidence. We also found that the original sample size correlates at most weakly with the potential for selective reporting (as measured by UMV), suggesting a larger sample size does not protect against selective reporting. Yet interestingly, 205 out of the 294 studies (counting labs with multiple DVs as separate studies) did not contain any significant (measured as $p \leq .05$) results in their multiverses, suggesting p-hacking null results into significance may be more difficult than expected considering the sheer number of potential outcomes per study. We discuss these results and the limitations of our own study in the remainder of the discussion.

Defining the Multiverse

Creating a multiverse is an inherently subjective endeavor given that researchers might disagree about which decisions are (approximately) equally reasonable (Steegen et al., 2016). For example,

Figure 5

Selective Reporting in Labs Results in Overestimates in Meta-Analysis



Note. Meta-analytic distributions and estimates after excluding analytic decisions that resulted in $n <$ 24 participants per experimental group at the study level. Selection mechanism = how effect sizes were selected at the study level, either by p-hacking ("Most significant," "Bounded significant," "Random significant," "Below α,"), preregistered decisions ("preregistered"), or random selection ("Random draw"). "Random draw" and "Random significant" are averages across 10^5 draws from the meta-analytic multiverse, whereas other selection mechanisms are a single outcome. M = median study multiverse size, N [lower quartile, upper quartile] = median study sample sizes across their multiverses, * = effect size sign changed (RRR01, RRR02, RRR05 Exit, RRR05 Neglect, RRR10) so that hypothesized effect size (and p-hacking direction) was positive for all meta-analyses. Code to reproduce the figure is available at http://osf.io/thuyk/. See the online article for the color version of this figure.

although we created our researcher DFs based on common practice in the associated literature, there are methodological arguments to carefully consider the meaning and impact of outliers, or use outlier robust statistics (e.g., Rousseeuw & Hubert, 2011) rather than excluding them based on rules of thumb. In the same vein, it may often be preferable to perform multiple or maximum likelihood imputation (e.g., Jakobsen et al., 2017) of missing data points rather than excluding them. For this reason, we have endeavored to structure our data such that a disagreeing reader familiar with R can explore the consequence of only including those of our researcher DFs they consider reasonable, as demonstrated in Supplement F in the online supplemental material.

Del Giudice & Gangestad (2021) critically discussed what it means for a researcher DF to be "reasonable." They rightly argued why some of the researcher DFs as used in previously published multiverse analyses might not be equivalent on prior grounds or might not show equivalence for reasons yet unknown. We admit that our researcher DF S1, which deletes items in measurement scales based on the lowest item-rest correlations, could perhaps be nonequivalent on psychometric grounds. On the one hand, deletion of the item with the lowest item-rest correlation could heighten the reliability of the scale, and hence increase genuine effects, if the deleted item were poorly performing. On the other hand, deletion would have little effect on the reliability and hence on actual effects if the item performed as well as other items in the scale. Hence, if there were some item(s) consistently performing poorly (in the psychometric sense) across labs in an RRR, this researcher DF could increase true effects and be of principled nonequivalence, at least for that RRR. Similarly, the researcher DF related to composite scores (S2) could create genuinely different effects under some conditions. In Supplement F in the online supplemental material, we repeated our analyses without the PCA option of this researcher DF, which showed less variation in outcomes, as expected if indeed the way composite scores are computed makes a difference. However, like all our DFs, we included this researcher's DF because we feel that it might be used in practice in unregistered studies.

One could also argue that multiverses with varying sample sizes are nonequivalent in the statistical sense of sampling variability (Del Giudice & Gangestad, 2021). Smaller sample sizes generally result in less statistical power to detect effects and hence larger *p*-values, which makes it more difficult to directly compare *p*-values across multiverses differing in sample size. We do not consider the issue of power to be a major concern for our study, as many of our analyses focus on effect size estimates and variability, and because most RRRs in our sample (apart from RRR1/RRR2, based on our reading of the original articles) appear to study null effects. In addition, although researcher DFs that lead to smaller sample sizes would increase the variability of outcomes regardless of true effect size, the strong sample overlap between multiverse samples in a given study creates covariation between multiverses that diminishes total sampling variability. In the end, which multiverses are considered reasonable will depend not only on individual researchers' beliefs, but also on which decisions their research community considers acceptable in terms of theoretical, methodological, and empirical standards (Del Giudice & Gangestad, 2021).

Notwithstanding selective reporting, researchers should recognize that researcher DFs, or decisions treated as if they were researcher DFs, by themselves create another layer of uncertainty in study estimates; we found a median UMV of 0.1*SD* in our SMD data, although this will differ depending on the research field and which DFs researchers find reasonable. As such, we advise researchers doing hypothesis-testing research to (a) preregister the (single) analysis they believe is optimal for testing their hypothesis, motivate why this is the case and report uncertainty estimates (e.g., confidence or credibility interval), and (b) include a multiverse analysis as a sensitivity analysis (following the advice of Del Giudice & Gangestad, 2021) and report their UMV. Although some may argue that preregistering a preferred analysis is contradictory if the options in the multiverse analysis are considered equivalent, pragmatically, we believe that most researcher DFs will not be exactly equivalent and that most researchers will have a preferred analysis that it would be useful to accompany with a sensitivity analysis. The goal of these recommendations is for the research process to be transparent so that results act as credible evidence despite the potential effect of researcher DFs on outcomes.

More generally, multilab collaboration, or regular, projects may wish to consider incorporating the multiverse perspective already in the design of their studies, identifying which of their decisions are largely arbitrary and collecting data on alternatives. Preregistration of research is likely to be helpful from this perspective, in addition to its transparency-enhancing properties, which are helpful when evaluating a study for selective reporting.

Exploring the Multiverse

The unique design of our project enabled us to examine the effect of researcher DFs (i.e., perform multiverse analysis) across multiple direct replication studies. We observed that (a) the same researcher DFs applied to direct replication studies resulted in widely varying distributions of effect sizes and (b) which researcher DF caused the variability within a study differed between direct replications. That is, the effect of researcher DFs both within and across direct replication studies appeared unsystematic. These results demonstrate that results of multiverse analysis in any single study, like other exploratory analyses, are not necessarily replicable in new data. We believe this point is underappreciated among many multiverse analysts. In addition, some researchers may be tempted to directly interpret the existence of researcher DFs and resulting UMV as evidence of "hidden moderators" (Van Bavel et al., 2016); currently unknown moderators that explain why effect sizes differ between studies. However, the existence of the multiverse does not by itself imply moderators as substantial variability and apparent moderating effects may be found through sampling variance alone.

That the effect of researcher DFs both within and across direct replication studies was generally nonsystematic also corroborates previous findings of ours (Olsson-Collentine et al., 2020) that differences in study results in social and cognitive psychology show little to no between-study heterogeneity, and supports the conclusion that the best explanation for differences between effect sizes in (direct) replication studies is typically the joint effect of sampling error and researcher DFs, possibly in combination with selective reporting.

When we have a substantive researcher DF that we suspect of being a moderator, it may be most useful to examine it from an empirical meta-analytic perspective. If we have a researcher DF at the study level (e.g., measurement scale) with sufficient variation between primary studies, it is possible to examine it as a moderator using meta-regression (e.g., Houwelingen et al., 2002. However, individual-level researcher DFs (e.g., age) are preferably examined in individual participant data (IPD) meta-analysis to avoid the ecological fallacy (e.g., Stewart & Tierney, 2002. In the case of multiple dependent variables, which might also be a researcher's DF, potential systematic differences could be examined in a multivariate meta-analysis (e.g., Jackson et al., 2011). As with multiverse analysis, such moderator analyses should primarily be considered exploratory and hypothesis-generating.

Researcher DFs in primary studies also add a layer of uncertainty to meta-analysis when those studies are meta-analyzed. Researcher DFs in primary studies can change both point estimates and the associated standard errors and can do so across multiple studies. Consequently, in a meta-analysis, they can influence not only the meta-analytic point estimate but also the between-study variance. That said, the standard deviations in point estimates over the meta-analytic multiverses in our meta-analyses were quite small, with an UMV of at most 0.04 *SD* amongst SMD effects. This is unsurprising: when researchers' decisions are truly random and in the absence of selective reporting and publication bias, researcher DFs in primary studies can be expected to cancel out across a sufficiently large number of studies. As such, researcher DFs in primary studies (and resulting multiverses) are not a major concern for meta-analysts in the absence of selective reporting. Unfortunately, the availability of such ideal data is not expected in most meta-analyses. Even with ideal data, in a meta-analysis with a small number of primary studies, as is common in medicine (Davey et al., 2011), researcher DFs are less likely to balance each other out and meta-analytic UMV may be a larger concern.

Encouragingly, in our data, the meta-analytic point estimates based on preregistered studies often fell close to the center of the meta-analytic distributions. Preregistration may generally lead to less effect size inflation (Schäfer & Schwarz, 2019) by decreasing the risk of selective reporting through increased transparency (although preregistrations are of varying quality or not always suitable, Bakker et al., 2020; Pham & Oh, 2021). However, the connection between multiverse analysis and preregistration may have been enhanced by the nature of our data: large collaborative projects, including researchers with adversary hypotheses, which may have resulted in a "wisdom of the crowd" selection of decisions amongst researchers DFs. Alternatively, if no decision within a multiverse has a systematic effect, results from any preregistered set of decisions from that multiverse would also be expected to coincide approximately with the mean of the multiverse distribution when analyzed across samples. Regardless, to account for UMV preregistered multilab collaborations (e.g., Moshontz et al., 2018) may offer a way forward in the absence of more concrete theory (Fried, 2020), although as we saw in our results even such data is not a guarantee for a point estimate at the center of the meta-analytic multiverse distribution. It is important to be aware that a preregistered set of decisions nonetheless only represents a single universe from the multiverse.

The extent of UMV in any given field depends on the multiverse created, and our estimates in this study may only apply to our nonrandom sample of social and cognitive psychology research. UMV in other fields could be either larger or smaller, but is unlikely to be nonexistent, and it may be worthwhile to study the UMV in different subfields to examine their susceptibility to selective reporting given normative research behavior. Generally, the extent of bias introduced by selective reporting will depend on the multiverse size and the UMV, and researcher DFs that affect these two factors to a larger extent will hence contribute more risk of bias to a study.

Selecting From the Multiverse

We do not mean to imply that exploration of researcher DFs is problematic per se. We view it as important to study the robustness of conclusions in the context of a sensitivity analysis, of which a multiverse analysis can be seen as an extensive (systematic) variant. Exploring factors that truly moderate an effect or association can be valuable as long as the exploration is transparently reported and employs rigorous statistical controls to guard against overfitting. What is problematic is the selective or incomplete reporting from the multiverse of statistical results. Hence, it is important to evaluate studies for risk of selective reporting when using them to make decisions (e.g., about setting up future research), or including them in systematic reviews (Appelbaum et al., 2018; as recommended by e.g., PRISMA and MARS: Moher et al., 2009). Both preregistrations and multiverse analyses will facilitate evaluation of a study's selective reporting risk by making research decisions more transparent, and there are many selective reporting protocols available (e.g., Page et al., 2018) that may assist in the evaluation.

There is a risk that researchers exploit (intentionally or not) researcher DFs to selectively report those results from the multiverse that most strongly support their hypothesis. In extreme cases, researcher DFs and *p*-hacking can provide evidence for any desired conclusion; in one lab, the effect size estimate changed by as much as $d = 1.27$. More realistically, we found a median UMV of 0.1 *SD* amongst 294 studies (counting labs with multiple DVs as separate studies). Nonetheless, a median UMV of 0.1 *SD* in a field still implies that selective reporting can turn a statistically nonsignificant effect into a significant effect.

For many studies in our data, this was not a concern. We found that in our data and given our researcher DFs, about 70% of study multiverses did not contain a single significant result (measured as $p < .05$) in the hypothesized direction. As most RRRs that made up our data had overall effect size estimates (based on preregistered outcomes) not significantly different from zero, this coincides with previous findings of ours that on average null results also tend to have very little heterogeneity (Olsson-Collentine et al., 2020). This suggests that it may be more difficult to *p*-hack null results into significance than many expect. We caution that this observation may no longer hold when applying other types of researcher DFs than we were able to do (e.g., this may not apply when researcher DF options are less correlated, as in the case of outcome switching), and that 30% of labs did contain multiverses with a mix of significant and nonsignificant effect size estimates.

Relatedly, we found a correlation not significantly different from zero between sample size and potential for selective reporting (as measured by UMV). This implies that sample size should not be taken to be protective against selective reporting, as also corroborated by a simulation study by Stefan & Schnbrodt (2022). We do note, however, that for genuine effects, larger sample sizes would increase power thereby lowering the need to selectively report outcomes based on the multiverse and hence less ensuing bias in estimated effects (?).

Our counter-factual design allowed us to see what biases could have emerged if the studies (and ensuing meta-analyses) had not been preregistered and could have been subjected to selective reporting based on the significance of the outcomes in the primary studies, as is the case for most meta-analyses. Our analyses demonstrate the substantial bias in the hypothesized direction that may be incorporated into meta-analytic effect size estimates due to selective reporting in primary studies. The possible inflation of average effect size will depend on the proportion of meta-analyzed studies at risk of bias and the strength of this bias. Evidence from Kvarven et al. (2019) based on social and cognitive psychology research suggests meta-analyses may sometimes estimate effect sizes to be as much as a third larger than in comparable multilab projects, although this also includes publication bias. Larger or smaller differences may be more typical in other fields.

For meta-analysts using retrospective data, minimizing the risk of bias in their included data (i.e., by only including preregistered data and evaluating it for selective reporting) may be the best option until the practices of multiverse analysis or the sharing of raw data become widespread. Our results corroborate recommendations from meta-analysis reporting protocols such as PRISMA and MARS (Appelbaum et al., 2018; Moher et al., 2009) to always evaluate primary studies for risk of bias (here, selective reporting), and we advise meta-analysts to study differences in outcome between studies identified as at high risk of selective reporting bias and those at low risk. In line with previous research on the topic (Carter et al., 2019; van Aert et al., 2016), our supplemental results show that existing publication bias methods should not be relied on to correct for *p*-hacking.

The large variance in the impact of researcher DFs across direct replication studies demonstrates that the study-specific effect of a researcher DF, and related bias induced by selective reporting, is difficult to predict and dependent on sampling error. Stefan & Schnbrodt (2022), who simulated the effects of many different *p*-hacking methods in single studies, reach a similar conclusion in their simulations: "Apart from the aggressiveness of *p*-hacking itself, our simulations showed that across all strategies, the severity of *p*-hacking also depends on the environment in which *p*-hacking takes place, for example, the correlation structure in the data" (p. 46). Our results using the RRR data indeed show that the correlations between multiverses create a variation that is generally smaller than the sampling variation one would expect under independent sampling.

That is not to say that we cannot draw some conclusions about the expected (average) impact of *p*-hacking different researcher DFs. The potential for effect size bias is larger when studies allow more analyses to be run and when more variance is created by included researcher DFs. Supplement G in the online supplemental material demonstrates that more overlapping (sub)samples created using alternative exclusions based on age created less subsequent UMV, which is expected given that more overlap creates higher correlations between alternative outcomes. As such, when considering effect size bias, we should (typically) be more concerned about researcher DFs in which the options are less correlated, although high false positive rates are possible in either case (see discussion by Friese & Frankenbach, 2020).

It may be insightful to do more complex modeling of selective reporting from a multiverse perspective, including nonintentional selective reporting, and we hope our data will also be useful to other researchers interested in more complex modeling of research bias. Our modeling of it in this study was relatively straightforward and we only attempted to model the outcomes of intentional selective reporting (*p*-hacking). Nonetheless, our biased selection methods applied to empirical RRR data are similar to those used in the simulations of a recent compendium of *p*-hacking methods Stefan & Schnbrodt (2022) and are on par with other recent simulation studies of *p*-hacking in a meta-analysis (Botella et al., 2021; Friese & Frankenbach, 2020).

Contrary to these two simulation studies (Botella et al., 2021; Friese & Frankenbach, 2020), we found that *p*-hacking with actual data and using fairly generic researcher DFs could cause substantial inflation of meta-analytic average effect sizes also when the average effect appears to be null. Both Friese & Frankenbach (2020) and Botella et al. (2021) run extensive simulation studies of the effect of *p*-hacking across many conditions and Friese & Frankenbach (2020) consider how it interacts with publication bias, something we did not do. We believe the difference in results is due to the choice in both of these simulation studies to *p*-hack results based on the common assumption that *p*-hacking leads to a peak of *p*-values below 0.05 (Hartgerink, 2017). In the case of Friese & Frankenbach (2020), results were *p*-hacked to a distribution with a mode of $p = .049$, and Botella et al. (2021), similarly *p*-hacked studies into the region $.025 < p \leq .05$. We know from previous studies that the type of *p*-hacking matters; incremental methods such as optional stopping that result in a peak below $p = .05$ have little effect on effect size inflation, whereas methods such as outcome reporting bias have a large effect and do not result in a peak (Francis, 2012; Kirkham et al., 2010; Stefan & Schnbrodt, 2022). As such, differences between our results and these simulation studies are likely explained by the incremental *p*-hacking methods used in these simulation studies as compared to our methods of selecting the lowest *p*-values or randomly selecting one of the significant outcomes.

Unfortunately, there exists little evidence on which method of selection researchers use in practice. Incremental *p*-hacking can still lead to concerning numbers of false positive results, as discussed by both Friese & Frankenbach (2020) and Botella et al. (2021), and it is important to discuss that not all types of *p*-hacking lead to concerning levels of effect size inflation. Nonetheless, our results show that suggesting selective reporting is not a concern for meta-analytic results is inaccurate when considering nonincremental *p*-hacking based on researcher DFs that we consider to be widely applicable across a range of empirical studies.

Under some assumptions related to the correlational structure of the overlapping data across multiverses, we can be confident that the UMV and hence the potential bias due to selective reporting in a study is less than the reported standard error. The effect sizes in a multiverse are dependent because they are based on the same sample. Due to this dependence, the UMV will normally be smaller than the standard error in a study for a fixed sample size and statistical model, since the variability based on independent data is larger than that of dependent data. In other words, if we know that the statistical model and sample size have not changed in a study and that there is no publication bias, then we can be confident that the UMV in that study is less than its standard error estimate.

Consequently, at a fixed standard error, the possible bias is always larger with publication bias than with selective reporting due to the dependency between effect sizes in the multiverse. This suggests that while *p*-hacking is likely more common than publication bias in the literature, being more resource efficient, the distortion in the literature may be larger from publication bias when it does occur. Finally, we note that the correlation between effect sizes within a multiverse also means that the independent sample false positive rate (typically .05) should be expected to be lower when sampling effect sizes within a dataset. We can observe this in our funnel plots, where substantially fewer than 1 out of the 20 effect sizes are significant for most labs (i.e., fall outside the funnel lines).

Limitations and Constraints on Generality

Although we have attempted to accompany all our claims and findings in this article together with their caveats, we wish to make explicit the limits of generalizability of claims based on the data and design of our study. The included effects are neither a representative nor a random sample of effects from psychology. We expect our conclusions to be robust for effects in social and cognitive psychology, but specific values that we report (e.g., median UMV of 0.1 among SMD effect sizes) may not generalize beyond our sample. The RRRs in our data overwhelmingly reported average results not significantly different from zero (the exception being RRR1/RRR2). This allowed us to examine selective reporting in its most critical context (i.e., in the likely absence of genuine effects), but means it would be good to focus future research efforts on studying multiverses with nonnull effects, as these likely create more heterogeneity in results across labs (see Olsson-Collentine et al., 2020), and hence we would also expect larger overall variability and UMV due to researcher DFs (see also Friese & Frankenbach, 2020). During the process of this project, new RRRs have been published some of which report nonnull effects and could be used for such further analyses (e.g., Elliott et al., 2021). In addition, we are aware of several projects currently in progress with similar designs that will collect new data and provide additional evidence on the impact of standard researcher DFs in different fields.

Researchers interested in the theoretical implications of the specific effects studied here should carefully consider which researcher DFs they find reasonable (see Del Giudice & Gangestad, 2021) before drawing conclusions. Our researcher DFs were chosen to match standard decisions in social and cognitive psychology and treated as researcher DFs, meaning they were not guided by the substantive theory of the studied effects, except as reflected by decisions made by the RRRs in data collection and analysis. Moreover, if for instance age is theoretically expected to moderate an effect it should preferably be tested formally instead of being used in a multiverse analysis. van Aert (2022) demonstrated how this can be done by looking at the effect of age across labs in RRR9 (McCarthy et al., 2018) using IPD meta-analysis, finding a small positive interaction ($p = .038$). Individual studies often do not have the power to detect moderating effects, which also affects multiverse analyses.

We found that in only about 30% of studies did significant effect sizes in the hypothesized direction emerge in their multiverses, despite the typically thousands of analyses in every study. This finding suggests it is more difficult to turn apparent null results into significant results than might be expected but is dependent on our selection of researcher DFs. Although we implemented an extensive number of researcher DFs that we consider representative of researcher DFs that could be used in practice across a range of social and cognitive psychological studies, our use of secondary (real) data means there were many researcher DFs (e.g., Wicherts et al., 2016) that we could not apply but that might be applied in real situations (e.g., outcome switching, which is known to have a large impact, or changing the analytic model). As such, our selection of DFs and the resulting multiverse variances are unlikely to represent a worst-case scenario. It is feasible that in real life more extreme statistical results are found. Generally, we can expect researcher DFs with a lower correlation between options, because of less sample overlap and/or weaker correlations between (in)dependent variables, to result in larger multiverse variance.

Finally, it may be informative to analyze other multilab replications studies than those we included in our study such as Many Labs 1 to 5 (e.g., Klein et al., 2018). The studies in our sample (mostly) studied a single effect across multiple labs, whereas the Many Labs projects study many effects at the same time across multiple labs. We examined the RRRs to be able to apply study-unique researcher DFs, but the Many Labs design would allow examining the impact of applying a single set of researcher DFs on a large sample of effects from social and cognitive psychology.

Conclusion

We have shown that researcher DFs offer a wide array of potential outcomes in relatively standard psychological studies and demonstrated how selective reporting based on these researcher DFs creates a bias in meta-analytic effect size estimates that may undermine the credibility of many meta-analyses. Preregistration is a methodological solution to researcher DFs enabling selective reporting, whereas a statistical solution is to perform multiverse analysis of results. These two transparency-enhancing practices can also be applied together, although our analyses of multiverses across direct replications highlight that multiverse analyses in single studies should not necessarily be expected to replicate in new data. Due to dependencies between effect sizes within multiverses, exploring multivariate approaches to multiverse analysis may be a useful next step in helping to address uncertainties and biases in primary studies due to researcher DFs.

References

Agnoli, F., Wicherts, J. M., Veldkamp, C. L. S., Albiero, P., & Cubelli, R. (2017). Questionable research practices among Italian research psychologists. *PLoS ONE*, *12*(3), Article e0172792. https://doi.org/10.1371/journal.pone.0172792

Alogna, V. K., Attaya, M. K., Aucoin, P., Bouwmeester, S., Brandimonte, M. A., Brown, C., Buswell, K., Carlson, C., Carlson, M., Chu, S., Cislak, A., Colarusso, M., Colloff, M. F., Dellapaolera, K. S., Delvenne, J.-F., Di Domenico, A., Drummond, A., Echterhoff, G., Edlund, J. E., … Zwaan, R. A. (2014). Registered replication report: Schooler and Engstler-Schooler (1990). *Perspectives on Psychological Science*, *9*(5), 556–578. https://doi.org/10.1177/1745691614545653

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA publications and communications board task force report. *The American Psychologist*, *73*(1), 3–25. https://doi.org/10.1037/amp0000191

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554. https://doi.org/10.1177/1745691612459060

Bakker, M., Veldkamp, C. L. S., van Assen, M. A. L. M., Crompvoets, E. A. V., Ong, H. H., Nosek, B. A., Soderberg, C. K., Mellor, D., & Wicherts, J. M. (2020). Ensuring the quality and specificity of preregistrations. *PLoS Biology*, *18*(12), Article e3000937. https://doi.org/10.1371/journal.pbio.3000937

Bakker, M., & Wicherts, J. M. (2014). Outlier removal, sum scores, and the inflation of the type I error rate in independent samples *t* tests: The power of alternatives and recommendations. *Psychological Methods*, *19*(3), 409–427. https://doi.org/10.1037/met0000014

Borenstein, M. (2009). Effect sizes for continuous data. In H. Cooper, L. Hedges, & J. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 221–235). Russel Sage Foundation.

Botella, J., Suero, M., Durán, J. I., & Blazquez, D. (2021). The small impact of *p*-hacking marginally significant results on the meta-analytic estimation of effect size. *Anales De Psicolog'ıa*, *37*(1), 178–187. https://doi.org/10.6018/analesps.433051

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

Bouwmeester, S., Verkoeijen, P. P. J. L., Aczel, B., Barbosa, F., Bègue, L., Brañas-Garza, P., Chmura, T. G. H., Cornelissen, G., Døssing, F. S., Espín, A. M., Evans, A. M., Ferreira-Santos, F., Fiedler, S., Flegr, J., Ghaffari, M., Glöckner, A., Goeschl, T., Guo, L., Hauser, O. P., … Wollbrant, C. E. (2017). Registered replication report: Rand, Greene, and Nowak (2012). *Perspectives on Psychological Science*, *12*(3), 527–542. https://doi.org/10.1177/1745691617693624

Breznau, N., Rinke, E. M., Wuttke, A., Adem, M., Adriaans, J., Alvarez-Benjumea, A., Andersen, H. K., Auer, D., Azevedo, F., Bahnsen, O., Balzer, D., Bauer, G., Bauer, P. C., Baumann, M., Baute, S., Benoit, V., Bernauer, J., Berning, C., Berthold, A., … Nguyen, H. H. V. (2021). Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *MetaArXiv*. https://doi.org/10.31222/osf.io/cd5j9

Burton, A., & Altman, D. G. (2004). Missing covariate data within cancer prognostic studies: A review of current reporting and proposed guidelines. *British Journal of Cancer*, *91*(1), 4–8. https://doi.org/10.1038/sj.bjc.6601907

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, *2*(2), 115–144. https://doi.org/10.1177/2515245919847196

Cheung, I., Campbell, L., LeBel, E. P., Ackerman, R. A., Aykutoğlu, B., Bahník, Š, Bowen, J. D., Bredow, C. A., Bromberg, C., Caprariello, P. A., Carcedo, R. J., Carson, K. J., Cobb, R. J., Collins, N. L., Corretti, C. A., DiDonato, T. E., Ellithorpe, C., Fernández-Rouco, N., Fuglestad, P. T., … Yong, J. C. (2016). Registered replication report: Study 1 from Finkel, Rusbult, Kumashiro, & Hannon (2002). *Perspectives on Psychological Science*, *11*(5), 750–764. https://doi.org/10.1177/1745691616664694

Coburn, K. M., & Vevea, J. L. (2019). *Weightr: Estimating weight-function models for publication bias* [Manual]. https://CRAN.R-project.org/package=weightr

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4–19. https://doi.org/10.1016/j.jesp.2015.07.006

Davey, J., Turner, R. M., Clarke, M. J., & Higgins, J. P. (2011). Characteristics of meta-analyses and their component studies in the Cochrane database of systematic reviews: A cross-sectional, descriptive analysis. *BMC Medical Research Methodology*, *11*(1), Article 160. https://doi.org/10.1186/1471-2288-11-160

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), 22–26. https://doi.org/10.1177/2515245920954925

Eerland, A., Sherrill, A. M., Magliano, J. P., Zwaan, R. A., Arnal, J. D., Aucoin, P., Berger, S. A., Birt, A. R., Capezza, N., Carlucci, M., Crocker, C., Ferretti, T. R., Kibbe, M. R., Knepp, M. M., Kurby, C. A., Melcher, J. M., Michael, S. W., Poirier, C., & Prenoveau, J. M. (2016). Registered replication report: Hart & Albarracín (2011). *Perspectives on Psychological Science*, *11*(1), 158–171. https://doi.org/10.1177/1745691615605826

Elliott, E. M., Morey, C. C., AuBuchon, A. M., Cowan, N., Jarrold, C., Adams, E. J., Attwood, M., Bayram, B., Beeler-Duden, S., Blakstvedt, T. Y., Büttner, G., Castelain, T., Cave, S., Crepaldi, D., Fredriksen, E., Glass, B. A., Graves, A. J., Guitard, D., Hoehl, S., … Voracek, M. (2021). Multilab direct replication of Flavell, Beach, and Chinsky (1966): Spontaneous verbal rehearsal in a memory task as a function of age. *Advances in Methods and Practices in Psychological Science*, *4*(2), Article 25152459211018187. https://doi.org/10.1177/25152459211018187

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Francis, G. (2012). The psychology of replication and replication in psychology. *Perspectives on Psychological Science*, *7*(6), 585–594. https://doi.org/10.1177/1745691612459520

Franco, A., Malhotra, N., & Simonovits, G. (2016). Underreporting in psychology experiments: Evidence from a study registry. *Social Psychological and Personality Science*, *7*(1), 8–12. https://doi.org/10.1177/1948550615598377

Fried, E. I. (2020). Lack of theory building and testing impedes progress in the factor and network literature. *Psychological Inquiry*, *31*(4), 271–288. https://doi.org/10.1080/1047840X.2020.1853461

Friese, M., & Frankenbach, J. (2020). *P*-hacking and publication bias interact to distort meta-analytic effect size estimates. *Psychological Methods*, *25*(4), 456–471. https://doi.org/10.1037/met0000246

Fusar-Poli, P., Radua, J., Frascarelli, M., Mechelli, A., Borgwardt, S., Fabio, F. D., Biondi, M., Ioannidis, J. P. A., & David, S. P. (2014). Evidence of reporting biases in Voxel-based morphometry (VBM) studies of psychiatric and neurological disorders. *Human Brain Mapping*, *35*(7), 3052–3065. https://doi.org/10.1002/hbm.22384

Goldacre, B., Drysdale, H., Dale, A., Milosevic, I., Slade, E., Hartley, P., Marston, C., Powell-Smith, A., Heneghan, C., & Mahtani, K. R. (2019). Compare: A prospective cohort study correcting and monitoring 58 misreported trials in real time. *Trials*, *20*(1), Article 118. https://doi.org/10.1186/s13063-019-3173-2

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., Brand, R., Brandt, M. J., Brewer, G., Bruyneel, S., Calvillo, D. P., Campbell, W. K., Cannon, P. R., Carlucci, M., Carruth, N. P., Cheung, T., Crowell, A., De Ridder, D. T. D., Dewitte, S., … Zwienenberg, M. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, *11*(4), 546–573. https://doi.org/10.1177/1745691616652873

Hartgerink, C. H. J. (2017). Reanalyzing Head et al. (2015): Investigating the robustness of widespread *p*-hacking. *PeerJ*, *5*, Article e3068. https://doi.org/10.7717/peerj.3068

Henry, L., & Wickham, H. (2020). *Purrr: Functional programming tools* [Manual]. https://CRAN.R-project.org/package=purrr

Houwelingen, H. C. van, Arends, L. R., & Stijnen, T. (2002). Advanced methods in meta-analysis: Multivariate approach and meta-regression. *Statistics in Medicine*, *21*(4), 589–624. https://doi.org/10.1002/(ISSN)1097-0258

Huntington-Klein, N., Arenas, A., Beam, E., Bertoni, M., Bloem, J. R., Burli, P., Chen, N., Grieco, P., Ekpe, G., Pugatch, T., Saavedra, M., & Stopnitzky, Y. (2021). The influence of hidden researcher decisions in applied microeconomics. *Economic Inquiry*, *59*(3), 944–960. https://doi.org/10.1111/ecin.v59.3

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

Jackson, D., Riley, R., & White, I. R. (2011). Multivariate meta-analysis: Potential and promise. *Statistics in Medicine*, *30*(20), 2481–2498. https://doi.org/10.1002/sim.v30.20

Jakobsen, J. C., Gluud, C., Wetterslev, J., & Winkel, P. (2017). When and how should multiple imputation be used for handling missing data in randomised clinical trials—A practical guide with flowcharts. *BMC Medical Research Methodology*, *17*(1), Article 162. https://doi.org/10.1186/s12874-017-0442-1

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

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

Kirkham, J. J., Dwan, K. M., Altman, D. G., Gamble, C., Dodd, S., Smyth, R., & Williamson, P. R. (2010). The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. *BMJ*, *340*, Article c365. https://doi.org/10.1136/bmj.c365

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š, Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., … Nosek, B. A. (2018). Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Kvarven, A., Strømland, E., & Johannesson, M. (2019). Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*, *4*, 423–434. https://doi.org/10.1038/s41562-019-0787-z

Kyzas, P. A., Loizou, K. T., & Ioannidis, J. P. A. (2005). Selective reporting biases in cancer prognostic factor studies. *JNCI: Journal of the National Cancer Institute*, *97*(14), 1043–1055. https://doi.org/10.1093/jnci/dji184

Madden, K., Tai, K., Ali, Z., Schneider, P., Singh, M., Ghert, M., & Bhandari, M. (2019). Published intimate partner violence studies often differ from their trial registration records. *Women & Health*, *59*(1), 13–27. https://doi.org/10.1080/03630242.2017.1421287

McCarthy, R. J., Skowronski, J. J., Verschuere, B., Meijer, E. H., Jim, A., Hoogesteyn, K., Orthey, R., Acar, O. A., Aczel, B., Bakos, B. E., Barbosa, F., Baskin, E., Bègue, L., Ben-Shakhar, G., Birt, A. R., Blatz, L., Charman, S. D., Claesen, A., Clay, S. L., … Yıldız, E. (2018). Registered replication report on Srull and Wyer (1979). *Advances in Methods and Practices in Psychological Science*, *1*(3), 321–336. https://doi.org/10.1177/2515245918777487

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Group, T. P. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, *6*(7), Article e1000097. https://doi.org/10.1371/journal.pmed.1000097

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., Grahe, J. E., McCarthy, R. J., Musser, E. D., Antfolk, J., Castille, C. M., Evans, T. R., Fiedler, S., Flake, J. K., Forero, D. A., Janssen, S. M. J., Keene, J. R., Protzko, J., Aczel, B., … Chartier, C. R. (2018). The psychological science accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science*, *1*(4), 501–515. https://doi.org/10.1177/2515245918797607

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

O'Donnell, M., Nelson, L. D., Ackermann, E., Aczel, B., Akhtar, A., Aldrovandi, S., Alshaif, N., Andringa, R., Aveyard, M., Babincak, P., Balatekin, N., Baldwin, S. A., Banik, G., Baskin, E., Bell, R., Bialobrzeska, O., Birt, A. R., Boot, W. R., Braithwaite, S. R., … Zrubka, M. (2018). Registered replication report: Dijksterhuis and Van Knippenberg (1998). *Perspectives on Psychological Science*, *13*(2), 268–294. https://doi.org/10.1177/1745691618755704

Olsson-Collentine, A., Wicherts, J., Bakker, M., & van Aert, R. C. M. (2019). *Meta multiverse OSF repository*. https://osf.io/j8yg2/

Olsson-Collentine, A., Wicherts, J., Bakker, M., & van Aert, R. C. M. (2022). *Meta multiverse replication material* [Data set]. Zenodo.

Olsson-Collentine, A., Wicherts, J. M., & van Assen, M. A. L. M. (2020). Heterogeneity in direct replications in psychology and its association with effect size. *Psychological Bulletin*, *146*(10), 922–940. https://doi.org/10.1037/bul0000294

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867–872. https://doi.org/10.1016/j.jesp.2009.03.009

Page, M. J., McKenzie, J. E., & Higgins, J. P. T. (2018). Tools for assessing risk of reporting biases in studies and syntheses of studies: A systematic review. *BMJ Open*, *8*(3), Article e019703. https://doi.org/10.1136/bmjopen-2017-019703

Page, M. J., Sterne, J. A. C., Higgins, J. P. T., & Egger, M. (2020). Investigating and dealing with publication bias and other reporting biases in meta-analyses of health research: A review. *Research Synthesis Methods*, *12*(2), 248–259. https://doi.org/10.1002/jrsm.v12.2

Palpacuer, C., Hammas, K., Duprez, R., Laviolle, B., Ioannidis, J. P. A., & Naudet, F. (2019). Vibration of effects from diverse inclusion/exclusion criteria and analytical choices: 9216 different ways to perform an indirect comparison meta-analysis. *BMC Medicine*, *17*(1), Article 174. https://doi.org/10.1186/s12916-019-1409-3

Patel, C. J., Burford, B., & Ioannidis, J. P. A. (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

Pham, M. T., & Oh, T. T. (2021). Preregistration is neither sufficient nor necessary for good science. *Journal of Consumer Psychology*, *31*(1), 163–176. https://doi.org/10.1002/jcpy.v31.1

Pigott, T. D., Valentine, J. C., Polanin, J. R., Williams, R. T., & Canada, D. D. (2013). Outcome-reporting bias in education research. *Educational Researcher*, *42*(8), 424–432. https://doi.org/10.3102/0013189X13507104

R Core Team. 2020). *R: A language and environment for statistical computing* [Manual]. R Foundation for Statistical Computing. https://www.R-project.org/

Rankin, J., Ross, A., Baker, J., O'Brien, M., Scheckel, C., & Vassar, M. (2017). Selective outcome reporting in obesity clinical trials: A cross-sectional review: Reporting outcomes in obesity clinical trials. *Clinical Obesity*, *7*(4), 245–254. https://doi.org/10.1111/cob.2017.7.issue-4

Revelle, W. (2020). *Psych: Procedures for psychological, psychometric, and personality research* [Manual]. https://CRAN.R-project.org/package=psych

Roese, N. J., & Vohs, K. D. (2012). Hindsight bias. *Perspectives on Psychological Science*, *7*(5), 411–426. https://doi.org/10.1177/1745691612454303

Rombach, I., Rivero-Arias, O., Gray, A. M., Jenkinson, C., & Burke, Ó. (2016). The current practice of handling and reporting missing outcome data in eight widely used PROMs in RCT publications: A review of the current literature. *Quality of Life Research*, *25*(7), 1613–1623. https://doi.org/10.1007/s11136-015-1206-1

Rongen, J. J., & Hannink, G. (2016). Comparison of registered and published primary outcomes in randomized controlled trials of orthopaedic surgical interventions. *The Journal of Bone and Joint Surgery*, *98*(5), 403–409. https://doi.org/10.2106/JBJS.15.00400

Rosati, P., Porzsolt, F., Ricciotti, G., Testa, G., Inglese, R., Giustini, F., Fiscarelli, E., Zazza, M., Carlino, C., Balassone, V., Fiorito, R., & D'Amico, R. (2016). Major discrepancies between what clinical trial registries record and paediatric randomised controlled trials publish. *Trials*, *17*(1), Article 430. https://doi.org/10.1186/s13063-016-1551-6

Rousseeuw, P. J., & Hubert, M. (2011). Robust statistics for outlier detection. *WIREs Data Mining and Knowledge Discovery*, *1*(1), 73–79. https://doi.org/10.1002/widm.v1.1

Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology*, *10*, Article 813. https://doi.org/10.3389/fpsyg.2019.00813

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., … Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. https://doi.org/10.1177/2515245917747646

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

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

Stanley, T. D., & Doucouliagos, H. (2014). Meta-Regression approximations to reduce publication selection bias. *Research Synthesis Methods*, *5*(1), 60–78. https://doi.org/10.1002/jrsm.v5.1

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

Stefan, A., & Schnbrodt, F. (2022). *Big little lies: A compendium and simulation of p-hacking strategies*. PsyArXiv. https://doi.org/10.31234/osf.io/xy2dk

Stewart, L. A., & Tierney, J. F. (2002). To IPD or not to IPD? Advantages and disadvantages of systematic reviews using individual patient data. *Evaluation & the Health Professions*, *25*(1), 76–97. https://doi.org/10.1177/0163278702025001006

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2007). *Using multivariate statistics*. Pearson.

van Aert, R. C. M. (2020). *Puniform: Meta-analysis methods correcting for publication bias* [Manual]. https://CRAN.R-project.org/package=puniform

van Aert, R. C. M. (2022). Analyzing data of a multilab replication project with individual participant data meta-analysis: A tutorial. *Zeitschrift Für Psychologie*, *230*(1), 60–72. https://doi.org/10.1027/2151-2604/a000483

van Aert, R. C. M., & van Assen, M. A. L. M. (2020). *Correcting for publication bias in a meta-analysis with the* p*-uniform* method*. MetaArXiv.

van Aert, R. C. M., Wicherts, J. M., & van Assen, M. A. L. M. (2016). Conducting meta-analyses based on *p* values: Reservations and recommendations for applying *p*-uniform and *p*-curve. *Perspectives on Psychological Science*, *11*(5), 713–729. https://doi.org/10.1177/1745691616650874

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Reply to Inbar: Contextual sensitivity helps explain the reproducibility gap between social and cognitive psychology. *Proceedings of the National Academy of Sciences*, *113*(34), E4935–E4936. https://doi.org/10.1073/pnas.1609700113

van Ginkel, J. R., Sijtsma, K., van der Ark, L. A., & Vermunt, J. K. (2010). Incidence of missing item scores in personality measurement, and simple item-score imputation. *Methodology*, *6*(1), 17–30. https://doi.org/10.1027/1614-2241/a000003

Verschuere, B., Meijer, E. H., Jim, A., Hoogesteyn, K., Orthey, R., McCarthy, R. J., Skowronski, J. J., Acar, O. A., Aczel, B., Bakos, B. E., Barbosa, F., Baskin, E., Bègue, L., Ben-Shakhar, G., Birt, A. R., Blatz, L., Charman, S. D., Claesen, A., Clay, S. L., … Yıldız, E. (2018). Registered replication report on Mazar, Amir, and Ariely (2008). *Advances in Methods and Practices in Psychological Science*, *1*(3), 299–317. https://doi.org/10.1177/2515245918781032

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, *60*(3), 419–435. https://doi.org/10.1007/BF02294384

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Voracek, M., Kossmeier, M., & Tran, U. S. (2019). Which data to meta-analyze, and how? A specification-curve and multiverse-analysis approach to meta-analysis. *Zeitschrift Für Psychologie*, *227*(1), 64–82. https://doi.org/10.1027/2151-2604/a000357

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., Albohn, D. N., Allard, E. S., Benning, S. D., Blouin-Hudon, E.-M., Bulnes, L. C., Caldwell, T. L., Calin-Jageman, R. J., Capaldi, C. A., Carfagno, N. S., Chasten, K. T., Cleeremans, A., Connell, L., DeCicco, J. M., … Zwaan, R. A. (2016). Registered replication report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, *11*(6), 917–928. https://doi.org/10.1177/1745691616674458

Wagenmakers, E.-J., Sarafoglou, A., & Aczel, B. (2022). One statistical analysis must not rule them all. *Nature*, *605*(7910), 423–425. https://doi.org/10.1038/d41586-022-01332-8

Wayant, C., Scheckel, C., Hicks, C., Nissen, T., Leduc, L., Som, M., & Vassar, M. (2017). Evidence of selective reporting bias in hematology journals: A systematic review. *PLoS One*, *12*(6), Article e0178379. https://doi.org/10.1371/journal.pone.0178379

Wicherts, J. M. (2012). *Cohort differences in big five personality factors over a period of 25 years*. Data Archiving and Networked Services (DANS). https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:51655

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid *p*-Hacking. *Frontiers in Psychology*, *7*, Article 1832. https://doi.org/10.3389/fpsyg.2016.01832

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag. https://ggplot2.tidyverse.org

Wickham, H., & Bryan, J. (2019). *Readxl: Read excel files* [Manual]. https://CRAN.R-project.org/package=readxl

Wickham, H., François, R., Henry, L., & Müller, K. (2020). *Dplyr: A grammar of data manipulation* [Manual]. https://CRAN.R-project.org/package=dplyr

Wickham, H., & Miller, E. (2020). *Haven: Import and export "SPSS", "stata" and "SAS" files* [Manual]. https://CRAN.R-project.org/package=haven

Zhu, H. (2019). *kableExtra: Construct complex table with "kable" and pipe syntax* [Manual]. https://CRAN.R-project.org/package=kableExtra

Received January 27, 2022
Revision received October 27, 2022
Accepted November 07, 2022 ■