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

Anton Olsson-Collentine[1], Robbie van Aert[1], Marjan Bakker[1], and Jelte Wicherts[1]

[1] **Department of Methodology and Statistics, Tilburg School of Social and Behavioral Sciences, Tilburg University, the Netherlands.**

Corresponding author:
Anton Olsson-Collentine[1]

Email address: anton@olssoncollentine.com

ABSTRACT

Researcher degrees of freedom refer to arbitrary decisions that exist in the execution and reporting of hypothesis-testing research. These allow for many possible outcomes from a single study. Selective reporting of results (e.g., *p*-hacking) from this ‘multiverse’ of outcomes can lead to inflated effect size estimates and false positive results. We studied the effects of researcher degrees of freedom and selective reporting across direct replication studies using empirical data, taking advantage of the unique opportunity offered by the open data of extensive multi-study projects in psychology (Registered Replication Reports). Our results show the substantial variability in effect sizes that researcher degrees of freedom can cause in a study, and how selective reporting combined with this variability can yield different conclusions and bias in meta-analysis. We also observed that the effect of a particular researcher degree of freedom was inconsistent across direct replication studies. We recommend hypothesis-testing researchers to preregister their preferred analysis and perform multiverse analysis. We propose that multiverse analysis be accompanied by an index (Underlying Multiverse Variability; UMV) that quantifies the robustness of the results.

Researcher Degrees of Freedom [DFs; Simmons, Nelson, and Simonsohn (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 modelling decisions (for more examples, see Jelte M. Wicherts et al. 2016). Researcher DFs allow for many possible outcomes in a single study; a 'multiverse' of statistical results (Steegen et al. 2016) 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; Huntington-Klein et al. 2020; Breznau et al. 2021).

However, 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 that they allow for selective reporting. 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), although we do not include in this definition the special case where no study results are reported (i.e., publication bias). Selective reporting is often focused on the significance of outcomes and can be intentional ('*p*-hacking') or happen unintentionally due to forgetfulness or hindsight and confirmation biases (Roese and Vohs 2012; Nickerson 1998). Selective reporting from the multiverse of statistical results is worrisome: It can allow researchers to present statistical evidence for almost any claim (Simmons, Nelson, and Simonsohn 2011), leading to an overrepresentation of false positive findings in the literature (Ioannidis 2005) and inflated effect size estimates (Ioannidis 2008) in the literature.

Unfortunately, selective reporting appears common amongst researchers. In psychology, about 50-60% of researchers admit to not reporting all dependent measures in a study (John, Loewenstein, and Prelec 2012; Agnoli et al. 2017), and in a study registry comparison 70% of studies did not report all outcome variables (Franco, Malhotra, and Simonovits 2016). Moreover, there is an 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 and Hannink 2016), obesity (Rankin et al. 2017), and cancer research (Kyzas, Loizou, and Ioannidis 2005). A recent study examining results in 67 trials published between October and November 2015 in 5 top journals from general medicine found that 42% of pre-specified 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 have their 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.

The effects of selective reporting have been studied in simulated data for meta-analysis (e.g., Botella et al. 2021; Carter et al. 2019; Friese and Frankenbach 2020), but simulated data may not be representative of observed data in practice. The effects of researcher DFs have been studied in observed data of individual studies (e.g., Silberzahn et al. 2018; Botvinik-Nezer et al. 2020; Huntington-Klein et al. 2020; Breznau et al. 2021), but not in meta-analysis. 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 recent multi-lab direct replication projects in psychology (Registered Replication Reports). In doing so, we illustrate the entire process of selective reporting, from the arbitrary decisions in primary studies that enable the practice to the consequences for meta-analysis.

METHODS

Registered Replication Reports (RRRs) each consist of a set of studies (labs) that collected data on an effect in psychology using the same pre-specified 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, we use the open data from these extensive studies to examine the variability in results across many alternative analytic outcomes. Our design allows 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 significance of the outcomes in the primary studies.

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 1) examining the underlying multiverse variability in effect size estimates at the primary study level, 2) examining the resulting multiverse variability at the meta-analytic level, and 3) 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 Underlying Multiverse Variability (UMV, statistically defined in 'the multiverses' section).

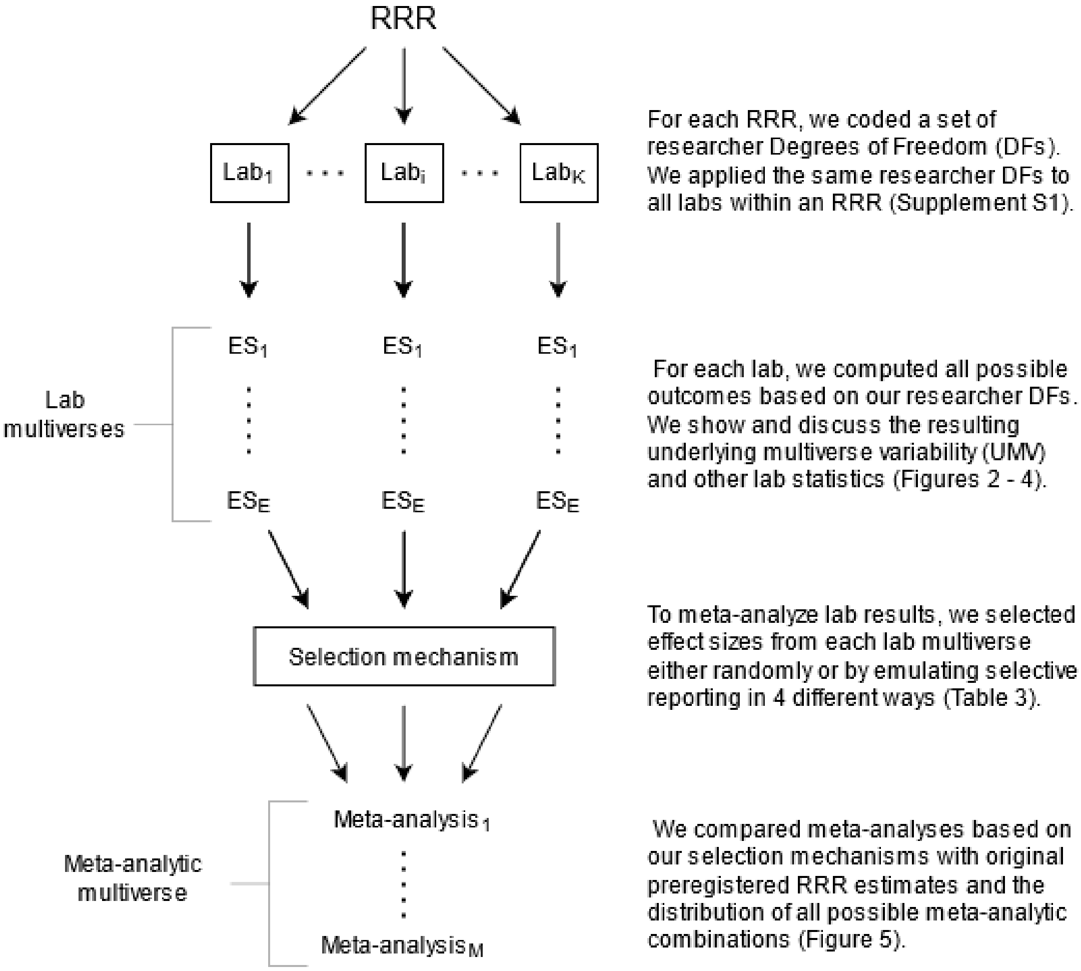
RRR
Lab1
Labi
LabK
For each RRR, we coded a set of researcher Degrees of Freedom (DFs). We applied the same researcher DFs to all labs within an RRR (Supplement S1).
Lab multiverses
ES1
ES1
ES1
ESE
ESE
ESE
For each lab, we computed all possible outcomes based on our researcher DFs. We show and discuss the resulting underlying multiverse variability (UMV) and other lab statistics (Figures 2 - 4).
Selection mechanism
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).
Meta-analytic multiverse
Meta-analysis1
Meta-analysisM
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).

Figure 1. Summary of the study design. For each lab in a Registered Replication Report (RRR) multiverses were computed, analyzed and used for meta-analysis. ES = Effect size. Each RRR consists of K labs, lab i = 1, 2, 3, ... 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-anlytic multiverse of size M. We approximate the meta-analytic multiverse by randomly sampling 10^5 meta-analyses from the meta-analytic multiverse.

Transparency and openness

All our code and data for this project are available on the Open Science Framework (OSF) at osf.io/j8yg2/. We refer directly to relevant files on the OSF using brackets and links in the sections below. We registered data cleaning code and the researcher DFs available in each RRR before proceeding to analysis (osf.io/h397y/). We only made minor code corrections and clarifications of researcher DFs after registration, fully detailed in Supplement S2 (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.

Data Collection

We included all RRRs published up until 2019-05-15 available in the journals ‘Perspectives on Psychological Science’ and ‘Advances in Methods and Practices in Psychological Science’ (see also Olsson-Collentine, Wicherts, and van Assen 2020). Three RRRs (RRR3, RRR5, RRR9) had multiple primary outcome variables (as explicitly identified in the accompanying publications). In total, we included 10 projects containing 14 primary outcome variables that could be meta-analyzed, consisting of 236 labs and 37,602 participants (Table 1).

Table 1.

Preregistered Multi-Lab Replication Projects

RP	Paper	Countries	Labs	Effects	N	Sample and Settings	Description of Effects
RRR1	Alogna et al. (2014)	10	31	1	4832	31/32 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	2932	22/23 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	1210	11/12 samples were undergraduate students mostly aged 18-25, one of which was online. 1 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 imperfect or perfect tense and then rated protagonist’s intentions (intentionality/intention attribution/detailed processing).
RRR4	Hagger et al. (2016)	10	24	1	3127	All samples consisted of in-lab undergraduate students	Ego depletion; Independent two-group experiment. Participants 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	2279	All samples consisted of in-lab undergraduate students aged 18-25	Commitment on neglect/exit; Independent two-group experiment with two outcome variables. Participants either primed to think about commitment to or independence from partner.
RRR6	Wagenmakers et al. (2016)	8	17	1	2542	All but one sample explicitly consisted of students and all took place in-lab. The last sample was recruited at university grounds.	Facial feedback hypothesis; Independent two-group experiment. Participants either induced to ’smile’ or ’pout’ by holding a pen in their mouth differently and simultaneously rated funniness of cartoons.
RRR7	Bouwmeester et al. (2017)	12	21	1	3669	All samples consisted of in-lab undergraduate students aged 18-34.	Intuitive cooperation; Independent two-group experiment. Economic game with money contribution to a common pool either under time pressure or time delay.
RRR8	O’Donnell et al. (2017)	13	40	1	7045	All samples consisted of in-lab undergraduate students aged 18-25	Professor priming; Independent two-group experiment. Participants primed with either a ’professor’ or ’hooligan’ stimuli. Outcome was percentage correct trivia answers.
RRR9	McCarthy et al. (2018)	13	26	2	6721	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	3245	All samples consisted of in-lab students aged 18-25	Moral reminder; Independent two-group experiment. Participants either recalled the Ten Commandments or books they had read. Outcome was degree of cheating when reporting results.

Note:
RP = Replication Project, Countries = number of lab country locations, Effects = number of primary effects studied, N = participants before exclusions, RRR = Registered Replication Report. Table adapted with permission from Olsson-Collentine et al. (2020). Code to reproduce table: osf.io/jehpy/

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 2 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 (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.

The Multiverses

A multiverse analysis consists 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. One assumption of multiverse analysis is that the alternative decisions are all (approximately) equally reasonable (Steegen et al. 2016). Consequently, we can see the distribution of effect sizes in the multiverse as random variability around a true score. This perspective assumes that no researcher DF has a systematic effect (i.e., is an actual moderator of the effect), a point we return to in the discussion.

How large the multiverse variability is depends on which researcher DFs are considered reasonable. We refer to the variability underlying a given set of researcher DFs as the Underlying Multiverse Variability (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 UMV. Other researchers focus on the range of effect sizes in the multiverse (e.g., Patel, Burford, and Ioannidis 2015), but we consider it more useful to treat multiverse variability in terms of the standard deviation, in line with how sampling error is defined.

Our coded researcher DFs each consisted of a decision that needs to be made and a number of associated potential options for that decision. 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. To do so, we proceeded in two steps: We 1) set up a list of 'common' researcher DFs, and then 2) 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. We created our list of common researcher DFs (Table 2) based on recommendations in statistical textbooks, common decisions by applied researchers as reported in research literature (see references in Table 2), data analytic decisions made by the included RRRs, and our own experience of decisions encountered in the literature.

Table 2.

List of common researcher degrees of freedom applied to Registered Replication Reports

Decision	Options	Explanation
S1. Post hoc scale length	a) No adjustment b) Drop the item with the lowest item-rest correlation c) Drop the two items with the lowest item-rest correlations	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 seems 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)
S2. Composite score	a) Unweighted average score b) Sum score c) PCA score: Varimax rotation, force two components and pick the first, requires at least 3 items.	For Likert-type scales with multiple items. Other DVs e.g., reaction time variability (RRR3), dichotomous correct/incorrect (RRR1&2), continuous measures (RRR7), single item DVs (RRR10) may need more unique choice options. We chose Varimax rotation to maximize variance between outcomes.
E1. Missingness DV	a) Any missing items -> list-wise deletion b) If $\leq$25% items missing then pair-wise deletion of missing items. Otherwise list-wise.	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 (Eekhout et al., 2012; Rombach et al., 2016; Burton & Altman, 2004). 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.
E2. Missingness E3-E4 variables	a) No exclusion b) Missing on any variable -> list-wise deletion	RRRs that excluded data based on a 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).
E3. Age	a) No exclusion b) Not 18-24 c) Not 18-23 d) Not 18-22 e) Not 18-21	Used by 9/10 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 et al, 2012). The oldest students in this dataset were 25. Bachelor psychology students are common samples for psychological research (e.g., Heinrich et al., 2010), but may be becoming less common with an increase in internet samples (Sassenberg & Ditrich, 2019)
E4. Language / Student / Ethnicity	a) No exclusion b) Exclude participants not belonging to the dominant category	Used by 3/10 RRRs. Demographic variables which 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 variables are available they are treated as separate exclusion criteria.
E5. Attention check	a) No exclusion 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	Attention checks are common in psychology, as evidenced by the more than 1500 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.
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. Tabachnik, Fidell and Ullman (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: osf.io/jehpy/.

Although there are many researcher DFs before analyzing the data (Jelte M. 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, our coded researcher DFs can be divided into two categories: a) how the composite score was computed from multiple indicators (decision points starting with 'S.' in Table 2), and b) different exclusion criteria (decisions starting with 'E.' in Table 2). 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). We registered the full list (coding protocol) of common, and unique, researcher DFs before analyzing any data.

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 design, and consisted of different exclusion criteria. We can separate between two types of 'unique' researcher DFs: either 1) the RRRs excluded participants based on some variable that was not defined in our list of common decisions or 2) 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 less than 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 used this variable for exclusions. Collecting data on a variable with no clear purpose thus adds researcher DFs, which we in this case used to create our multiverses.

The RRRs ended up with between 2 (RRR3) and 10 (RRR7) unique researcher DFs, each decision with 2-6 associated options. After registering the coding protocol, we coded the applicability of the 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 pre-screened 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 S1 (osf.io/wj38n/).

We computed resulting effect sizes from all possible combinations of decisions for each lab in an RRR (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*, p. 226, Borenstein 2009) and computed log odds ratios for RRR1 and RRR2. Effect sizes were originally analyzed unstandardized in all RRRs except for

RRR4, but we wanted to draw conclusions across RRRs and this required standardizing the effect sizes.

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, van Dijk, and Wicherts 2012), in our primary analyses. Three labs in RRR2 (L09, L17, 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 S3.

Analysis

To explore the effects of researcher DFs on research output, we 1) examined the variance in effect size estimates at the lab level (lab multiverses) and 2) 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 and the standard deviations in effect size resulting from variation in 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 S1). 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, 3) then we always applied them in the order 1, 2, 3 regardless of chosen option, rather than also varying the order (e.g., 2, 1, 3). This fixed order does not affect final effect size estimates but makes it impossible to compute the impact of a single researcher DF across all possible researcher DF combinations. 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 697×10ˆ33 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 alpha,’ ‘Random significant,’ ‘Bounded significant’) with different motivations. All selection mechanisms were based on statistical significance, and we used a two-tailed test with α =.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 alpha’). 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 alpha’ condition, we selected, for each lab in an RRR, the outcome with a *p*-value closest below .05 (or, if there were no *p*-values below .05, the lowest value). These two approaches (‘most significant,’ ‘below alpha’) 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 1) randomly drawing 100 possible outcomes and 2) 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 [link to R-script]. 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.

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 pre-registered 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 signficant	Yes	Yes	Select the effect size in the multiverse with the smallest p-value.
Below alpha	Yes	Yes	Select the effect size in the multiverse with a p-value closest below $p = 0.05$. If no p-value 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 table: osf.io/jehpy/.

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, Strømland, and Johannesson 2019). We applied three publication bias correction methods [PET-PEESE, 3PSM, and p-uniform*; Stanley and Doucouliagos (2014); Vevea and Hedges (1995); van Aert and van Assen (2020)] to examine their applicability to selective reporting in the absence of publication bias.

All our data and code are freely available on the OSF in the repository 'Meta multiverse' at osf.io/j8yg2/.

RESULTS

After excluding conditions that resulted in less than 24 participants per experimental group, 8/14 RRR multiverses decreased in size, see 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.

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	Percentage remaining
RRR01	5	5	116 [107, 125]	3,840	3,840	100%
RRR02	5	5	88 [84, 98]	3,840	3,840	100%
RRR03 Attribution	6	2	84 [82, 84]	3,840	3,816	99%
RRR03 Intention	6	2	84 [82, 84]	3,840	3,840	100%
RRR03 Process	7	2	84 [83, 85]	7,680	7,680	100%
RRR04	5	3	76 [68, 90]	23,040	20,160	88%
RRR05 Exit	6	9	82 [70, 94]	2,488,320	1,503,904	60%
RRR05 Neglect	6	9	82 [70, 94]	2,488,320	1,540,176	62%
RRR06	7	5	96 [77, 111]	122,880	61,440	50%
RRR07	3	10	74 [63, 95]	2,621,440	525,680	20%
RRR08	8	4	79 [65, 102]	115,200	19,200	17%
RRR09 Behavior	8	4	169 [114, 218]	46,080	46,080	100%
RRR09 Hostility	8	4	168 [114, 218]	46,080	46,080	100%
RRR10	5	5	90 [77, 107]	11,520	3,808	33%

Note:
Meta-analytic distributions and estimates after excluding analytic choices that resulted in <24 particants per experimental group at the study-level. DFs = Researcher 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: osf.io/jehpy/.

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 16/24 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 S4 (osf.io/2htc6/).

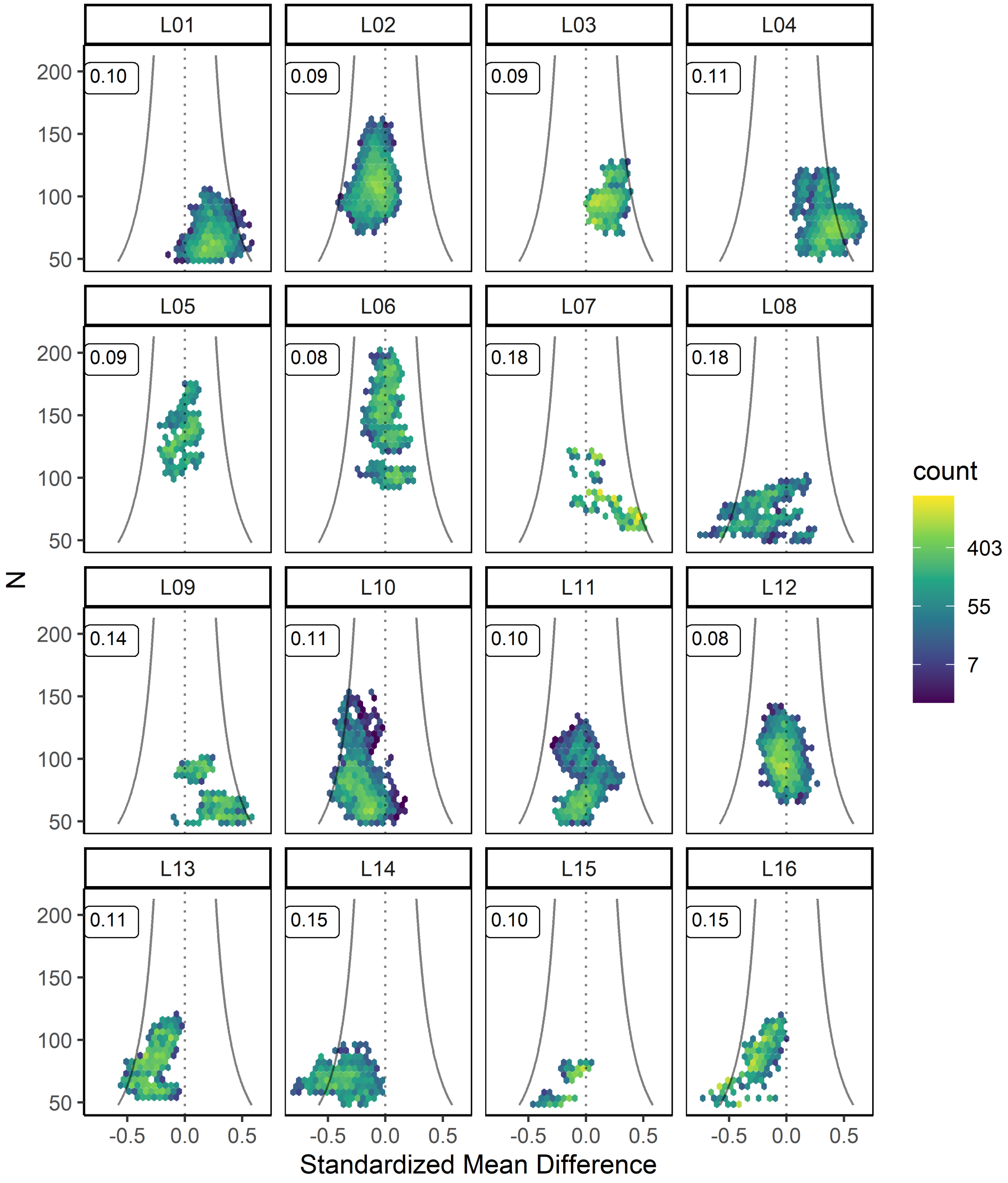

Figure 2. Arbitrary decisions in research cause underlying multiverse variability (UMV) in effect size estimates. Funnel plots showing the effect sizes based on the multiverses in 16 labs for RRR04, after removing cases where $n < 24$ in either experimental group. Values in upper left corner of each facet are UMV for each lab. For legibility, 16/24 RRR04 labs are shown; the figure including all labs is available in Supplement S4 (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: osf.io/thuyk/.

Overall in Figure 2 few (median = 0.87%, interquartile range = 0 - 3%) possible outcomes within labs were statistically significant (indicated by observations falling outside the funnel lines), but there were labs with a higher proportion of significant outcomes (e.g., L14 = 25%, L04 = 24%, L16 = 17%). The median underlying multiverse variability (UMV) across labs in Figure 2 was 0.1SD, 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 0.11SD (IQR = 0.08 - 0.14) for SMD effect sizes and 0.07SD (IQR = 0.04 - 0.12) for log 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 (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 - 0.37). Hence, a large sample size does not ensure a small UMV.

There can be substantial variation between labs 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 design across labs and the same researcher DF being applied, no two bar plots look identical and it differs between labs which researcher DF leads to the most variance. 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. There are also researcher DFs that tend to contribute little to variation in effect sizes; changing exclusion criteria regarding Ethnicity (E4.2), Language (E4.1) and whether to exclude participants with missingness on the age or demographic variables (E2) has little effect in most labs in 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' methods section), which explains the larger range of effect sizes in Figure 2.

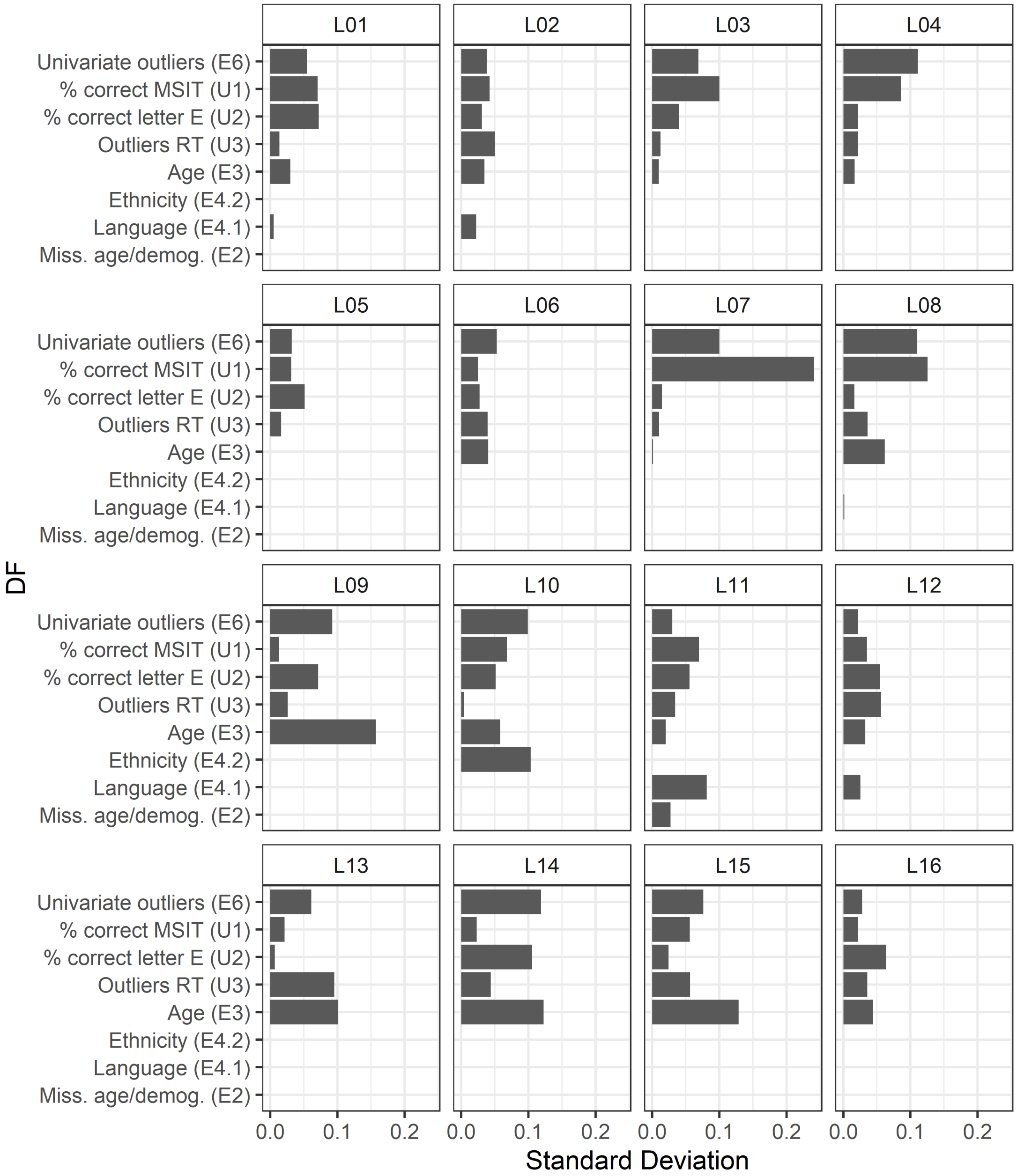

Figure 3. The same arbitrary decisions have a different effect in comparable studies. Standard deviation (SD) in effect size estimates in 16 labs in RRR04 resulting from applying different researcher degrees of freedom (DFs) individually, after removing cases where $n < 24$ in either experimental group. For legibility, 16/24 RRR04 labs are shown; the figure including all labs is available in Supplement S4 (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 S1 osf.io/wj38n/). The Y-axis is ordered by median SD across labs. Code to reproduce figure: osf.io/thuyk/

Some researcher DFs strongly contributed to the UMV. Figure 4 shows the SDs in estimated effect sizes in labs resulting from applying each researcher DF individually, disaggregated across RRRs and then aggregated into their respective categories. Excluding participants on age appears to have had a consistently strong effect (median upper panel, log OR = 0.05, lower panel, *d* = 0.03, likely reflecting that age is a variable applicable to all participants and with many (5) associated options (see Table 2 E3). 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), highlighting the importance of measurement.

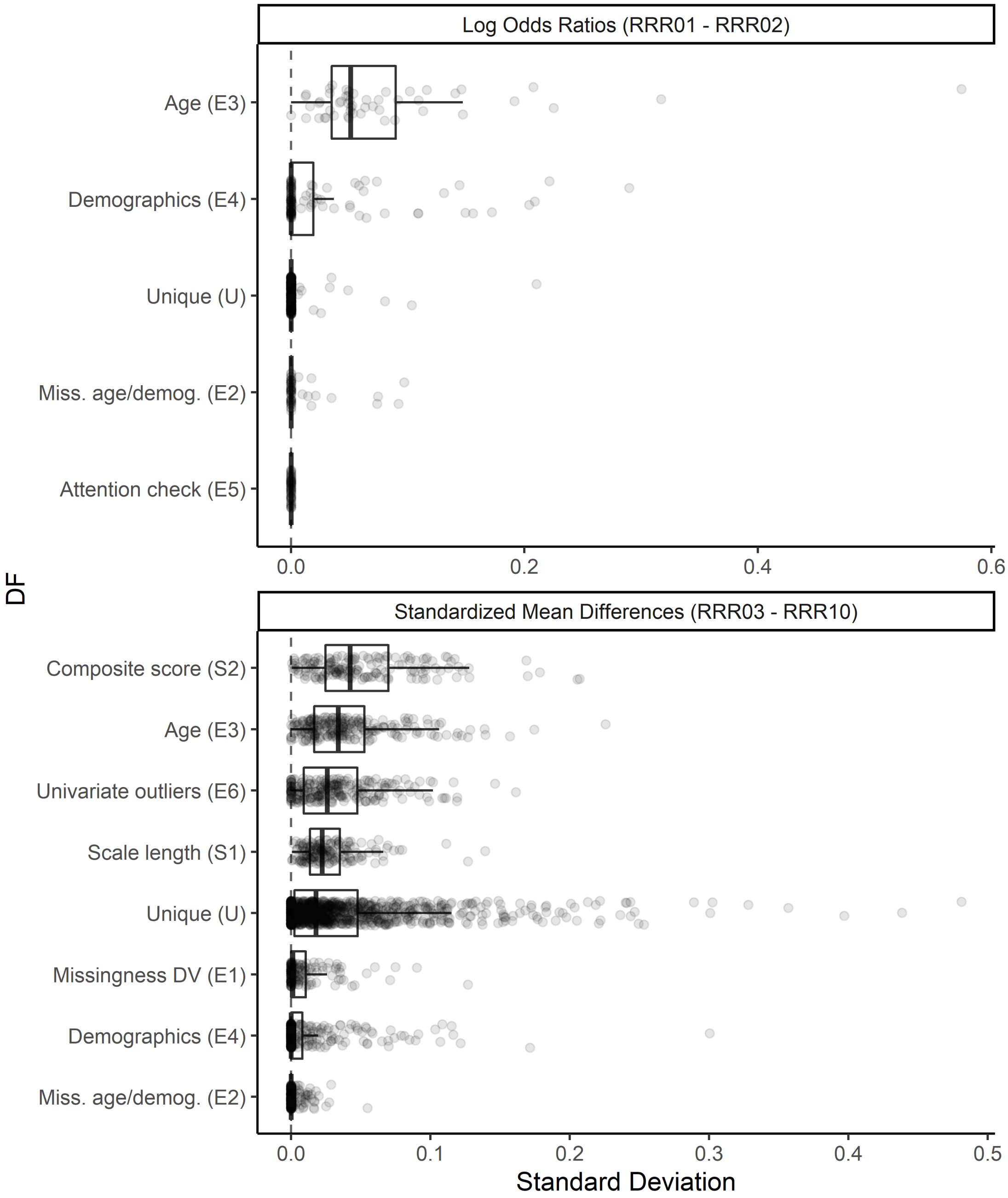

Figure 4. Some arbitrary decisions tend to create more effect size variability than others. Standard deviation (SD) in effect size estimates in labs resulting from different researcher degrees of freedom

(DFs) applied individually. The top panel shows results for Registered Replication Reports (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 Registered Replication Reports (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: osf.io/thuyk/.

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 S4). This same researcher DF in RRR07 resulted in 4 out of the 5 highest effect size SDs in Figure 4. The remaining observation (third from the right) belonging to RRR09 Hostility, lab 1, and resulting 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 RRR01/02 only being applicable to a few of the constituent labs (see Supplement S1). 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 lead to different 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 0.02SD (RRR09 Hostility) to 0.04SD (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 5).

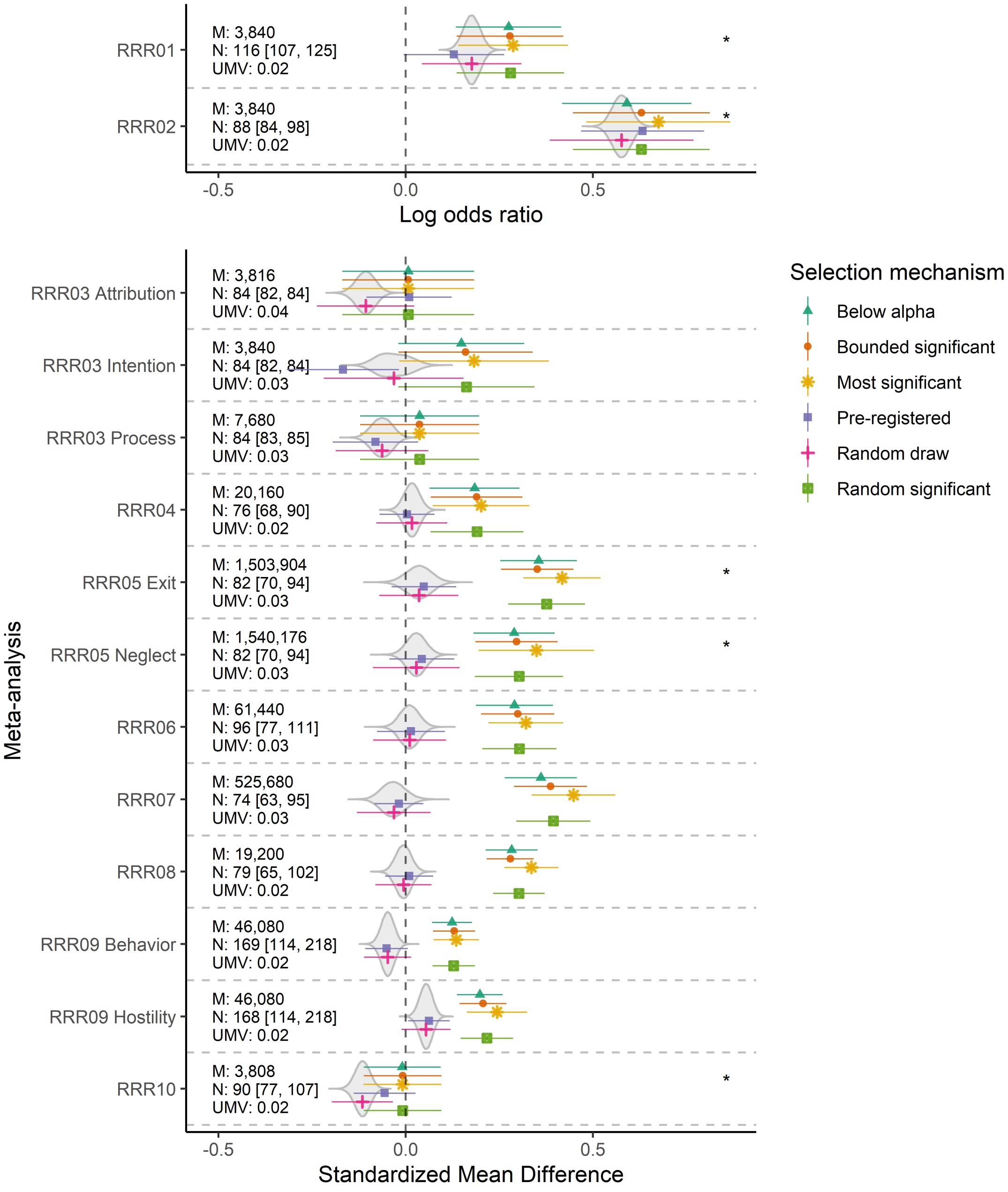

Figure 5. Selective reporting in labs results in overestimates in meta-analysis. Meta-analytic distributions and estimates after excluding analytic decisions that resulted in $n < 24$ participants per experimental group at the study-level. Selection mechanism = how effect sizes were selected at the study-level, either by *p*-hacking ("Most significant," "Bounded significant," "Random significant," "Below alpha,"), 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 figure: osf.io/thuyk/.

When researcher DFs were combined with 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 (median 0.05% of the possible outcomes), and our biased selection procedures resulting in the same results if there were no significant outcomes in a lab (134 / 294 labs across all RRRs, counting labs with multiple DVs as separate measurements).

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 S5), 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, Wicherts, and van Assen 2016).

DISCUSSION

In this article, we have studied the joint effect of researcher DFs and selective reporting on reported effect size estimates both within and across multiple direct replication studies, taking advantage of the unique opportunity offered us by studies of the type Registered Replication Report (RRR) to do so using empirical data. We made use of data from ten of these preregistered multi-lab direct replication projects in psychology, containing a total of 14 primary outcome variables. We identified arbitrary decision points 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 what could have emerged had the primary studies not been preregistered. Our results showed that researcher DFs can lead to substantial variability in effect sizes in primary studies, but that the distribution of this variability differs considerably between direct replication studies even when applying the same researcher DFs. Furthermore, which researcher DF caused variability within a study differed across between direct replications, suggesting that the effect of any researcher DF in a study is unpredictable. Nonetheless, we demonstrated that selective reporting in primary studies can severely bias meta-analyses even in the absence of publication bias.

A multiverse of outcomes

Researcher DFs allow many possible outcomes to be reported from a single study. The median UMV, the standard deviation in effect sizes based on coded researcher DFs, was about 0.1SD in our SMD data. The UMV will differ depending on research field and which DFs researchers find reasonable. However, regardless of preferences we believe most researchers will agree that in most research there are alternative reasonable decisions that could be made, and that these decisions may affect outcomes. As such, researchers should recognize that researcher DFs create another layer of uncertainty in their estimates. We advise researchers doing hypothesis-testing research to 1) 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 2) include a multiverse analysis as a sensitivity analysis and report their UMV. 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.

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 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 relatively low, with an UMV of at most 0.04SD amongst SMD effects. It may be that when researcher's decisions are truly random, and in the absence of selective reporting and publication bias, researcher DFs in primary studies are not a major concern for meta-analysts as the decisions would be expected to cancel out across studies. Unfortunately, such ideal data is not realistic to expect 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, 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 and 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 and Oh (2021)]. If preregistered results generally are close to multiverse means also in larger datasets than ours, it may be that preregistration and multiverse analysis can be used individually as somewhat interchangeable credibility-enhancing tools, each with their own advantages and disadvantages. 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. To account for UMV such preregistered multi-lab 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.

Exploring the multiverse

Creating a multiverse is an inherently subjective endeavor given that researchers might disagree about which decisions are (approximately) equally reasonable. For example, although we created our researcher DFs based on common practice in the associated literature, there are strong methodological arguments to consider carefully the meaning and impact of outliers, or use outlier robust statistics (e.g., Rousseeuw and 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. Which multiverses are considered reasonable will be affected not only by individual researchers' beliefs, but also by which decisions their research community consider acceptable in terms of theory, methodology, and empirical standards.

Fundamentally, disagreement about the appropriateness of a particular multiverse amounts to disagreement about whether a (or multiple) researcher DFs or associated options is a moderator of (i.e., have a systematic effect on) the outcome of interest. Some researchers may be tempted to directly interpret the existence of researcher DFs and resulting UMV as evidence of 'hidden moderators' (Van Bavel 2016); currently unknown moderators that explain why effect sizes differ between studies. However, the existence of the multiverse does not by itself imply moderators. The design of our project enabled us to examine the effect of researcher DFs across multiple direct replication studies. We observed that 1) the same researcher

DFs applied to direct replication studies resulted in widely varying distributions of effect sizes, and 2) 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 was generally non-systematic. These results corroborate previous findings of ours (Olsson-Collentine, Wicherts, and van Assen 2020) that differences in study results in social and cognitive psychology are rarely systematic across direct replication studies (i.e., show little to no between-study heterogeneity), and that the best explanation for differences between effect sizes in (direct) replication studies is 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, Arends, and Stijnen 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 and Tierney 2002). In the case of multiple dependent variables, which might also be a researcher DF, potential systematic differences could be examined in multivariate meta-analysis (e.g., Jackson, Riley, and White 2011). As with multiverse analysis, such moderator analyses should primarily be considered exploratory and hypothesis-generating.

Simply computing the UMV for a researcher DF may be useful when its different options are expected to give the same statistical result. For example, with psychometrically sound scales our alternative ways of computing scale composites (unweighted average score across items, sum score, or first component from PCA) should have been very nearly equivalent (e.g., McNeish and Wolf 2020). Yet they resulted in a considerable UMV, which may reflect inadequate attention to psychometric scale development amongst our (non-random) sample of social and cognitive multi-lab studies (see also Shaw et al. 2020). That said, these multi-lab projects were replication projects, and there is an understandable tension between exact replication and design improvements. More generally, multi-lab 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.

Selective reporting from the multiverse

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 0.1SD amongst 236 labs. Nonetheless, a median UMV of 0.1SD in a field still implies that selective reporting can turn a statistically non-significant effect into a significant effect. The extent of UMV in a field depends on the multiverse created, and our estimates may only apply to our non-random sample of social and cognitive psychology research. UMV in other fields could be either larger or smaller, but is unlikely to be non-existent, and it may be worthwhile to study the UMV in different subfields to examine their susceptibility to selective reporting. Generally, the extent of bias introducible 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.

We do not mean to imply that exploration of researcher DFs are 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. 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 (as recommended by e.g., PRISMA and MARS: Moher et al. 2009; Appelbaum et al. 2018). 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, McKenzie, and Higgins 2018) that assist in making research more transparent.

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. Especially when a meta-analysis includes primary studies at risk of selective reporting, the average effect size estimate may be biased in the hypothesized direction. 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, Strømland, and Johannesson (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 multi-lab projects, although this also includes publication bias. Larger or smaller differences may be more typical in other fields. Our results corroborate recommendations from meta-analysis reporting protocols such as PRISMA and MARS (Moher et al. 2009; Appelbaum et al. 2018) 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, Wicherts, and van Assen 2016), our supplemental results show that publication bias methods should not be relied on to correct for *p*-hacking.

Under some assumptions, we can be confident that the UMV 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 always 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 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. Studies published as so-called Registered Reports (Chambers 2019) fulfill these criteria. We note that 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.

It may be insightful to do more complex modelling of selective reporting from a multiverse perspective. Our modelling of it in this study was relatively straightforward. Our purpose here was not to model selective reporting extensively, but rather explore specifications to see how extreme the resulting bias in meta-analytic estimates might be, and we only attempted to model the outcomes of intentional selective reporting (*p*-hacking). Contrary to two recent simulation studies (Botella et al. 2021; Friese and Frankenbach 2020) we found that *p*-hacking could cause substantial inflation of meta-analytic average effect sizes. We know from previous studies that the type of *p*-hacking matters; incremental methods such as optional stopping have little effect on effect size inflation, whereas outcome reporting bias has a large effect (Francis 2012; Kirkham et al. 2010). As such, differences are likely explained by the incremental *p*-hacking methods used in these simulation studies as compared to our methods of (primarily) selecting the lowest *p*-values, possibly in combination with our use of more realistic data. We hope our data will also be useful to other researchers interested in more complex modelling of research bias.

Conclusion

Researcher DFs cause unpredictable variability in effect size estimates that makes interpretation of results more difficult in single studies, yet from the perspective of the broader literature and meta-analysis, researcher DFs are primarily a concern because they allow for selective reporting. More concrete theory, statements about the constraints on generality (Simons, Shoda, and Lindsay 2017) or agreed upon field standards will decrease researcher DFs and the resulting UMV. However, we do not believe that researchers DFs will ever fully disappear from a research field (nor that it would be desirable), meaning it will remain important for researchers to preregister their research when suitable, evaluate research for selective reporting, whenever feasible perform multiverse analysis (Steegen et al. 2016) and make their raw data available.

COMPETING INTERESTS

The authors have no financial or non-financial competing interests.

REFERENCES

Aert, Robbie C. M. van. 2020. *Puniform: Meta-analysis Methods Correcting for Publication Bias*. Manual. https://CRAN.R-project.org/package=puniform.

Aert, Robbie C. M. van, and Marcel A. L. M. van Assen. 2020. "Correcting for Publication Bias in a Meta-Analysis with the P-uniform* Method," August. https://doi.org/10.31222/osf.io/zqjr9.

Aert, Robbie C. M. van, Jelte M. Wicherts, and Marcel A. L. M. van Assen. 2016. "Conducting Meta-Analyses Based on *p* Values: Reservations and Recommendations for Applying *p* -Uniform and *p* -Curve." *Perspectives on Psychological Science* 11 (5): 713–29. https://doi.org/10.1177/1745691616650874.

Agnoli, Franca, Jelte M. Wicherts, Coosje L. S. Veldkamp, Paolo Albiero, and Roberto Cubelli. 2017. "Questionable Research Practices Among Italian Research Psychologists." *PLOS ONE* 12 (3): e0172792. https://doi.org/10.1371/journal.pone.0172792.

Alogna, V. K., M. K. Attaya, P. Aucoin, Š Bahník, S. Birch, A. R. Birt, B. H. Bornstein, et al. 2014. "Registered Replication Report: Schooler and Engstler-Schooler (1990)." *Perspectives on Psychological Science*, September. https://doi.org/10.1177/1745691614545653.

Appelbaum, Mark, Harris Cooper, Rex B. Kline, Evan Mayo-Wilson, Arthur M. Nezu, and Stephen M. Rao. 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, Marjan, Annette van Dijk, and Jelte M. Wicherts. 2012. "The Rules of the Game Called Psychological Science." *Perspectives on Psychological Science* 7 (6): 543–54. https://doi.org/10.1177/1745691612459060.

Bakker, Marjan, Coosje L. S. Veldkamp, Marcel A. L. M. van Assen, Elise A. V. Crompvoets, How Hwee Ong, Brian A. Nosek, Courtney K. Soderberg, David Mellor, and Jelte M. Wicherts. 2020. "Ensuring the Quality and Specificity of Preregistrations." *PLOS Biology* 18 (12): e3000937. https://doi.org/10.1371/journal.pbio.3000937.

Bakker, Marjan, and Jelte M. Wicherts. 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): 409–27. https://doi.org/10.1037/met0000014.

Borenstein, Michael. 2009. "Effect Sizes for Continuous Data." In *The Handbook of Research Synthesis and Meta-Analysis*, edited by H. Cooper, L. Hedges, and J. Valentine, 2nd ed., 221–35. New York: Russel Sage Foundation.

Botella, Juan, Manuel Suero, Juan I. Durán, and Desirée Blazquez. 2021. "The Small Impact of p-Hacking Marginally Significant Results on the Meta-Analytic Estimation of Effect Size." *Anales de Psicología* 37 (1): 178–87. https://doi.org/10.6018/analesps.433051.

Botvinik-Nezer, Rotem, Felix Holzmeister, Colin F. Camerer, Anna Dreber, Juergen Huber, Magnus Johannesson, Michael Kirchler, et al. 2020. "Variability in the Analysis of a Single Neuroimaging Dataset by Many Teams." *Nature*, May, 1–7. https://doi.org/10.1038/s41586-020-2314-9.

Bouwmeester, S., P. P. J. L. Verkoeijen, B. Aczel, F. Barbosa, L. Bègue, P. Brañas-Garza, T. G. H. Chmura, et al. 2017. "Registered Replication Report: Rand, Greene, and Nowak (2012)." *Perspectives on Psychological Science* 12 (3): 527–42. https://doi.org/10.1177/1745691617693624.

Breznau, Nate, Eike Mark Rinke, Alexander Wuttke, Muna Adem, Jule Adriaans, Amalia Alvarez-Benjumea, Henrik Kenneth Andersen, et al. 2021. "Observing Many Researchers Using the Same Data and Hypothesis Reveals a Hidden Universe of Uncertainty," March. https://doi.org/10.31222/osf.io/cd5j9.

Burton, A, and D G Altman. 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, Evan C., Felix D. Schönbrodt, Will M. Gervais, and Joseph Hilgard. 2019. "Correcting for Bias in Psychology: A Comparison of Meta-Analytic Methods:" *Advances in Methods and Practices in Psychological Science*, June. https://doi.org/10.1177/2515245919847196.

Chambers, Chris. 2019. "What's Next for Registered Reports?" *Nature* 573 (7773, 7773): 187–89. https://doi.org/10.1038/d41586-019-02674-6.

Cheung, I., L. Campbell, E. P. LeBel, R. A. Ackerman, B. Aykutoğlu, Š. Bahník, J. D. Bowen, et al. 2016. "Registered Replication Report: Study 1 From Finkel, Rusbult, Kumashiro, & Hannon (2002)." *Perspectives on Psychological Science* 11 (5): 750–64. https://doi.org/10.1177/1745691616664694.

Coburn, Kathleen M., and Jack L. Vevea. 2019. *Weightr: Estimating Weight-Function Models for Publication Bias*. Manual. https://CRAN.R-project.org/package=weightr.

Curran, Paul G. 2016. "Methods for the Detection of Carelessly Invalid Responses in Survey Data." *Journal of Experimental Social Psychology* 66 (September): 4–19. https://doi.org/10.1016/j.jesp.2015.07.006.

Davey, Jonathan, Rebecca M Turner, Mike J Clarke, and Julian PT Higgins. 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). https://doi.org/10.1186/1471-2288-11-160.

Eekhout, Iris, R. Michiel de Boer, Jos W. R. Twisk, Henrica C. W. de Vet, and Martijn W. Heymans. 2012. "Missing Data: A Systematic Review of How They Are Reported and Handled." *Epidemiology* 23 (5): 729–32. https://doi.org/10.1097/EDE.0b013e3182576cdb.

Eerland, A., A. M. Sherrill, J. P. Magliano, R. A. Zwaan, J. D. Arnal, P. Aucoin, S. A. Berger, et al. 2016. "Registered Replication Report: Hart & Albarracín (2011)." *Perspectives on Psychological Science*, January. https://doi.org/10.1177/1745691615605826.

Flake, Jessica K., Jolynn Pek, and Eric Hehman. 2017. "Construct Validation in Social and Personality Research: Current Practice and Recommendations." *Social Psychological and Personality Science* 8 (4): 370–78. https://doi.org/10.1177/1948550617693063.

Francis, Gregory. 2012. "The Psychology of Replication and Replication in Psychology." *Perspectives on Psychological Science* 7 (6): 585–94. https://doi.org/10.1177/1745691612459520.

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 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, Eiko I. 2020. "Lack of Theory Building and Testing Impedes Progress in The Factor and Network Literature." *Psychological Inquiry* 31 (4): 271–88. https://doi.org/10.1080/1047840X.2020.1853461.

Friese, Malte, and Julius Frankenbach. 2020. "P-Hacking and Publication Bias Interact to Distort Meta-Analytic Effect Size Estimates." *Psychological Methods* 25 (4): 456–71. https://doi.org/10.1037/met0000246.

Fusar-Poli, Paolo, Joaquim Radua, Marianna Frascarelli, Andrea Mechelli, Stefan Borgwardt, Fabio Di Fabio, Massimo Biondi, John P. A. Ioannidis, and Sean P. David. 2014. "Evidence of Reporting Biases in Voxel-Based Morphometry (VBM) Studies of Psychiatric and Neurological Disorders." *Human Brain Mapping* 35 (7): 3052–65. https://doi.org/10.1002/hbm.22384.

Ginkel, Joost R. van, Klaas Sijtsma, L. Andries van der Ark, and Jeroen K. Vermunt. 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.

Goldacre, Ben, Henry Drysdale, Aaron Dale, Ioan Milosevic, Eirion Slade, Philip Hartley, Cicely Marston, Anna Powell-Smith, Carl Heneghan, and Kamal R. Mahtani. 2019. "COMPare: A Prospective Cohort Study Correcting and Monitoring 58 Misreported Trials in Real Time." *Trials* 20 (1): 118. https://doi.org/10.1186/s13063-019-3173-2.

Hagger, M. S., N. L. D. Chatzisarantis, H. Alberts, C. O. Anggono, C. Batailler, A. R. Birt, R. Brand, et al. 2016. "A Multilab Preregistered Replication of the Ego-Depletion Effect." *Perspectives on Psychological Science* 11 (4): 546–73. https://doi.org/10.1177/1745691616652873.

Hartgerink, Chris H. J. 2017. "Reanalyzing Head Et Al. (2015): Investigating the Robustness of Widespread p-Hacking." *PeerJ* 5: e3068. https://doi.org/10.7717/peerj.3068.

Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2010. "The Weirdest People in the World?" *Behavioral and Brain Sciences* 33 (2-3): 61–83. https://doi.org/10.1017/S0140525X0999152X.

Henry, Lionel, and Hadley Wickham. 2020. *Purrr: Functional Programming Tools*. Manual. https://CRAN.R-project.org/package=purrr.

Houwelingen, Hans C. van, Lidia R. Arends, and Theo Stijnen. 2002. "Advanced Methods in Meta-Analysis: Multivariate Approach and Meta-Regression." *Statistics in Medicine* 21 (4): 589–624. https://doi.org/10.1002/sim.1040.

Huntington-Klein, Nick, Andreu Arenas, Emily Beam, Marco Bertoni, Jeffrey Bloem, Pralhad H. Burli, Naibin Chen, et al. 2020. "The Influence of Hidden Researcher Decisions in Applied Microeconomics," May. https://papers.ssrn.com/abstract=3602409.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *PLoS Medicine* 2 (8): e124. https://doi.org/10.1371/journal.pmed.0020124.

———. 2008. "Why Most Discovered True Associations Are Inflated." *Epidemiology* 19 (5): 640–48. https://doi.org/10.1097/EDE.0b013e31818131e7.

Jackson, Dan, Richard Riley, and Ian R. White. 2011. "Multivariate Meta-Analysis: Potential and Promise." *Statistics in Medicine* 30 (20): 2481–98. https://doi.org/10.1002/sim.4172.

Jakobsen, Janus Christian, Christian Gluud, Jørn Wetterslev, and Per Winkel. 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): 162. https://doi.org/10.1186/s12874-017-0442-1.

John, Leslie K., George Loewenstein, and Drazen Prelec. 2012. "Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling." *Psychological Science* 23 (5): 524–32.

https://doi.org/10.1177/0956797611430953.

Kirkham, J. J, K. M Dwan, D. G Altman, C. Gamble, S. Dodd, R. Smyth, and P. R Williamson. 2010. "The Impact of Outcome Reporting Bias in Randomised Controlled Trials on a Cohort of Systematic Reviews." *BMJ* 340 (February): c365–65. https://doi.org/10.1136/bmj.c365.

Kvarven, Amanda, Eirik Strømland, and Magnus Johannesson. 2019. "Comparing Meta-Analyses and Preregistered Multiple-Laboratory Replication Projects." *Nature Human Behaviour*, December. https://doi.org/10.1038/s41562-019-0787-z.

Kyzas, Panayiotis A., Konstantinos T. Loizou, and John P. A. Ioannidis. 2005. "Selective Reporting Biases in Cancer Prognostic Factor Studies." *JNCI: Journal of the National Cancer Institute* 97 (14): 1043–55. https://doi.org/10.1093/jnci/dji184.

Madden, Kim, Kerry Tai, Zak Ali, Patricia Schneider, Mahip Singh, Michelle Ghert, and Mohit Bhandari. 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, Randy J., John J. Skowronski, Bruno Verschuere, Ewout H. Meijer, Ariane Jim, Katherine Hoogesteyn, Robin Orthey, et al. 2018. "Registered Replication Report on Srull and Wyer (1979)." *Advances in Methods and Practices in Psychological Science* 1 (3): 321–36. https://doi.org/10.1177/2515245918777487.

McNeish, Daniel, and Melissa Gordon Wolf. 2020. "Thinking Twice about Sum Scores." *Behavior Research Methods* 52 (6): 2287–2305. https://doi.org/10.3758/s13428-020-01398-0.

Moher, David, Alessandro Liberati, Jennifer Tetzlaff, Douglas G. Altman, and The PRISMA Group. 2009. "Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement." *PLOS Medicine* 6 (7): e1000097. https://doi.org/10.1371/journal.pmed.1000097.

Moshontz, Hannah, Lorne Campbell, Charles R. Ebersole, Hans IJzerman, Heather L. Urry, Patrick S. Forscher, Jon E. Grahe, et al. 2018. "The Psychological Science Accelerator: Advancing Psychology Through a Distributed Collaborative Network." *Advances in Methods and Practices in Psychological Science* 1 (4): 501–15. https://doi.org/10.1177/2515245918797607.

Nickerson, Raymond 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, Michael, Leif D. Nelson, Evi Ackermann, Balazs Aczel, Athfah Akhtar, Silvio Aldrovandi, Nasseem Alshaif, et al. 2018. "Registered Replication Report: Dijksterhuis and van Knippenberg (1998):" *Perspectives on Psychological Science*, February. https://doi.org/10.1177/

1745691618755704.

Olsson-Collentine, Anton, Jelte M. Wicherts, and Marcel A. L. M. van Assen. 2020. "Heterogeneity in Direct Replications in Psychology and Its Association with Effect Size." *Psychological Bulletin*, July. https://doi.org/10.1037/bul0000294.

Oppenheimer, Daniel M., Tom Meyvis, and Nicolas Davidenko. 2009. "Instructional Manipulation Checks: Detecting Satisficing to Increase Statistical Power." *Journal of Experimental Social Psychology* 45 (4): 867–72. https://doi.org/10.1016/j.jesp.2009.03.009.

Page, Matthew J., Joanne E. McKenzie, and Julian P. T. Higgins. 2018. "Tools for Assessing Risk of Reporting Biases in Studies and Syntheses of Studies: A Systematic Review." *BMJ Open* 8 (3): e019703. https://doi.org/10.1136/bmjopen-2017-019703.

Page, Matthew J., Jonathan A. C. Sterne, Julian P. T. Higgins, and Matthias Egger. 2020. "Investigating and Dealing with Publication Bias and Other Reporting Biases in Meta-analyses of Health Research: A Review." *Research Synthesis Methods*, November, jrsm.1468. https://doi.org/10.1002/jrsm.1468.

Patel, Chirag J., Belinda Burford, and John P. A Ioannidis. 2015. "Assessment of Vibration of Effects Due to Model Specification Can Demonstrate the Instability of Observational Associations." *Journal of Clinical Epidemiology* 68 (9): 1046–58. https://doi.org/10.1016/j.jclinepi.2015.05.029.

Pham, Michel Tuan, and Travis Tae Oh. 2021. "Preregistration Is Neither Sufficient nor Necessary for Good Science." *Journal of Consumer Psychology* 31 (1): 163–76. https://doi.org/10.1002/jcpy.1209.

Pigott, Therese D., Jeffrey C. Valentine, Joshua R. Polanin, Ryan T. Williams, and Dericka D. Canada. 2013. "Outcome-Reporting Bias in Education Research." *Educational Researcher* 42 (8): 424–32. https://doi.org/10.3102/0013189X13507104.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Manual. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rankin, J., A. Ross, J. Baker, M. O'Brien, C. Scheckel, and M. Vassar. 2017. "Selective Outcome Reporting in Obesity Clinical Trials: A Cross-Sectional Review: Reporting Outcomes in Obesity Clinical Trials." *Clinical Obesity* 7 (4): 245–54. https://doi.org/10.1111/cob.12199.

Revelle, William. 2020. *Psych: Procedures for Psychological, Psychometric, and Personality Research*. Manual. Evanston, Illinois. https://CRAN.R-project.org/package=psych.

Robinson, David, Alex Hayes, and Simon Couch. 2020. *Broom: Convert Statistical Objects into Tidy Tibbles*. Manual. https://CRAN.R-project.org/package=broom.

Roese, Neal J., and Kathleen D. Vohs. 2012. "Hindsight Bias." *Perspectives on Psychological Science* 7 (5): 411–26. https://doi.org/10.1177/1745691612454303.

Rombach, Ines, Oliver Rivero-Arias, Alastair M. Gray, Crispin Jenkinson, and Órlaith Burke. 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): 1613–23. https://doi.org/10.1007/s11136-015-1206-1.

Rongen, Jan J., and Gerjon Hannink. 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): 403–9. https://doi.org/10.2106/JBJS.15.00400.

Rosati, Paola, Franz Porzsolt, Gabriella Ricciotti, Giuseppina Testa, Rita Inglese, Ferruccio Giustini, Ersilia Fiscarelli, et al. 2016. "Major Discrepancies Between What Clinical Trial Registries Record and Paediatric Randomised Controlled Trials Publish." *Trials* 17 (1): 430. https://doi.org/10.1186/s13063-016-1551-6.

Rousseeuw, Peter J., and Mia Hubert. 2011. "Robust Statistics for Outlier Detection." *WIREs Data Mining and Knowledge Discovery* 1 (1): 73–79. https://doi.org/10.1002/widm.2.

Sassenberg, Kai, and Lara Ditrich. 2019. "Research in Social Psychology Changed Between 2011 and 2016: Larger Sample Sizes, More Self-Report Measures, and More Online Studies." *Advances in Methods and Practices in Psychological Science*, April, 251524591983878. https://doi.org/10.1177/2515245919838781.

Schäfer, Thomas, and Marcus A. Schwarz. 2019. "The Meaningfulness of Effect Sizes in Psychological Research: Differences Between Sub-Disciplines and the Impact of Potential Biases." *Frontiers in Psychology* 10 (April). https://doi.org/10.3389/fpsyg.2019.00813.

Shaw, Mairead, Leonie J. R. Cloos, Raymond Luong, Sasha Elbaz, and Jessica Kay Flake. 2020. "Measurement Practices in Large-Scale Replications: Insights from Many Labs 2." *Canadian Psychology/Psychologie Canadienne* 61 (4): 289–98. https://doi.org/10.1037/cap0000220.

Silberzahn, R., E. L. Uhlmann, D. P. Martin, P. Anselmi, F. Aust, E. Awtrey, Š. Bahník, et al. 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): 337–56. https://doi.org/10.1177/2515245917747646.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22 (11): 1359–66. https://doi.org/10.1177/0956797611417632.

Simons, Daniel J., Yuichi Shoda, and D. Stephen Lindsay. 2017. "Constraints on Generality (COG): A Proposed Addition to All Empirical Papers." *Perspectives on Psychological Science* 12 (6): 1123–28. https://doi.org/10.1177/1745691617708630.

Stanley, T. D., and Hristos Doucouliagos. 2014. "Meta-Regression Approximations to Reduce Publication Selection Bias: T. D. STANLEY AND H. DOUCOULIAGOS." *Research Synthesis Methods* 5 (1): 60–78. https://doi.org/10.1002/jrsm.1095.

Steegen, Sara, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. 2016. "Increasing Transparency Through a Multiverse Analysis." *Perspectives on Psychological Science* 11 (5): 702–12. https://doi.org/10.1177/1745691616658637.

Stewart, Lesley A., and Jayne F. Tierney. 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., L. S. Fidell, and J. B. Ullman. 2007. *Using Multivariate Statistics*. Vol. 5. Boston: MA: Pearson.

Van Bavel, Jay J. 2016. "Contextual Sensitivity Helps Explain the Reproducibility Gap Between Social and Cognitive Psychology." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2820883.

Verschuere, Bruno, Ewout H. Meijer, Ariane Jim, Katherine Hoogesteyn, Robin Orthey, Randy J. McCarthy, John J. Skowronski, et al. 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, Jack L., and Larry V. Hedges. 1995. "A General Linear Model for Estimating Effect Size in the Presence of Publication Bias." *Psychometrika* 60 (3): 419–35. https://doi.org/10.1007/BF02294384.

Viechtbauer, Wolfgang. 2010. "Conducting Meta-Analyses in *R* with the **Metafor** Package." *Journal of Statistical Software* 36 (3). https://doi.org/10.18637/jss.v036.i03.

Wagenmakers, E.-J., T. Beek, L. Dijkhoff, Q. F. Gronau, A. Acosta, R. B. Adams, D. N. Albohn, et al. 2016. "Registered Replication Report: Strack, Martin, & Stepper (1988)." *Perspectives on Psychological Science* 11 (6): 917–28. https://doi.org/10.1177/1745691616674458.

Wayant, Cole, Caleb Scheckel, Chandler Hicks, Timothy Nissen, Linda Leduc, Mousumi Som, and Matt Vassar. 2017. "Evidence of Selective Reporting Bias in Hematology Journals: A Systematic Review." Edited by Lise Lotte Gluud. *PLOS ONE* 12 (6): 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, Jelte M., Coosje L. S. Veldkamp, Hilde E. M. Augusteijn, Marjan Bakker, Robbie C. M. van Aert, and Marcel A. L. M. van Assen. 2016. “Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking.” *Frontiers in Psychology* 7. https://doi.org/10.3389/fpsyg.2016.01832.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, and Jennifer Bryan. 2019. *Readxl: Read Excel Files*. Manual. https://CRAN.R-project.org/package=readxl.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2020. *Dplyr: A Grammar of Data Manipulation*. Manual. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, and Evan Miller. 2020. *Haven: Import and Export ’SPSS’, ’Stata’ and ’SAS’ Files*. Manual. https://CRAN.R-project.org/package=haven.

Zhu, Hao. 2019. *kableExtra: Construct Complex Table with ’Kable’ and Pipe Syntax*. Manual. https://CRAN.R-project.org/package=kableExtra.