Heterogeneity in direct replications in psychology and Its association with effect size

Olsson-Collentine, Anton; Wicherts, Jelte M.; van Assen, Marcel A.L.M.

Published in:
Psychological Bulletin

DOI:
10.1037/bul0000294

Publication date:
2020

Document Version
Peer reviewed version

Citation for published version (APA):
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

Heterogeneity in direct replications in psychology and its association with effect size

Anton Olsson-Collentine[1], Jelte M. Wicherts[1], & Marcel A.L.M. van Assen[1,2]

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

[2] Department of Sociology, Faculty of Social and Behavioural Sciences, Utrecht University, the Netherlands

Author note

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. Correspondence concerning this article should be addressed to Anton Olsson-Collentine, Warandelaan 2, 5037 AB Tilburg. E-mail: j.a.e.olssoncollentine@uvt.nl.

Abstract

We examined the evidence for heterogeneity (of effect sizes) when only minor changes to sample population and settings were made between studies and explored the association between heterogeneity and average effect size in a sample of 68 meta-analyses from thirteen pre-registered multi-lab direct replication projects in social and cognitive psychology. Amongst the many examined effects, examples include the Stroop effect, the “verbal overshadowing” effect, and various priming effects such as “anchoring” effects. We found limited heterogeneity; 48/68 (71%) meta-analyses had non-significant heterogeneity, and most (49/68; 72%) were most likely to have zero to small heterogeneity. Power to detect small heterogeneity (as defined by Higgins, 2003) was low for all projects (mean 43%), but good to excellent for medium and large heterogeneity. Our findings thus show little evidence of widespread heterogeneity in direct replication studies in social and cognitive psychology, suggesting that minor changes in sample population and settings are unlikely to affect research outcomes in these fields of psychology. We also found strong correlations between observed average effect sizes (standardized mean differences and log odds ratios) and heterogeneity in our sample. Our results suggest that heterogeneity and moderation of effects is unlikely for a zero average true effect size, but increasingly likely for larger average true effect size.

Keywords: heterogeneity, meta-analysis, direct replication, psychology, many labs

Word count: 208

Public Significance Statement

This paper suggests that for direct replications in social and cognitive psychology research, small variations in design (sample settings and population) are an unlikely explanation for differences in findings of studies. Differences in findings of direct replications are particularly unlikely if the overall effect is (close to) zero, whereas these differences are more likely if the overall effect is larger.

Heterogeneity in direct replications in psychology and its association with effect size

Empirical research is typically portrayed as proceeding in two stages. First, belief in the existence of an effect is established. Second, the effect's generalizability is examined by exploring its boundary conditions (Simons et al., 2017). In the first stage, inferential statistics (including testing of statistical hypotheses, confidence intervals, or Bayesian analyses) are used to minimize the risk that a discovery is due to sampling error. In the second stage, one may ask to what extent the effect depends on a particular choice of four contextual factors; the 1) sample population, 2) settings, 3) treatment variables and 4) measurement variables (e.g., Campbell & Stanley, 2015). This extent is often explored through replications of the original study that are either as similar as possible to the original (called 'direct' or 'exact' replications) or with some deliberate variation on conceptual factors (so-called 'conceptual' or 'indirect' replications; Zwaan et al., 2017), and once sufficient studies have accumulated through meta-analysis. In meta-analysis, the heterogeneity of an effect size (henceforth referred to as heterogeneity) is a measure of an effect's susceptibility to changes in these four factors. An effect strongly dependent on one or more of the four factors, unless controlled for, should exhibit high heterogeneity. In this paper we examine the heterogeneity in replication studies in psychology, focusing on direct replications, and explore a proposed relationship between effect size and heterogeneity.

The possibility of heterogeneity can create controversy in the interpretation of replication results. The proclamation of a 'failure' to replicate an effect (by the reader’s preferred definition) is sometimes taken to suggest that the original finding was merely a false positive, due to '*p*-hacking' (Simmons et al., 2011) or publication bias (Inzlicht et al., 2015). Unsurprisingly, some researchers take offense (e.g., Baumeister, 2016), interpreting such implications as attacks on their abilities as researchers. An alternative explanation for non-replication, often espoused by the original authors (e.g, IJzerman et al.,, 2015; Strack, 2016), is that the effect is more heterogeneous than (perhaps implicitly) claimed originally, meaning that the true effect varies across contextual factors as described earlier. From this perspective, non-replication implies (possibly previously unknown) predictors of effect size, so called 'hidden moderators' (Van Bavel, 2016), the discovery of which can be seen as an opportunity for theoretical advancement (Simons et al., 2017; McShane et al., 2019). To attenuate the risk of heated discussions on the (non)existence of an effect, original authors have been recommended to pre-specify the degree of heterogeneity that would make them lose interest in the effect (e.g., by declaring 'constraints on generality'; Simons et al., 2017).

It is commonly believed that heterogeneity is the norm in psychology. In support of this notion, recent large scale reviews of meta-analyses in psychology (Stanley et al., 2017; Van Erp, et al., 2017) report median heterogeneity levels that can best be described as 'large' (see the section ‘Quantifying heterogeneity’). In comparison, the median heterogeneity estimate in medicine (Ioannidis et al., , 2007) would be considered 'small' by the same standard. It may simply be that effects in psychology are more heterogeneous than those of medicine. However, meta-analyses in psychology also typically include more studies than those in medicine, and it could be that they tend to include studies from a much broader spectrum. That is, varying on more contextual factors (sample population, settings, treatment variables, measurement variables) or varying more on these four factors than what is typical in medicine. The median number of studies (effect sizes) per meta-analysis in the psychology sample of Van Erp et al. (2017) was 12, whereas in medicine it was only 3 (Davey et al., 2011). It is difficult to separate these explanations (intrinsically more heterogeneity, or psychology including studies from a broader spectrum?). To facilitate doing so, in this paper we focus on meta-analyses of only direct replications, which are exempt from the potential problem of including too disparate studies. Our sample consists of all pre-registered multi-lab direct replication projects in psychology available on *curatescience.org* up until *2019-10-25*. By only including pre-registered multi-lab studies we also avoid the issue of publication bias, which can have a large and unpredictable effect on the assessment of heterogeneity (Augusteijn et al., 2018), as well as on the assessment of the effect size itself (e.g., Dickersin, 2005; Simes 1986).

Heterogeneity is often considered a primary outcome in meta-analysis for good reasons. As described above, unaccounted for heterogeneity suggests that a theory is unable to predict all contextual factors of importance to its claims and its existence affects the interpretation of replication outcomes. Moreover, unaccounted for heterogeneity can have practical consequences not to be ignored. This is readily evident for medicine, where in the case of heterogeneity an intervention, such as a medication, that is successful for some may have direct negative health consequences for others. The same is true of mental health interventions in psychology. Heterogeneity can also have major consequences for topics such as child development, education, and business performance, where research often impacts policy recommendations. A newly implemented policy to, say, help socialize children (e.g., in a day care), improve learning outcomes in education or employee satisfaction in business, which works only in some contexts or for some individuals and not others (i.e., is heterogeneous) could have an overall null or even negative impact instead of positive. Awareness of heterogeneity thus affects the cost-benefit analysis of whether to implement a particular policy. In other words, heterogeneity should be no less of a concern for psychologists than for medical practitioners.

Heterogeneity also affects meta-analytic techniques used to statistically summarize findings on a certain topic. Heterogeneity alters the interpretation of meta-analytic estimates as either *the* true effect size (under homogeneity) or the average of the true effect sizes (under heterogeneity), though one may question the usefulness of interpreting the average true effect size in the presence of heterogeneity (Simonsohn, 2017), just as it may be questionable to interpret an average main effect in the context of an interaction effect (Aiken et al., 1991). In addition, techniques that attempt to correct for publication bias in their estimate tend to fail in the presence of heterogeneity (Carter et al., 2019; McShane et al., 2016; Stanley, 2017; van Aert, 2018; van Aert et al., 2016; van Assen et al., 2015), which is problematic if we believe publication bias is widespread in psychology (Cooper et al., 1997; Franco et al., 2014, 2016; although see Stanley et al. 2018 and van Aert et al., 2019 for opposing conclusions) . To conclude, heterogeneity or its absence provides vital information for the implementation of research in practice, the advancement of theory, and the interpretation of research outcomes.

Quantifying Heterogeneity

Assessing heterogeneity can be problematic due to its inherent uncertainty. Heterogeneity is often measured by the I^2 index (Higgins, 2003; Higgins & Thompson, 2002). It can be interpreted as the percentage of variability in observed effect sizes in a meta-analysis that is due to heterogeneity amongst the true effect sizes (that is, sensitivity to contextual factors) rather than sampling variance, and ranges from 0-100%. More formally, $I^2 = \hat{\tau}^2 \,/\, (\hat{\tau}^2 + \hat{\sigma}^2)$, where $\hat{\tau}^2$ is the estimated between-studies variance and $\hat{\sigma}^2$ is an estimate of the 'typical' within-studies variance, and I^2 is set to zero if negative. An alternative but related index of heterogeneity is H^2 (Higgins & Thompson, 2002), with $H^2 = 1/(1- I^2)$ or (for the DerSimonian-Laird estimator) $H^2 = Q/(K\text{-}1)$. As opposed to I^2, H^2 is not truncated (when $Q < K – 1$), and H^2 ranges from zero to infinity, with higher values signaling more heterogeneity, with a value of 1 indicating homogeneity.

The I^2 index has several advantages when using it for meta-research as in our paper. First, it has an easy and intuitive interpretation as it is between 0 and 100%. Second, well-known rules of thumb (Higgins, 2003) exist to interpret values of I^2 as small (25%), medium (50%), or large (75%). As with all rules of thumb these should be used with caution. We do not use these labels normatively, but just as examples of “small”, “medium” and “large” heterogeneity. Third, I^2 can be computed for any effect size metric (correlations, standardized mean differences, odds ratios, etc.), without having to transform effect sizes to a specific metric. And finally, most large meta-meta analyses also employ I^2, which allows for comparing results of different meta-meta analyses. Two well-known examples of such large scale meta-meta analyses are Ioannidis et al., (2007) in medicine, and Van Erp et al., (2017) in psychology. Because of these advantages we employ I^2 (and its relative H^2) as one of our heterogeneity indices in our paper.

However, I^2 also has two important disadvantages. First, I^2 is not an absolute but a relative measure of heterogeneity, as it is dependent on the primary studies' sample sizes (Borenstein et al., 2017; Rücker et al., 2008). For instance, keeping constant $\hat{\tau}^2$, multiplying all primary studies' sample sizes with 3 will increase I^2 from small to medium (25% to 50%) or medium to large (50% to 75%), and multiplying with 9 will turn a small I^2 into a large I^2. Note that this characteristic of I^2 also implies that values 25, 50, 75% cannot be normatively used as labels for small, medium, large heterogeneity, respectively. Second, even though heterogeneity of all different effect sizes (correlations, standardized mean differences, odds ratios) are placed on the same I^2 scale, one can argue that I^2 values originating from different effect size metrics cannot be directly compared as they are based on different distributions and assumptions. Hence, these two disadvantages also call for another assessment of effect size heterogeneity, and estimators of τ seem to be the most promising alternatives, although τ estimates also cannot be compared across effect size types.

The Pearson's correlation and their Fisher-transformed counterparts could be a viable alternative for a common effect size metric. It is possible to transform effect sizes such as mean differences to point-biserial correlations, which are simply Pearson's correlations as applied to dichotomous data (see e.g., Borenstein, 2009; Schmidt & Hunter, 2015). However, there are potential concerns with transforming effect sizes to either Pearson or Fisher-transformed correlations. Our analyses revealed two undesirable characteristics of (transformed to) point-biserial correlations, making them inappropriate for answering our main research questions on heterogeneity of effect size in multi-lab direct replications and its association to average effect size. First, values of $\hat{\tau}$ are restricted for larger values of average effect size as the point-biserial correlation gets closer to 1, implying a possible unwanted negative association between average effect size and effect size heterogeneity. Second, while transforming from one metric to the point-biserial correlation, strong assumptions need to be made. Both undesirable characteristics may lead to serious distortions of heterogeneity assessment. For instance, if $\hat{\tau}_A > \hat{\tau}_B$ for two meta-analyses A and B based on exactly the same sample sizes and assessed on the same metric (e.g., standardized mean differences), then after transforming to point-biserial correlations the order of heterogeneity assessments may be reversed (see Supplement A for an illustration using our data). This issue is alleviated by using the Fisher-transformation, although violations of non-monotonicity may still be observed (see Supplement A). These findings suggest that researchers should carefully consider whether it is advisable to combine or transform effect sizes from different effect size metrics in a meta-analysis.

Another alternative estimator of heterogeneity is using τ based on the original effect size metrics. Estimates of τ can then not be compared across meta-analyses based on different metrics, but can be straightforwardly compared across meta-analyses based on the same metric, without having the disadvantages detailed above (negative association between average effect size and heterogeneity, strong assumptions) of using a common effect size metric. Hence, in addition to I^2 we also report results of τ based on the original metric. The consequence for our analysis on the association between heterogeneity and average effect sizes is that we only estimate this association for standardized mean difference and log odds ratios, since other effect size types (correlations, Cohen's *q*) were rare in our dataset (see Methods section).

Uncertainty and Statistical Power of Heterogeneity Assessment

Tests of heterogeneity typically have low statistical power in many practical situations (Huedo-Medina et al., 2006; Jackson, 2006). This complicates the discussion of heterogeneity, because while I^2 always provides an estimate of heterogeneity, this estimate is often accompanied by high uncertainty and by wide confidence intervals (Ioannidis et al., 2007). For example, Ioannidis reports that in a large set of Cochrane meta-analyses, all meta-analyses with I^2 point estimates of 0% had upper 95% confidence intervals that exceeded I^2 estimates of 33%, exceeding what Higgins (2003) defined as 'small' heterogeneity. In addition, under homogeneity the *Q*-statistic has a central chi-square distribution (von Hippel, 2015), a distribution that is right-skewed with 40-50% of observations falling above the expected value (for all $k > 3$). As point estimates of both τ and

I^2 are related (this relation is one-to-one, i.e., $Q > df$ implies $\tau^2 > 0$ and $I^2 > 0$), a meta-analysis of 4 or more studies will also have close to 50% of estimates exceeding 0, even in the absence of true heterogeneity.

To simplify interpretation of estimates of τ and I^2, we will report both these estimates as well as their confidence intervals, and report the results of power analyses of the Q-test of heterogeneity assuming zero/small/medium/large heterogeneity (here defined as I^2 = 0/25/50/75%, respectively). Conducting power analyses is necessary as a high frequency of zero estimates of τ and I^2 as well as a high frequency of confidence intervals including 0 can be the result of, for instance, either (i) a high frequency of true homogeneity, or (ii) a high frequency of true heterogeneity but combined with low statistical power. We need to be able to distinguish between these two cases. The power analyses additionally provide information to researchers on how many labs and participants may be needed to assess certain heterogeneity, based on real data rather than only simulations (Huedo-Medina et al., 2006; Jackson, 2006) and in a context highly similar to that of future multi-lab projects (e.g., Registered Replication Reports, ManyBabies, the Psych Science Accelerator).

Association Between Effect Size and Heterogeneity

Effect size is likely associated with heterogeneity. Intuitively, it makes sense to believe that if the meta-analytic effect size is zero there is nothing to moderate (i.e., no heterogeneity). However, a null or near null (average) effect size estimate may arise from failure to consider contextual factors ('hidden moderators'; Van Bavel, 2016) and does not by itself imply the absence of heterogeneity. To the contrary, a large meta-analytic effect size can be expected to be associated with more heterogeneity. To explain further, consider first the definition of heterogeneity.

Heterogeneity is defined as the standard deviation of the true (study-level) effect sizes. True effect size, however, may refer to two possibly very different entities. First, it may refer to the effect size that is obtained for a study with an infinite sample size (or having the complete population of subjects in one’s study; in any case, sampling error equals 0).

Second, the true effect size of a single study may refer to an effect size obtained with an infinite sample size but that is also corrected for unreliability of the measurements. We need to distinguish both entities when assessing and interpreting the average true effect size and true effect size heterogeneity, and their estimates.

Estimates of effect size heterogeneity always attempt to 'partial out' sampling error. Whether heterogeneity estimates also partial out measurement error depends on whether effect sizes were corrected for unreliability beforehand (in which case standard errors must also be corrected; Schmidt & Hunter, 2015, p. 314-320). Typically measurement error is not corrected for when estimating individual study effect sizes in a meta-analysis (although the field of Industrial-Organizational psychology is an exception to this rule). None of the thirteen multi-lab projects did so in any of the 68 meta-analyses. We therefore also do not attempt to correct for measurement error when estimating average effect size and effect size heterogeneity, and true (study-level) effect sizes in our paper refer to the effect size obtained by that study if sample size were infinite (i.e., the first entity). Below we illustrate how measurement error may result in a positive association between effect size as thus conceptualized, and heterogeneity.

To illustrate, consider a meta-analysis of say, the correlation between neuroticism and procrastination (e.g., Steel, 2007). Each included study would need to measure the two variables somehow, possibly the same way across studies in the meta-analysis. However, because of individual differences and differences in study samples, measurement reliabilities may differ across studies either due to sampling variance (that the sample happens to be more or less homogeneous) or to differences in contextual factors (e.g., sampling population, measurement variables). This means that even if the underlying true effect size (after correcting for measurement error; second entity above) is the same, the correlation between the two variables will differ between studies (see also Schmidt & Hunter, 2015). Assuming no association exists between reliability and true effect size (second entity above), differences in observed study effect sizes will increase with the underlying true effect size, resulting in more variability being ascribed to heterogeneity. More formally, an observed correlation r_{xy} can be expressed as the product of the true correlation or effect size (second entity), ρ_{xy}, multiplied by the square root of the measurement reliabilities for X ($R_{xx\prime}$) and Y ($R_{yy\prime}$): $r_{xy} = \rho_{xy} \times \sqrt{R_{xx\prime}} \times \sqrt{R_{yy\prime}}$. As such, keeping constant study differences in $\sqrt{R_{xx\prime}} \times \sqrt{R_{yy\prime}}$ while increasing true effect size ρ_{xy} (second entity) increases heterogeneity of effect sizes (first entity). Table 1 illustrates this relationship using three values of r_{xy} and true study-level effect sizes. We therefore explore with a correlational analysis if a positive association exists between effect size and heterogeneity in the sample of pre-registered multi-lab replication projects in psychology.

Table 1.

Variation in observed effect sizes as a function of true effect size and measurement reliability.

		Observed Effect Sizes			
		Study 1	**Study 2**	**Study 3**	
Meta-Analysis	ρ_{xy}	$\sqrt{R_{xx'}} \times \sqrt{R_{yy\prime}}$. **= .60**	$\sqrt{R_{xx'}} \times \sqrt{R_{yy\prime}}$. **= .70**	$\sqrt{R_{xx'}} \times \sqrt{R_{yy\prime}}$. **= .80**	***SD* (*ES*)**
I	0.00	0.00	0.00	0.00	0.00
II	0.30	0.18	0.21	0.24	0.03
III	0.50	0.30	0.35	0.40	0.05

Note. The values under Study 1, 2 and 3 are observed effect sizes for that study given its measurement reliability $\sqrt{R_{xx'}} \times \sqrt{R_{yy\prime}}$ and the true effect size ρ_{xy} when within-study sample size is infinite. *SD* (*ES*) is the standard deviation of the observed effect sizes for meta-analysis I, II and III, equivalent to heterogeneity given infinite within-study sample sizes. Code to reproduce table: osf.io/gtfjn

The Pre-registered Multi-lab Replication Projects

Table 2 lists the thirteen replication projects, with a total of 68 primary outcome variables, we used to examine heterogeneity and the correlation between effect size and heterogeneity in psychology. These “Many Labs” and “Registered Replication Report” (RRR) projects are a recent phenomenon in psychological science where multiple labs collaborate to replicate one or multiple effects from the psychological literature. Fundamental to these projects is that they are pre-registered and that each collaborating lab uses the exact same materials (possibly with language translations), so that essentially the only difference between participating labs is that they run the study in different locations and using different people. This also means that heterogeneity estimates based on these data only reflect this type of variation in sample population and settings. The projects are often done to examine the robustness of seminal findings with high impact and typically in discussion with the original authors. The principal difference between Many Labs and RRR projects is that the Many Labs include multiple distinct psychological effects (all run in one session), whereas the RRRs focus only on one effect. That we report multiple effects for three of the RRRs in Table 2 is because they used multiple primary outcome variables.

We would consider most of the effects described in Table 2 to belong to social and cognitive psychology (and Many Labs 2 explicitly selected effects from these domains). As an example, RRR8 (O’Donnell et al., 2017) replicated an experiment examining the link between priming of social categories (soccer hooligan/professor) and objective knowledge performance (a trivia quiz). Priming can be viewed as the idea that brief (often subconscious) exposure to a concept should activate related concepts or behavior. The experiment replicated by the RRR8 authors has been cited more than 800 times, and the manipulation (“professor priming”) is well known in the field of social psychology (O’Donnell et al., 2017). However, O’Donnel et al. report that when RRR8 was organized there had been increasing debate over the validity of priming effects in the past years, including of the “professor priming” effect. RRR8 was set up in response to this controversy. Many of the studied effects (as in the case of O’Donnell et al.) used priming (23 effects) in their design. Others asked participants to imagine different situations (14 effects) or to react to slightly different statements (six effects), with the remainder using a variety of approaches (see Table 2). We present only a succinct summary of the studied results and direct readers to the original multi-lab publications for more detailed descriptions.

In reference to meta-analyses of direct replications, McShane with several co-authors (McShane et al., 2016; 2019) have argued that if we were to expect heterogeneity to be absent or minimal anywhere, it would be in pre-registered multi-lab projects with a common protocol (such as Klein et al., 2014). They further argue that the fact that heterogeneity has been reported even under such circumstances is an indication of widespread heterogeneity in psychology, although McShane (personal communication, July 19, 2019) acknowledges that expected heterogeneity in multi-lab replication projects is much smaller than in large scale meta-analyses in psychology. However, In the case of multi-lab direct replication projects, studies still vary on two contextual factors (sample population and settings) and if we believe an effect is sensitive to changes in these two factors we might also expect to find some heterogeneity.

As all thirteen projects in our dataset were (relatively) large-scale and pre-registered, our dataset arguably represents the best meta-analytic data currently available in psychology. To better interpret the heterogeneity estimates we also estimate power of each project to find zero/small/medium/large heterogeneity by the definitions of Higgins (2003). Consequently, our analyses will provide information on how two contextual factors (sample population and settings) may affect consistency or heterogeneity of effects in psychology, and on the precision of its estimate.

Table 2.

Pre-registered multi-lab replication projects

RP	Paper	Countries	*K* (US)	Effects	*N*	Sample and Settings	Description of Effects
ML1	Klein et al. (2014)	10	36 (25)	16	5975	26/36 samples were primarily university students, 3 general population and 7 undescribed. 9/36 samples were online, including all the general population ones.	Two correlational effects: 'Gender math attitude' compared implicit attitudes (IAT) towards math between genders and 'IAT correlation math' correlated implicit attitudes with self-reported measures. The remainder were experiments with two independent groups. The groups were primed in some way (Anchoring 1-4; low vs. high category scales; norm of reciprocity; flag priming; currency priming), asked to imagine slightly different situations (Sunk costs; gain vs. loss framing; gambler's fallacy; imagined contact) or asked their agreement with statements presented differently (Allowed vs. forbidden; quote attribution).
ML2	Klein et al. (2018)	35	115 (21)	28	6570	79/125 samples were collected in person (typically in labs), remainder online. Mean age in two rounds of data collection were 22.67 and 23.34 years.	Most effects were experiments with two independent groups. Often participants were primed in some way (Structure & Goal Pursuit, Priming Consumerism, Incidental Anchors, Position & Power, Moral Cleansing, Priming Warmth) or asked to imagine slightly different situations (SMS & Well-Being, Less is Better, Moral Typecasting, Intentional Side-Effects, Tempting Fate, Affect & Risk, Trolley Dilemma 1, Framing, Trolley Dilemma 2, Disgust & Homophobia, Choosing or Rejecting). Some groups saw slightly different statements (Correspondence Bias, Intuitive Reasoning), were asked to perform slightly different tasks (Direction & SES, Actions are Choices), or had to read a text with a clear vs. unclear font (Incidental Disfluency). Two correlational effects measured the correlations of Moral Foundations with political leaning, and Social Value Orientation with family size. Two effects examined order effects (Assimilation & Contrasts, Direction & Similarity). Finally, in False Consensus 1 and 2, participants made a binary choice and estimated how many people had made the same choice.
ML3	Ebersole et al. (2016)	2	21 (19)	10	2845	20/21 samples were university students, 1 general population which was also the only online sample.	Several effects were experiments with two independent groups. The groups were either primed in some way (Power and perspective; warmth perceptions; subjective distance interaction), saw slightly different statements (Elaboration likelihood interaction; credentials interaction) or experienced different situations (weight embodiment). Examined interactions were between treatment conditions and participant characteristics. One priming effect (metaphor) compared two treatment groups with a control. One effect was correlational: 'conscientiousness and persistence' was measured by an unsolvable anagram task and self-report respectively. The Stroop task is a within-person experiment with two conditions and the 'Availability' effect asks participants to judge whether some letters are more common in the first or third position.
RRR1	Alogna et al. (2014)	10	32 (17)	1	4117	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	23 (14)	1	2442	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.

Table 2 continued

RP	**Paper**	**Countries**	***K* (US)**	**Effects**	***N***	**Sample and Settings**	**Description of Effects**
RRR3	Eerland et al. (2016)	2	12 (10)	3	1187	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	23 (7)	1	2872	All samples consisted of in-lab undergraduate students	Ego depletion; Independent two-group experiment. Participants either assigned to a cognitively demanding task or a neutral, and performance was then measured in a subsequent cognitive task.
RRR5	Cheung et al. (2016)	5	16 (9)	2	2071	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 (8)	1	1894	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 (5)	1	3596	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	23 (9)	1	4493	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	22 (4)	2	5610	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	19 (4)	1	2294	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’d read. Outcome was degree of cheating when reporting results.

*Note***.** For studies with several effects the number of participants is the average across effects, rounded to the closest whole number. *N* = Participants used for primary analyses by original authors (i.e., after exclusions). RP = Replication Project, *K* (US) = no. primary studies (number of US studies), ML = Many Labs, RRR = Registered Replication Report. Code to reproduce table: osf.io/gtfjn

Method

All code and data for this project are available on the Open Science Framework (OSF) at osf.io/4z3e7. We refer directly to relevant files on the OSF using brackets and links in the sections below. We ran all analyses using R version 3.4.3 (R Core Team, 2017).

Data Collection

For the purposes of this project (as described in the introduction) we collected meta-analyses of only pre-registered direct replications in psychology. We defined a meta-analysis of “direct” replications as a meta-analysis of a set of studies with no differences in treatment or measurement variables. This type of multi-lab studies have only recently become popular in psychology, and as typically large collaborations on well-known and/or highly debated topics (see section ‘The pre-registered multi-lab projects’) each publication garners wide attention. We set out to include all such pre-registered multi-lab projects in psychology with published data. To decrease the risk of missing any published projects we made use of the webpage curatescience.org. Curatescience.org is a crowdsourced project to keep track of replications and transparency of research and so well-attuned to the purpose of finding replication studies with available data. In addition it includes a section with a “curated list of large scale replication efforts” which was intended to be “as comprehensive and inclusive as possible” (LeBel, personal communication, November 12, 2019). We included all multi-lab projects from this list. Originally, we included projects published before 2018-03-31, but updated our dataset in the process of revision with 3 additional projects that were published between 2018-03-31 and 2019-10-25, for a total of 13 projects containing 68 meta-analyses of primary effects.

We downloaded and collated summary data from the thirteen pre-registered multi-lab replication projects in psychology (Table 2). Data from all thirteen projects were available on the Open Science Framework (osf.io) and downloaded between 2018/02/01 and 2019-10-25.

Although some projects (e.g. RRR4) reported results from several outcome variables, we only included primary outcome variables as explicitly stated in accompanying publications, resulting in a total of 68 meta-analyses. For each meta-analysis we extracted (osf.io/mcj5d) summary data (e.g., means and standard deviations) at the level of the lab as specified by the original authors for their primary analysis (i.e., typically after exclusions). We extracted information on the country of each lab, whether participants were physically in the lab for the study, total number of participants per lab, type of effect size, and additional information related to each meta-analysis (see codebook; osf.io/yn9fb). Extracted data were in a variety of formats: Excel (Many labs 1, RRR1 & RRR2), CSV (Many labs 3, RRR3, RRR4, RRR5, RRR6), and as PDF tables (RRR7). In three cases (RRR5, RRR6, and RRR9) it was necessary to download the raw data to extract summary data. In two cases (RRR8 and RRR10) there was summary data available as a CSV file, but without all the information we needed. For these, it was necessary to download the raw data and make minor code edits to extract the standard deviations. Although a particular lab may have participated in several projects, the lab indicator was typically not the same across projects. Even so, we kept the original lab indicators to facilitate comparing observations in our dataset with the original datasets. Finally, we collated the summary data for all meta-analyses into one dataset for analysis (osf.io/mcj5d)

Heterogeneity Across Meta-analyses

To examine heterogeneity of each of the 68 effects, we computed meta-analytic estimates for all 68 effects in our dataset (Table 3). In our primary analysis we ran all analyses as specified by the replication authors (osf.io/q9vwb). In contrast, since the replication authors sometimes transformed effect sizes (e.g., odds ratios to standardized mean differences; ML1) in our analysis of the association between heterogeneity and average effect size we did not always follow the replication authors specifications (see section “Association between effect size and heterogeneity”). Here we describe how effect sizes in Table 3 were estimated. The effect size of the original study, which was the focus of the replication effort, was not included in these meta-analyses. All meta-analyses were estimated with random-effects models and the Restricted Maximum Likelihood (REML) estimator using the R-package metafor (Viechtbauer, 2010), though with a variety of outcome variables: product moment correlations (*r*), differences in correlations (Cohen's *q*), standardized mean differences (*SMD*), raw mean differences (*MD*), and risk differences (*RD*). Many Labs 1 transformed effect sizes measured as odd ratios into standardized mean differences when meta-analyzing under the assumption that responses followed logistic distributions (Sánchez-Meca et al., 2003; Viechtbauer, 2010). Two projects (RRR5 and RRR7) used the Knapp and Hartung adjustment of the standard errors (Knapp & Hartung, 2003) and Many Labs 3 correlations were corrected for bias (Hedges, 1989; Viechtbauer, 2010). Many Labs 3 meta-analyzed (see osf.io/yhdau) several effects that were not originally measured as correlations (Availability, Metaphor; Stroop effect, Elaboration likelihood interaction, Subjective distance interaction, Credentials interaction) but were nonetheless transformed to and analyzed as product-moment correlations. It is not clear from the Many Labs 3 documents how they transformed the dichotomous (Availability, Metaphor) or within-person (Stroop effect) outcomes to product-moment correlations and their standard errors. Interaction effect sizes appear to have been transformed from the original partial η^2 by taking the square root. Many Labs 2 transformed all effect sizes, except two measured as Cohen's *q*, into product-moment correlations for analysis by computing the non-central confidence intervals for each test statistic and then transforming these into product-moment correlations using the R-package "compute.es" (Hasselman, personal communication, October 14, 2019).

In each meta-analysis we estimated τ, I^2 and their 95% confidence intervals. The R-package metafor uses a general expression for I^2 (equation 9 in Higgins & Thompson 2002) and estimates its confidence interval using the *Q*-profile method (Jackson, Turner, Rhodes, & Viechtbauer, 2014). We used this information together with our power analyses (described below) to examine the extent of heterogeneity across meta-analyses.

Simulation of Type I Error Rate and Power.

In order to facilitate interpretation of our results, we estimated type I error and power of the *Q*-test of heterogeneity (Cochran, 1954) for each of the 68 meta-analyses under zero/small/medium/large heterogeneity (I^2 = 0/25/50/75% respectively). In addition, we approximated the probability density function of I^2 across meta-analyses at each of these four heterogeneity levels and compared them with the observed frequency distribution of the observed I^2 (respectively $\hat{\tau}$) estimates of the 68 meta-analyses. Hence, five distributions of I^2 were obtained; four simulated and one observed. To do so we simulated results of I^2 for each meta-analysis given its number of studies (*K*), sample sizes of those studies (vector N_k), and each of the four heterogeneity levels (osf.io/mw4aq). We directly simulated the distribution of I^2 for correlation, Cohen's *q*, standardized mean difference, and mean difference effect size measures, but not for risk differences. We treated risk differences as mean differences using the study sample sizes to compute study precision, because treating them as risk differences would require strong assumptions on the probability of success in both treatment groups, assumptions which would greatly affect the outcomes of the simulation. For the same reason we treated the four effects of Many Labs 1 which were measured as odds ratios (and then transformed into standardized mean differences) as standardized mean differences. Many labs 2 and Many Labs 3 effects which were reported as correlations were treated as such.

As our concern was heterogeneity, for convenience we set the average true effect size to zero in our simulations of heterogeneity. This should not affect the results for correlations or mean differences, as estimates of effect size and heterogeneity for these measures are unrelated (i.e., changing the value of one estimate does not directly affect the formula and value of the other estimate). For standardized mean differences we expect negligible effects on the results, because while these estimates of effect size are positively correlated to their standard errors, the within-study variance σ^2 was kept constant across studies. As a sensitivity analysis we also ran all I^2 analyses assuming 'medium' effect sizes (Cohen, 1988) and indeed found the same average power at the different heterogeneity levels, see Supplement B.

In case the observed effect size was a correlation, one run of a simulation proceeded as follows. First, we randomly sampled *K* true correlations ρ_i from a normal distribution with mean 0 and heterogeneity (standard deviation) τ. Second, for each of the *K* true correlations we sampled one Fisher-transformed (Fisher, 1915; 1921) observed correlation from a normal distribution with mean ρ_i and variance $1/(N_i - 3)$. Finally, we fitted a random-effects meta-analysis with REML and estimated I^2 for that run. In the simulations, we varied the between-studies standard deviation τ between 0.000 and 0.50 in increments of 0.005, and used 1,000 runs at each step to approximate the distribution of I^2 at that value for true heterogeneity. For Cohen's *q*, we proceeded identically, except that variance was computed as $1/(n_t - 3)$ + $1/(n_c - 3)$ where n_t and n_c were the observed treatment and control sample sizes for each study.

For mean differences (and hence also for risk differences) we assumed a within-study variance of one for both treatment and control groups, $\sigma_c^2 = \sigma_t^2 = 1$. For each run we then set the population mean of the control condition to 0 and sampled *K* treatment population means μ_k from $N(0, \tau)$. Subsequently, K sample means for both control and treatment conditions were sampled, with $\overline{x}_c \sim N(0{,}1/\sqrt{n_c})$ and $\overline{x}_t \sim N(\mu_k, 1/\sqrt{n_t})$.. Group variances were sampled using $s_c^2 \sim \chi^2(n_c - 1)/(n_c - 1)$ and $s_t^2 \sim \chi^2(n_t - 1)/(n_t - 1)$. Finally, we fitted a random-effects meta-analysis with REML and estimated I^2 for that run. For standardized mean differences (and odds ratios) we proceeded identically, except that in the final step we asked metafor to transform the effect size into a standardized mean difference (Hedge's g) in fitting the random-effects model. As with correlations, the distribution of I^2 was approximated for values of τ from 0 to .5 in steps of .005, using 1,000 runs at each step.

To approximate the statistical power of all 68 meta-analyses at zero, small, medium, and large heterogeneity we continued as follows. For each of the 68 meta-analyses we selected the values of τ which yielded the average value of I^2 in the simulations closest to 25% (small), 50% (medium), and 75% (large). For these values of τ and for $\tau = 0$ (homogeneity) we ran a simulation with 10,000 runs, and for each run I^2 was calculated and the *Q*-test of heterogeneity was performed, yielding estimates of type I error (in case of homogeneity) and power (for heterogeneity) for each of the 68 meta-analyses. We considered a result significant when $p \leq 0.05$ for the *Q*-test. The distributions of I^2 for zero, small, medium, large heterogeneity, which we compared to the observed distribution of 68 effect sizes, was generated by pooling the 68 distributions of 10,000 I^2 values in each category of heterogeneity. Hence these I^2 distributions can be considered a mixture distribution of 68 distributions, using equal weights across all 68 meta-analyses.

Association Between Effect Size and Heterogeneity

We examined the association between average (meta-analytic) effect size and $\hat{\tau}$, I^2, and the closely related H^2, for effect sizes on the log odds ratio metric (10 effects) and the standardized mean differences metric (Hedges g; 43 effects). We avoided transforming effect sizes for this analysis because transforming effect sizes will distort this association (see Supplement A). Hence we only used effect sizes that were originally measured as mean differences or binary outcomes with two groups (risk differences, odds ratios). There were too few product moment correlation effect sizes (4) and differences in correlations (2) to warrant estimating a correlation to these effect types. Many Labs 3 reported correlations, which they treated as product-moment correlations, as summary statistics for several effects (Availability, Metaphor; Stroop effect; interactions), which were not originally measured in this metric. These effect sizes were excluded from the analysis, as were three effects from Many Labs 2 for the same reason (Choosing or Rejecting; Direction & Similarity; Actions are Choices). The four effects (Allowed vs. forbidden, Gain vs. loss framing, Norm of reciprocity, Low vs. high category scales) that were transformed by Many Labs 1 into standardized mean differences we computed as (log) odds ratios.

In our analyses we computed the association of estimates of average effect size with three different heterogeneity estimates: $\hat{\tau}$, I^2 and the closely related H^2 (Higgins & Thompson, 2002). All estimates were obtained with the REML estimator in metafor. We added the H^2 index as a robustness check to avoid the truncation at zero of the I^2 index when computing correlations between estimates of effect size and effect size heterogeneity. However, to avoid truncation we had to compute H^2 as $H^2 = Q / (k - 1)$. This expression of H^2 is strictly only correct when using the DerSimonian-Laird estimator of $\hat{\tau}$, and readers should be aware of this when interpreting the results of H^2. To describe the association between average effect size and heterogeneity due to variation in sample population and settings, we report both Pearson's product moment correlations and, as the association may be nonlinear, Spearman's rank order correlations. For these statistics we also report 95% bootstrap confidence intervals using the percentile method (osf.io/u2t3r).

Results

Table 3 presents the meta-analytic effect size estimates of τ and I^2 with confidence intervals for each of the 68 included effects, as well as simulated type I error and statistical power for zero, small, medium, and large true heterogeneity (defined in terms of I^2 = 0/25/50/75%, respectively).

Table 3.

Heterogeneity across primary effects and statistical power of thirteen multi-lab replication projects, ordered with respect to estimated heterogeneity

									Type I Error Rate & Statistical Power			
									Level of heterogeneity			
RP	**Effect**	***K***	**Effect type**	**Effect size estimate**	**I^2 (%)**	**I^2 95% *CI***	**$\hat{\tau}$**	**$\hat{\tau}$ 95% *CI***	**Zero**	**Small**	**Medium**	**Large**
ML2	Intentional Side-Effects*	59	*r*	0.67	93.47	[91.66, 96.51]	0.148	[0.129, 0.205]	0.05	0.48	0.98	1.00
ML1	Anchoring 3 – Everest*	36	*SMD*	2.41	91.29	[86.61, 95.23]	0.693	[0.544, 0.956]	0.05	0.42	0.92	1.00
ML2	Direction & SES	64	*r*	0.20	88.77	[84.14, 92.15]	0.247	[0.202, 0.301]	0.05	0.53	0.99	1.00
ML1	Allowed vs. forbidden†	36	*SMD*	1.93	75.56	[60.32, 85.46]	0.496	[0.348, 0.685]	0.05^b	0.46^b	0.92^b	1.00^b
ML1	Anchoring 2 – Chicago*	36	*SMD*	2.00	75.36	[61.11, 87.15]	0.358	[0.257, 0.533]	0.04	0.40	0.92	1.00
ML2	Moral Typecasting*	60	*r*	0.45	72.94	[61.69, 82.76]	0.110	[0.085, 0.147]	0.05	0.58	0.98	1.00
ML2	Intuitive Reasoning*	57	*r*	0.40	66.48	[54.38, 80.87]	0.103	[0.080, 0.150]	0.05	0.54	0.98	1.00
ML2	Less is Better*	57	*r*	0.39	64.74	[48.82, 76.96]	0.099	[0.071, 0.133]	0.05	0.57	0.97	1.00
ML2	Moral Foundations	60	*r*	0.13	64.74	[49.11, 75.70]	0.091	[0.066, 0.118]	0.05	0.55	0.98	1.00
ML2	Correspondence Bias*	58	*r*	0.69	64.69	[46.20, 73.07]	0.064	[0.044, 0.078]	0.05	0.57	0.98	1.00
ML1	Anchoring 4 – Babies*	36	*SMD*	2.53	64.67	[45.67, 83.33]	0.298	[0.202, 0.492]	0.05	0.42	0.91	1.00
ML2	Actions are Choices	57	*r*	-0.11	63.90	[46.77, 75.97]	0.061	[0.043, 0.081]	0.05	0.52	0.98	1.00
ML2	Trolley Dilemma 1†	59	*r*	0.59	54.07	[31.83, 66.16]	0.080	[0.050, 0.102]	0.05	0.54	0.99	1.00
ML1	Quote Attribution*	36	*SMD*	0.31	52.05	[24.63, 76.25]	0.164	[0.090, 0.282]	0.05	0.45	0.91	1.00
ML2	Social Value Orientation	54	*r*	0.03	50.22	[28.21, 67.88]	0.069	[0.043, 0.100]	0.05	0.52	0.98	1.00
ML2	False Consensus 2*	58	*r*	0.41	43.15	[18.07, 62.64]	0.063	[0.034, 0.093]	0.05	0.58	0.98	1.00
ML1	Anchoring 1 – NYC*	36	*SMD*	1.21	40.23	[10.62, 73.94]	0.152	[0.064, 0.311]	0.05	0.44	0.91	1.00
ML1	IAT correlation math	35	*r*	0.39	40.05	[3.93, 64.97]	0.056	[0.014, 0.094]	0.05	0.40	0.92	1.00
RRR3	Grammar on intentionality*	12	*MD*	-0.25	38.06	[0.00, 85.72]	0.227	[0.000, 0.708]	0.06	0.26	0.68	0.96
ML2	Priming Warmth*	47	*r*	-0.01	36.76	[8.16, 62.73]	0.082	[0.032, 0.140]	0.05	0.51	0.97	1.00
ML2	Tempting Fate*	59	*r*	0.11	36.49	[5.91, 53.57]	0.065	[0.021, 0.091]	0.05	0.58	0.98	1.00
ML3	Subjective Distance interaction	21	*r*	0.02	33.51	[0.00, 76.78]	0.059	[0.000, 0.151]	0.05	0.28	0.83	0.99
ML1	Gender math attitude*	35	*SMD*	0.57	28.06	[0.00, 67.34]	0.112	[0.000, 0.258]	0.05	0.41	0.91	1.00
ML2	Choosing or Rejecting	41	*r*	-0.06	26.49	[0.00, 52.42]	0.047	[0.000, 0.083]	0.06	0.46	0.94	1.00

Table 3 continued

RP	**Effect**	***K***	**Effect type**	**Effect size estimate**	**I^2 (%)**	**I^2 95% *CI***	**$\hat{\tau}$**	**$\hat{\tau}$ 95% *CI***	**Zero**	**Small**	**Medium**	**Large**
ML2	Incidental Anchors*	49	*r*	0.03	24.94	[0.00, 54.71]	0.056	[0.000, 0.107]	0.05	0.49	0.97	1.00
ML3	Credentials interaction	21	*r*	0.02	24.03	[0.00, 73.82]	0.046	[0.000, 0.137]	0.05	0.30	0.80	1.00
ML1	Gambler's Fallacy*	36	*SMD*	0.61	22.85	[0.00, 69.16]	0.090	[0.000, 0.248]	0.05	0.41	0.90	1.00
ML2	Moral Cleansing*	52	*r*	0.01	22.29	[0.00, 51.55]	0.047	[0.000, 0.090]	0.05	0.53	0.98	1.00
ML1	Imagined Contact*	36	*SMD*	0.12	20.60	[0.00, 62.50]	0.080	[0.000, 0.202]	0.05	0.44	0.91	1.00
ML1	Low vs. high category scales†	36	*SMD*	0.88	19.20	[0.00, 49.95]	0.155	[0.000, 0.318]	0.05[b]	0.44[b]	0.92[b]	1.00[b]
RRR9	Hostility priming – Behavior*	22	*MD*	-0.08	18.03	[0.00, 56.25]	0.096	[0.000, 0.233]	0.05	0.34	0.82	1.00
RRR9	Hostility priming – Hostility*	22	*MD*	0.08	17.73	[0.00, 59.61]	0.079	[0.000, 0.207]	0.05	0.30	0.81	1.00
RRR8	Professor priming*	23	*MD*	0.14	17.43	[0.00, 64.79]	0.857	[0.000, 2.538]	0.06	0.33	0.82	1.00
ML1	Norm of reciprocity†	36	*SMD*	-0.36	17.21	[0.00, 47.51]	0.091	[0.000, 0.190]	0.05[b]	0.44[b]	0.91[b]	1.00[b]
ML2	False Consensus 1*	59	*r*	0.48	15.88	[0.00, 40.52]	0.032	[0.000, 0.061]	0.05	0.57	0.98	1.00
ML2	Assimilation & Contrast	59	*q*	-0.07	15.12	[0.00, 33.35]	0.078	[0.000, 0.131]	0.05	0.52	0.98	1.00
ML3	Metaphor	20	*r*	0.14	13.03	[0.00, 57.02]	0.047	[0.000, 0.141]	0.06	0.31	0.81	0.99
RRR1	Verbal overshadowing 1†	32	*RD*	-0.03	12.23	[0.00, 46.51]	0.032	[0.000, 0.081]	0.05[b]	0.34[b]	0.82[b]	0.99[b]
ML2	Priming Consumerism*	54	*r*	0.07	11.97	[0.00, 49.10]	0.035	[0.000, 0.093]	0.05	0.54	0.97	1.00
ML2	Trolley Dilemma 2†	60	*r*	0.13	11.90	[0.00, 33.23]	0.036	[0.000, 0.069]	0.05	0.57	0.98	1.00
ML1	Sunk Costs*	36	*SMD*	0.29	9.18	[0.00, 45.93]	0.050	[0.000, 0.145]	0.05	0.44	0.93	1.00
ML2	Framing†	55	*r*	0.22	5.92	[0.00, 36.47]	0.025	[0.000, 0.075]	0.06	0.55	0.98	1.00
ML2	Position & Power	59	*r*	0.01	3.09	[0.00, 42.19]	0.016	[0.000, 0.074]	0.05	0.58	0.98	1.00
ML2	Disgust & Homophobia	59	*q*	0.04	3.05	[0.00, 30.32]	0.035	[0.000, 0.131]	0.05	0.54	0.98	1.00
RRR7	Intuitive-cooperation*	21	*MD*	-0.39	2.80	[0.00, 39.28]	0.911	[0.000, 4.321]	0.06	0.32	0.81	1.00
ML2	SMS & Well-Being	59	*r*	-0.01	1.84	[0.00, 29.80]	0.013	[0.000, 0.063]	0.05	0.55	0.98	1.00
ML3	Availability	21	*r*	0.04	0.51	[0.00, 56.09]	0.006	[0.000, 0.095]	0.05	0.33	0.82	1.00
ML2	Incidental Disfluency*	66	*r*	-0.02	0.01	[0.00, 27.41]	0.001	[0.000, 0.061]	0.05	0.56	0.99	1.00
ML1	Gain vs. loss framing†	36	*SMD*	-0.66	0.01	[0.00, 55.57]	0.002	[0.000, 0.205]	0.05[b]	0.44[b]	0.91[b]	1.00[b]
ML3	Power and Perspective*	21	*SMD*	0.03	0.01	[0.00, 57.17]	0.002	[0.000, 0.198]	0.05	0.32	0.82	1.00
RRR3	Grammar on intention attribution*	12	*MD*	0.00	0.00[a]	[0.00, 70.62]	0.001	[0.000, 0.185]	0.06	0.24	0.66	0.97
ML3	Conscientiousness and persistence	21	*r*	0.02	0.00[a]	[0.00, 61.42]	0.000[a]	[0.000, 0.104]	0.05	0.35	0.80	1.00

Table 3 continued

RP	Effect	K	Effect type	Effect size estimate	I^2 (%)	I^2 95% CI	$\hat{\tau}$	$\hat{\tau}$ 95% CI	Zero	Small	Medium	Large
RRR3	Grammar on detailed processing*	12	*MD*	-0.10	0.00	[0.00, 54.49]	0.000	[0.000, 0.246]	0.06	0.21	0.68	0.97
RRR5	Commitment on neglect*	16	*MD*	-0.05	0.00	[0.00, 53.18]	0.000	[0.000, 0.208]	0.06	0.28	0.75	0.99
ML3	Warmth Perceptions*	21	*SMD*	0.01	0.00	[0.00, 47.10]	0.000	[0.000, 0.158]	0.06	0.39	0.91	1.00
RRR4	Ego depletion*	23	*SMD*	0.00	0.00	[0.00, 46.91]	0.000	[0.000, 0.169]	0.05	0.33	0.84	1.00
RRR10	Moral reminder*	19	*MD*	0.11	0.00	[0.00, 44.13]	0.000	[0.000, 0.392]	0.06	0.31	0.79	0.99
ML1	Flag Priming*	36	*SMD*	0.02	0.00	[0.00, 36.23]	0.000	[0.000, 0.118]	0.05	0.43	0.92	1.00
ML1	Money Priming*	36	*SMD*	-0.02	0.00	[0.00, 33.18]	0.000	[0.000, 0.110]	0.05	0.48	0.92	1.00
RRR2	Verbal overshadowing 2†	23	*RD*	-0.15	0.00	[0.00, 32.36]	0.000	[0.000, 0.065]	0.05[b]	0.31[b]	0.82[b]	0.99[b]
ML3	Weight Embodiment*	20	*SMD*	0.03	0.00	[0.00, 29.97]	0.000	[0.000, 0.122]	0.06	0.34	0.83	1.00
RRR6	Facial Feedback hypothesis*	17	*MD*	0.03	0.00	[0.00, 25.13]	0.000	[0.000, 0.164]	0.06	0.27	0.79	0.99
ML2	Affect & Risk	60	*r*	-0.04	0.00	[0.00, 21.08]	0.000	[0.000, 0.056]	0.05	0.57	0.99	1.00
ML3	Elaboration likelihood interaction	20	*r*	0.00	0.00	[0.00, 18.62]	0.000	[0.000, 0.042]	0.05	0.31	0.79	1.00
RRR5	Commitment on exit*	16	*MD*	-0.06	0.00	[0.00, 17.44]	0.000	[0.000, 0.089]	0.06	0.29	0.74	0.99
ML3	Stroop effect	21	*r*	0.41	0.00	[0.00, 13.61]	0.000	[0.000, 0.027]	0.05	0.30	0.80	1.00
ML2	Structure & Goal Pursuit	52	*r*	-0.01	0.00	[0.00, 1.91]	0.000	[0.000, 0.013]	0.05	0.53	0.97	1.00
ML2	Direction & Similarity	49	*r*	0.01	0.00	[0.00, 0.00]	0.000	[0.000, 0.000]	0.05	0.54	0.97	1.00

*Note***.** Effects were estimated in metafor using REML. The following effects are odds ratios transformed into standardized mean differences: ‘Allowed vs. forbidden’, ‘Gain vs. loss framing’, ‘Norm of reciprocity’, ‘Low vs. high category scales’. All ML2 meta-analyses with effect type 'r' except 'Moral foundations' and 'Social Value Orientation' were transformed to correlations from a variety of effect sizes. RP = Replication Project, *K* = no. primary studies, $\hat{\tau}$ = between studies standard deviation of effect size, *CI* = confidence intervals. Statistical power was simulated, where Zero = simulated type 1 error, and the other headers represent simulated power under small/medium/large heterogeneity (I^2 = 25/50/75%) respectively. ML = Many Labs, RRR = Registered Replication Report, *SMD* = Standardized Mean difference (Hedge’s *g*), *MD* = Mean Difference, *RD* = Risk Difference, *r* = correlation, *q* = Cohen’s *q*. Code to reproduce table: osf.io/gtfjn

[a] Value rounded down to zero, [b] Odds ratio or risk difference simulated as (standardized) mean difference, * *SMD* effect size in Figure 3, † Log odds ratio effect size in Figure

Heterogeneity Estimates and Confidence Intervals

There is limited evidence for widespread heterogeneity across the examined effects. Rounding I^2 estimates to their closest value of 0/25/50/75% and under the specifications of the original authors 12/68 (18%) meta-analyses have I^2 estimates that best correspond to large heterogeneity (I^2 = 75%), 7/68 (10%) to medium heterogeneity (I^2 = 50%), 18/68 (26%) to small heterogeneity (I^2 = 25%) and 31/68 (46%) to zero heterogeneity (I^2 = 0%). The between studies standard deviation estimates ($\hat{\tau}$) shows a similar pattern, although interpretation is more difficult due to the differences in scale and lack of guidelines. For the two largest groups of effect size measures (correlations and SMDs) the largest $\hat{\tau}$ is .25 and 0.69, respectively, and their quartiles .014/.047/.068 and <0.001/0.090/0.160. The 48 meta-analyses that had confidence intervals of I^2 containing 0 (71%), also had confidence intervals of $\hat{\tau}$ that contained 0. Moreover, the sixteen (24%) meta-analyses with estimated I^2 = 0 also had $\hat{\tau}$ = 0 (note: two meta-analyses had I^2< .005 and were rounded down when printed in Table 3, and one of these also had a $\hat{\tau}$ < .0005 which was rounded down, see table footnote). The percentage of heterogeneity estimates larger than 0 (52/68; 76%) suggests heterogeneity for at least some meta-analyses, as this percentage is higher than the expected frequency of non-zero estimates under homogeneity (47%, or about 32/68), based on the chi-square distribution and average *K* (29) across projects. Hence our results on the assessment of heterogeneity are essentially the same using I^2 or $\hat{\tau}$.

I^2 and Power

Figure 1 shows how estimated I^2 varies across all 68 meta-analyses as a function of true heterogeneity (averaged across all simulation runs). Figure 1 makes clear that I^2 is particularly sensitive to changes in heterogeneity for small heterogeneity, and that estimates of I^2 may differ considerably across projects for the same value of true heterogeneity. This can largely be attributed to differences in the sample sizes of the studies incorporated in a meta-analyses (with larger sample sizes resulting in larger estimates of I^2). For example, the cluster of lines at the bottom all belong to RRR3, the replication project with the lowest average sample size per study (99; see Table 2). This illustrates why only relying on I^2 can be problematic, and why also reporting $\hat{\tau}$ is recommended, despite the fact that the between studies standard deviation (τ) is not measured on the same scale across different effect size measures and estimates are not directly comparable across effect types.

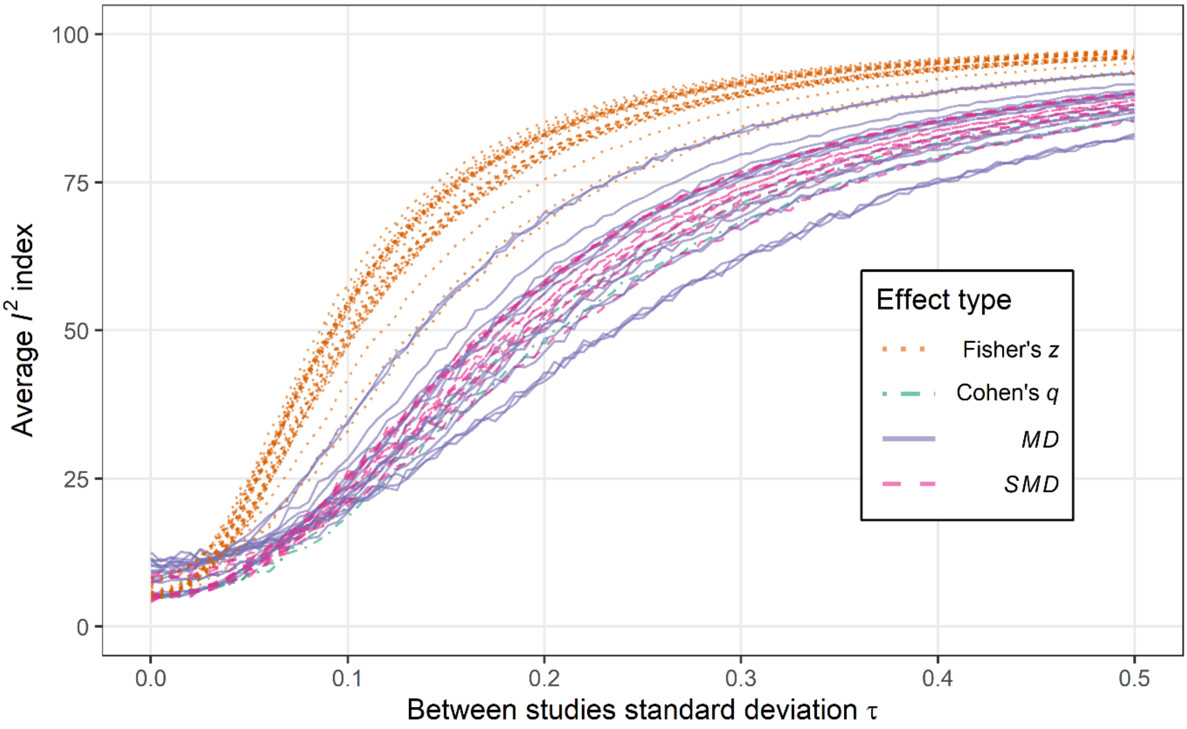

Figure 1. Result of simulation relating I^2-values to between studies standard deviation. Each line represent one of 68 effects. Tau (τ) is not directly comparable across effect size measures. *MD* = Mean Difference, *SMD* = Standardized Mean Difference. Code to reproduce figure: osf.io/u2t3r

Estimated type I error and power for zero/small/medium/large heterogeneity as defined by Higgins (2003) are shown for each meta-analysis in Table 3. In all cases the type I error is approximately nominal, as compared to the expected 5% error rate. Power to detect small heterogeneity was low, ranging from 21% to 58%, with an average of 43%. Power to detect medium heterogeneity was generally very good, with an average of 90% power, but goes down to as low as 66 - 68% for several meta-analyses with low *K* (i.e., meta-analyses from RRR3). Power to detect strong heterogeneity was excellent across the board. To conclude, even though for most projects the number of included studies (median 22) and number of participants (median 96 per study) was relatively large, only power to detect medium or larger heterogeneity was good to excellent, whereas power to detect small heterogeneity was unacceptably low. Hence, even large multi-lab projects struggle to distinguish zero from small heterogeneity when defined as $I^2 = 0$ vs. 25%.

Figure 2 shows the distribution of I^2 at different heterogeneity levels and the distribution of the observed I^2 estimates (bars) using original model and effect size specifications (as detailed in the methods section). The shortest bars in the observed distribution correspond to a frequency of one heterogeneity estimate. The considerable overlap of the theoretical (simulated) probability density functions illustrate that it will be particularly difficult to distinguish zero heterogeneity (i.e., homogeneity) from small heterogeneity (here, $I^2 = 25\%$), and why confidence intervals for I^2 are often wide. Given the distribution of observed I^2 estimates in Table 3 and Figure 2, the majority of observed I^2 estimates are most likely to have zero or zero to small heterogeneity. For I^2 only for twelve meta-analyses there seems to be substantial evidence that they originate from medium or large true effect size heterogeneity, as they fall outside the densities of zero and small true effect size heterogeneity.

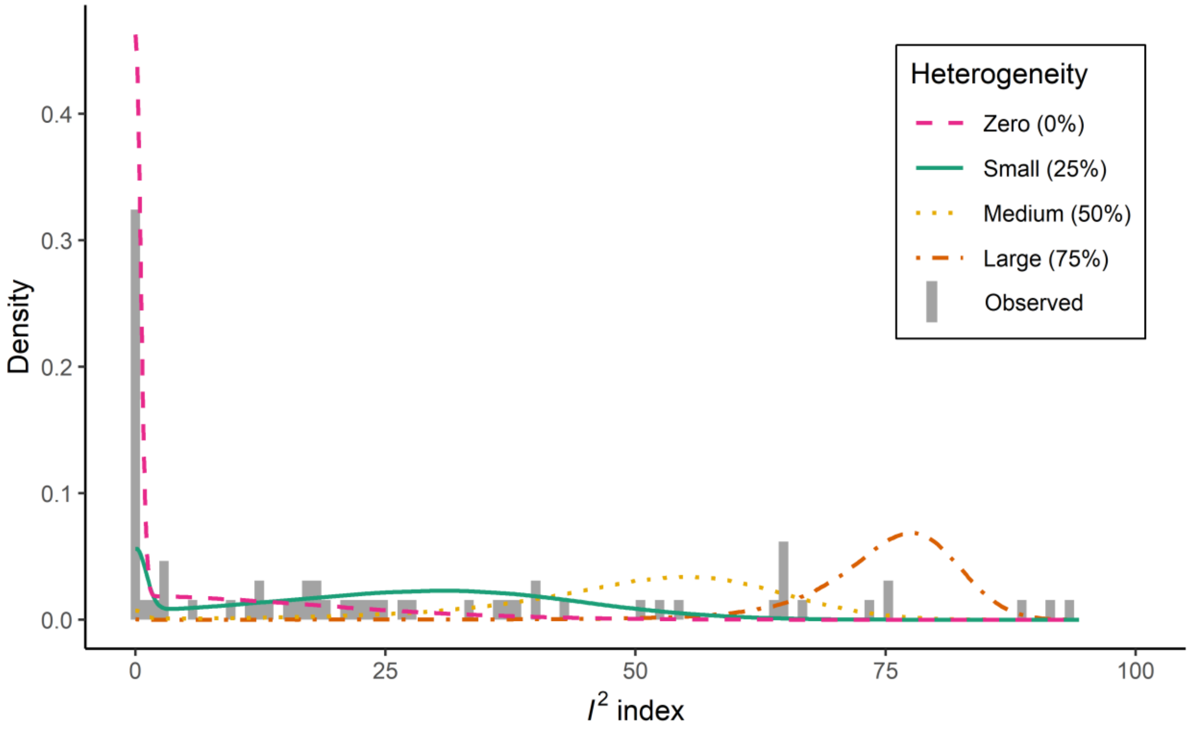

Figure 2. Simulated I^2 densities across 68 meta-analyses for zero, small, medium, and large heterogeneity according to the definitions of Higgins (2003), and the distribution of the observed I^2 estimates (bars) for the 68 meta-analyses. Each simulated density consists of approximately 680,000 estimates. Code to reproduce figure: osf.io/u2t3r

Heterogeneity and Effect Sizes

Larger estimated effect sizes appear to be associated with higher heterogeneity estimates. Our data show a strong correlation between absolute effect size and heterogeneity due to changes in sample population and settings (standardized mean differences and log odds ratios; Figure 3). Amongst the 43 meta-analyses based on standardized mean differences (lower graphs in panels A, B, and C in Figure 3), Pearson's correlations varied from .66 to .79 depending on the measure of heterogeneity ($r_{\hat{\tau}}$ (41) = .77, *p* < .001, bootstrap 95% *CI* [.57, .91]; r_{I2} (41)= .79, *p* < .001, bootstrap 95% *CI* [.63, .90]; r_{H2} (41) = .66, *p* < .001, 95% bootstrap *CI* [.37, .88). Results are similar for the 10 meta-analyses which could be computed as (log) odds ratios (upper graph in panels A, B, and C in Figure 3), although the lower number of effect sizes lead to less precision than for standardized mean differences as can be seen in the wider confidence intervals ($r_{\hat{\tau}}$ (8) = .91, bootstrap 95% *CI* [-.02, .98]; r_{I2} (8) = .90, bootstrap 95% *CI* [-.03, .98]; r_{H2} (8) = .85, bootstrap 95% *CI* [.17, .98]). Excluding Anchoring effects (the 1st, 3rd, 4th, and 6th largest effect sizes amongst average standardized mean differences) as robustness check results in only slightly lower Pearson's correlations between average standardized mean difference effect size and estimated heterogeneity ($r_{\hat{\tau}}$ (37) = .74, *p* < .001, bootstrap 95% *CI* [.48, .92]; r_{I2} (37) = .73, *p* < .001, bootstrap 95% *CI* [.52, .90]; r_{H2} (37) = .64 *p* < .001, bootstrap 95% *CI* [.27, .91]). Also Spearman's rank-order correlation across all average SMDs resulted in similar correlations ($r_{\hat{\tau}}$= .79, *p* < .001, bootstrap 95% *CI* [.62, .88]; r_{I2}= .79, p < .001, bootstrap 95% *CI* [.61, .88]; r_{H2}= .75, p < .001, bootstrap 95% *CI* [.55, .85]).

Finally, amongst the 22 standardized mean differences with average (meta-analytic) effect size not significantly different from zero (given alpha = 0.05), the average estimate of heterogeneity was $\hat{\tau}$ = 0.018, with 14/22 estimates exactly equal to zero (average I^2 = 3.80% and average H^2 = 0.90). These results are in line with zero true heterogeneity for all meta-analyses in this subset, corroborating the proposition that heterogeneity is not to be expected when average effect size is zero. There was only a single log odds ratio with an average effect size not significantly different from zero (Verbal overshadowing 1, *p* = .060, $\hat{\tau}$ = 0.132, I^2 = 11.81, H^2 = 1.05).

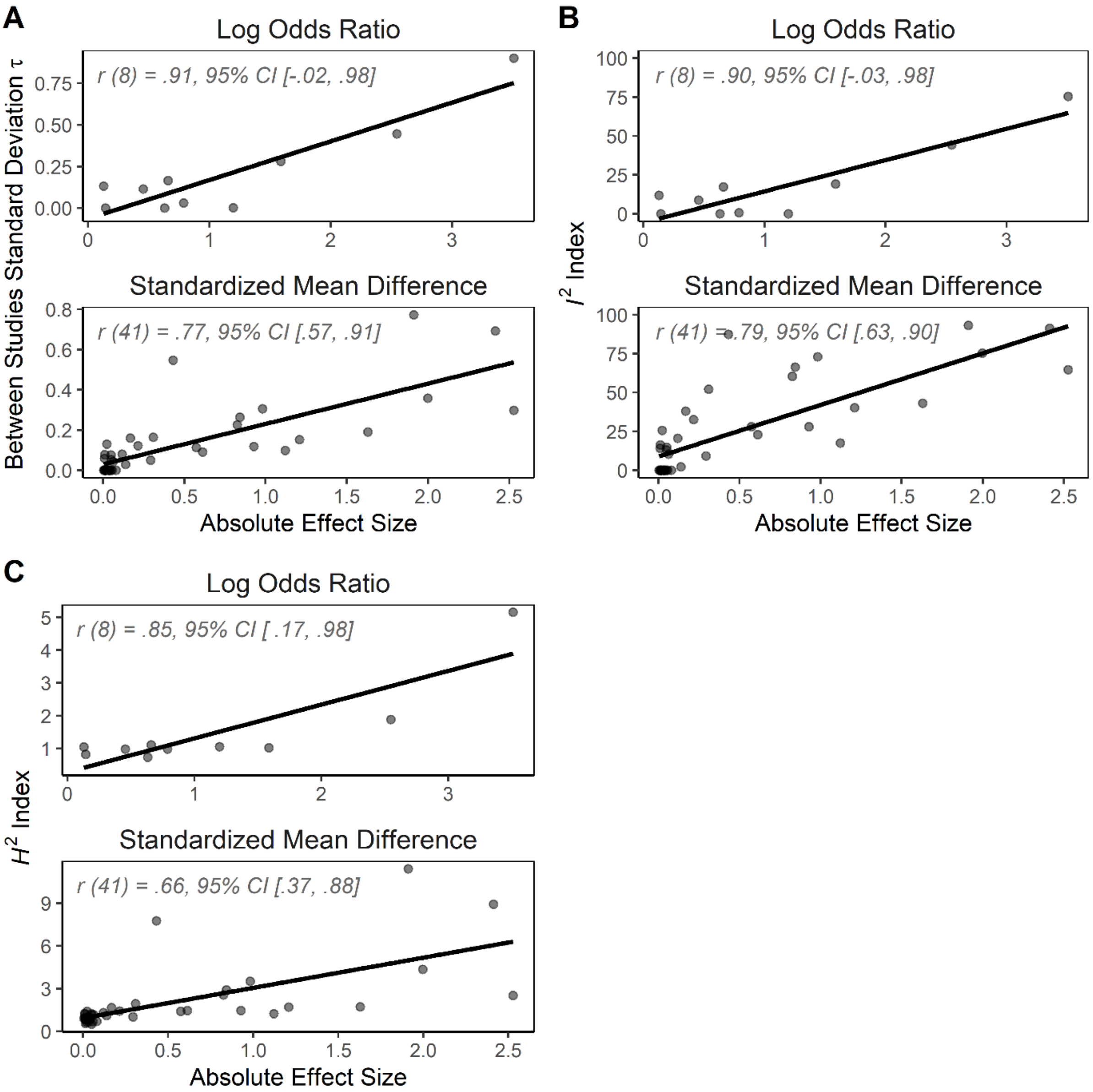

Figure 3. The Pearson correlation between absolute effect size and A) $\hat{\tau}$, B) I^2, and C) H^2 respectively for 43 effects that were measured as mean or standardized mean differences and 10 effects measured as odds ratios or risk differences from 13 pre-registered multi-lab replication projects. Effects reported as mean differences were standardized and odds ratios/risk differences computed as log odds ratios. *r* = product-moment correlation, square

brackets contain 95% bootstrapped percentile confidence intervals. Code to reproduce figure: osf.io/u2t3r

Discussion

We examined the evidence for widespread sensitivity of effect sizes to minor changes in sample population and settings (heterogeneity) in social and cognitive psychology and the correlation between average effect size and this heterogeneity, in a sample of thirteen pre-registered multi-lab direct replication projects in psychology. These thirteen projects examined a total of 68 primary outcome variables and arguably represent the best meta-analytic data currently available in psychology. To aid interpretation we also estimated power of each project to find zero/small/medium/large heterogeneity as defined by Higgins (2003) and approximated the distributions of I^2 under these four heterogeneity levels. Our results showed that most meta-analyses in our sample likely had zero to small heterogeneity, that power to distinguish between zero and small heterogeneity was low for all projects, and that heterogeneity due to changes in sample population and settings was strongly correlated with effect size for standardized mean differences and (log) odds ratios.

In addition to most effects showing no or small heterogeneity, some effects that showed evidence for medium to large heterogeneity were effects that might have been expected to be sensitive to changes in sampling population. That is, most replication projects included a large number of US labs (see Table 2) and some effects that demonstrated heterogeneity used designs where a heterogeneous US-related response would be unsurprising, also within the US. : They either asked questions about the US (anchoring effects), persons related to the US (Quote attribution) or issues that are well-known to generate strong debate in the US (i.e., free speech; allowed vs. forbidden). For instance, someone living close to Chicago is more likely to know the population of Chicago, thereby likely generating heterogeneity in the anchoring effect concerning the population size of

Chicago. We must note, however, that this observation is based on our ad hoc reasoning, and exploratory analyses.

Implications

Our finding that heterogeneity appears to be generally small or non-existent is an argument against so called 'hidden moderators', or unexpected contextual sensitivity. Indeed, our results imply that effects cannot simply be assumed to vary extensively "across time, situations and persons" (Iso-Ahola, 2017, p. 14) and that we should not expect "minor, seemingly arbitrary and even theoretically irrelevant modifications in procedures" (Coyne, 2016, p. 6) to have large impact on effect size estimates. That is, our results suggest that minor changes to sample population and settings are unlikely to affect research outcomes in social and cognitive psychology.

Nonetheless, a few cases in our sample had large heterogeneity estimates. There was no clear pattern in experimental design (as described in Table 2) to indicate when to expect minimal or large heterogeneity. For example, amongst priming effects (the largest subgroup, 23/68 experimental designs) there were both effects with large heterogeneity estimates (Anchoring 1 – 4) and zero (e.g., Structure & Goal Pursuit, Commitment on exit). The same was true when participants were asked to imagine slightly different situations (14/68 experimental designs) where ‘Intentional Side-Effects’ had the largest heterogeneity estimate (I^2) of all meta-analyses, yet several meta-analyses had zero estimates (e.g., Elaboration likelihood interaction, Affect & Risk).

What heterogeneity to expect due to minor changes to sample settings and population seems more dependent on the particular effect rather than on research design features. Researchers should thus carefully consider whether their particular topic is susceptible to changes in context (as also recommend by Simons et al., 2017). For example, a researcher working with anchoring effects might wish to carefully consider minor changes to sample settings and population as heterogeneity for these effects was large, whereas this appears less important for someone researching the Stroop effect. When information on heterogeneity for a particular effect is lacking (i.e., Table 3 only presents results for 68 effects) the appropriate default expectation seems to be that there will be no or very little heterogeneity due to minor changes in sample settings and population, given that this is what we found amongst most effects in our sample, particularly for zero effect sizes. In general, we believe the evidence presented in Table 3 can be useful for researchers seeking to understand why certain research results do or do not replicate. The exact implications for replicability under different frameworks for defining replication await exploration in future work. We cannot and do not generalize our conclusions to conceptual replications, as these studies may vary from original studies in aspects that are expected to yield different effect sizes, anticipated by theory.

In view of the fact that most effects in our sample likely had zero to small heterogeneity, the lack of power to distinguish between these two heterogeneity levels is of concern. That heterogeneity is small is not the same as being negligible, as even small heterogeneity may have consequences for implementing interventions, the advancement of theory, and the interpretation of research outcomes including replication studies. A suggestion to double the already very impressive number of participating labs and individuals of the largest replication projects in our sample seems unrealistic. However, initiatives like the Psychological Science Accelerator, which is a globally distributed network of over 500 psychology laboratories, now allow for more powerful multi-lab projects than those reported in this paper (Moshontz et al., 2018). Regardless, the good news is that sufficient power to detect medium and large heterogeneity is realistically achievable for many large multi-lab replication projects. As these projects' designs and methods are usually carefully controlled, we conclude that large (preferably preregistered) multi-lab studies are very valuable for increasing understanding of psychological phenomena. .

Heterogeneity amongst the studied effects was positively associated with effect size for standardized mean differences and (log) odds ratios. For both standardized mean differences and log odds ratios the correlation was similarly strong (ranging from .66 - .91 across heterogeneity and effect size measures). There are thus both theoretical reasons, related to the measurement reliability of estimates, and empirical reasons to expect larger effect sizes to exhibit comparatively more heterogeneity when using observed effect sizes in a meta-analysis.

For our own sample of meta-analyses, however, we have no evidence that the association between heterogeneity and effect size is (at least partly) explained by differences in measurement reliability amongst labs. Measurement reliabilities were not reported by the projects we examined, and downloading, cleaning and computing them from the raw data is outside of the scope of this paper. However, , the strong similarity of research materials across replication studies does imply smaller differences in measurement reliability than typically found in ‘regular’ meta-analyses in psychology, as these regular meta-analyses include studies with different measurements of the variables involved. We therefore hypothesize that differential measurement reliabilities across studies in the same meta-analysis may at least partially explain why heterogeneity in meta-analysis in psychology is typically larger than those found in multi-lab replication studies. For applied meta-analysts, differential measurement error is thus yet another potential explanation for observed heterogeneity. However, we want to stress that correcting for measurement error when estimating effect size is not an easy fix to the problem of accurately estimating heterogeneity of effect sizes; as both effect size and estimates of reliability are imprecise (i.e., subject to sampling error), attempting to correct for measurement error may also introduce heterogeneity, rather than reduce it.

The positive association between estimates of average effect size and heterogeneity cannot be a statistical artefact resulting from labs with small samples sizes examining large true effect sizes (e.g., because of *a priori* power analyses). Adequate *a priori* power analyses would imply a *negative (*spurious*)* association between estimates of effect size and heterogeneity, which is the opposite of what we found. We expand on this argument in Supplement C.

Our correlational results also suggest that if there is a null overall (meta-analytic) effect size, then it is likely the effect does not exist in any sample population or setting. This is suggested by our finding of no evidence of heterogeneity in the subset of 22 standardized mean differences with an average (meta-analytic) effect size that was not significantly different from zero. However, we advise caution with generalizing this implication. This implication holds for direct replications only, and may not hold for conceptual replications that differ more on sample population and setting (or that differ in a different way than our subset of meta-analyzed studies; see Table 2), or along other dimensions (treatment and measurement variables). Strictly speaking, it is also possible that the effect exists only in a particular subset of the population (e.g., elderly with low education), although we do not believe this is *a priori* likely as an average non-zero effect size would be expected in such a case. Another important caution is that an average effect size may not be (significantly) different from zero because of a lack of statistical power. This is not an issue for the multi-lab projects included in our paper, as their typical statistical power exceeds 0.99 to detect a small true effect size (i.e., Hedges g = 0.20), due to very large sample sizes.

Finally, our analysis reported in Supplement A demonstrates that transforming effect sizes to another metric may generally be inadvisable since the monotonicity principle can be violated (that is, the order of heterogeneity estimates of different meta-analyses may change due to the transformation). In view of this finding researchers may wish to carefully consider whether combining or transforming effect sizes from different effect size metrics in a meta-analysis is advisable.

Limitations and Future Research Directions

There are some limits to the generalizability of claims based on the data in our study. Primarily, the included effects are neither a representative nor random sample of effects in psychology and as such do not support making strong claims about average heterogeneity levels in psychology. We would consider most of the effects described in Table 2 to belong to social and cognitive psychology (and Many Labs 2 explicitly selected effects from these domains). Although these are large subfields in psychology, the lack of effects from other disciplines means our results may not generalize to disciplines such as developmental, clinical or educational psychology. Relatedly, at least the Many Labs studies (which examined many effects in a single session) selected effects partly based on their brevity, and hence we cannot exclude the possibility that our conclusions may be more applicable to these kinds of effects. In addition, we only considered meta-analyses that varied two contextual factors (sample population and settings) that may cause heterogeneity, keeping constant two other ones (treatment and measurement variables), which may have resulted in both lower heterogeneity estimates as well as a stronger relationship between effect size and heterogeneity estimates in our paper.

Moreover, our results may depend on the type of variation in sample population and settings across labs. Most samples consisted of college undergraduates and took place in a lab or online (Table 2) and it may be that there would be more variation across studies when using for example children or an organizational setting. In connection to this, it is possible that our use of a single webpage (*curatescience.org*) has led us to miss some multi-lab direct replication projects, although we believe it unlikely that we have missed many (if any) multi-lab direct replication projects, due to their relatively recent popularity and highly publicized nature. Although we are confident to have included the vast majority of relevant projects currently published, the relatively small number of meta-analyses in our sample means the association between heterogeneity and effect size might be an artifact of the data. However, the exclusion of the rather extreme anchoring effects from our analysis only slightly reduced the correlation between effect size and heterogeneity. We also found similar results across two different effect size measures, although the low number of meta-analyses with log odds ratios (10) meant confidence intervals for that measure were wide. Relatedly, we should stress that while our results point towards most meta-analyses having zero to small heterogeneity, many confidence intervals are very wide and congruent with a large range of actual heterogeneity.

Our results and the limitations of our data provide some guidance in directions of future research. The 68 meta-analyses studied here suggest that zero to small heterogeneity is the standard for direct replications in social and cognitive psychology, but it would still be desirable to examine heterogeneity in a larger sample of meta-analyses of direct replications. We are enthusiastic about the possibilities to do so in the near future, thanks to the many ongoing multi-lab initiatives in psychology (Registered Replication Reports, ManyBabies, the Psychological Science Accelerator). Relatedly, a larger sample of meta-analyses would enable testing whether the correlation between heterogeneity and effect size is generally as strong as what we found in our sample of standardized mean differences. Ideally there would be a continually updating analysis of heterogeneity and the association between heterogeneity, given the many multi-lab projects in psychology likely to be published in the next few years. Moreover, the spread of the multi-lab format to disciplines other than social and cognitive psychology (e.g., ManyBabies; developmental psychology) will enable researchers to examine whether our conclusions also apply to direct replications in other fields of psychology, and for direct replications varying other aspects of sample population and settings than those varied here, and/or other contextual factors. Finally, it may be worthwhile to attempt to disentangle the contribution of reliability to the correlation between heterogeneity and effect size from other aspects of measurement that are likely to contribute, such as range restrictions (Schmidt & Hunter, 2015).

Conclusions

To conclude, in the arguably best meta-analytic data currently available in psychology, most effects likely had zero to small heterogeneity arising from minor variation in sample population and settings, and this heterogeneity was strongly correlated with effect size for standardized mean differences and (log) odds ratios. Despite a relatively large number of studies and participants in each meta-analysis, power was too low to distinguish between zero and small heterogeneity in all cases. Our results suggest that minor changes to sample population and settings are unlikely to affect research outcomes in social and cognitive psychology.

References

Aiken, L. S., West, S. G., & Reno, R. R. (1991). *Multiple regression: Testing and interpreting interactions*. Sage.

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahník, Š., Birch, S., Birt, A. R., Bornstein, B. G., 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., … 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

Augusteijn, H.E.M., van Aert, R.C.M., & van Assen, M.A.L.M. (in press)). The Effect of Publication Bias on the Assessment of Heterogeneity. *Psychological Methods.*

Baumeister, R. F. (2016). Charting the future of social psychology on stormy seas: Winners, losers, and recommendations. *Journal of Experimental Social Psychology*, *66*, 153–158. https://doi.org/10.1016/j.jesp.2016.02.003

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.

Borenstein, M., Higgins, J. P., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. *Research Synthesis Methods, 8(1)*, 5-18.

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

Campbell, D. T., & Stanley, J. C. (2015). *Experimental and quasi-experimental designs for research*. Ravenio Books.

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., Aykutoglu, B., Bahnik, S., Bower, 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., ... & Carcedo, R. J. (2016). Registered Replication Report: Study 1 from Finkel, Rusbult, Kumashiro, & Hannon (2002). *Perspectives on Psychological Science, 11*(5), 750-764.

Cochran, W. G. (1954). The Combination of Estimates from Different Experiments. *Biometrics*, *10*(1), 101. https://doi.org/10.2307/3001666

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. https://doi.org/10.4324/9780203771587

Cooper, H., DeNeve, K., & Charlton, K. (1997). Finding the missing science: the fate of studies submitted for review by a human subjects committee. *Psychological Methods*, *2*(4), 447.

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2009). The handbook of research synthesis and meta-analysis (2nd ed.). Russell Sage Foundation. Coyne, J. C. (2016). Replication initiatives will not salvage the trustworthiness of psychology. *BMC Psychology*, *4*(1). https://doi.org/10.1186/s40359-016-0134-3

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). https://doi.org/10.1186/1471-2288-11-160

Dickersin, K. (2005). Publication bias: Recognizing the problem, understanding its origins and scope, and preventing harm. *Publication bias in meta-analysis: Prevention, assessment and adjustments*, 11-33.

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B. B., Boucher, L., Brown, E. R., Budiman, N. I., Cairo, A. H., Capaldi, C. A, Chartier, C. R., Chung, J. M, Cicero, D. C., Coleman, J. A., Conway, J. G., ... & Brown, E. R. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology, 67, 68-82*.

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., Poroer, 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

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*(3), 891–904. https://doi.org/10.1007/s11192-011-0494-7

Fisher, R. A. (1915). Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. *Biometrika*, *10*(4), 507. https://doi.org/10.2307/2331838

Fisher, R. A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, *1*, 3–32.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502-1505.

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.

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

Hedges, L. V. (1989). An unbiased correction for sampling error in validity generalization studies. *Journal of Applied Psychology*, *74*(3), 469–477. https://doi.org/10.1037//0021-9010.74.3.469

Higgins, J. P. T. (2003). Measuring inconsistency in meta-analyses. *BMJ*, *327*(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539–1558. https://doi.org/10.1002/sim.1186

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I² index? *Psychological Methods*, *11*(2), 193–206. https://doi.org/https://doi.org/10.1037/1082-989X.11.2.193

IJzerman, H., Szymkow, A., & Parzuchowski, M. (2015). Warmer Hearts, and Warmer, but Noisier Rooms: Communality Does Elicit Warmth, but Only for Those in Colder Ambient Temperatures Commentary on Ebersole et al. (2016). *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2698810

Inzlicht, M., Gervais, W., & Berkman, E. (2015). News of Ego Depletion’s Demise is Premature: Commentary on Carter, Kofler, Forster, & Mccullough, 2015. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2659409

Ioannidis, J. P. A., Patsopoulos, N. A., & Evangelou, E. (2007). Uncertainty in heterogeneity estimates in meta-analyses. *BMJ*, *335*(7626), 914–916. https://doi.org/10.1136/bmj.39343.408449.80

Iso-Ahola, S. E. (2017). Reproducibility in Psychological Science: When Do Psychological Phenomena Exist? *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00879

Jackson, D. (2006). The power of the standard test for the presence of heterogeneity in meta-analysis. *Statistics in Medicine*, *25*(15), 2688–2699. https://doi.org/10.1002/sim.2481

Jackson, D., Turner, R., Rhodes, K., & Viechtbauer, W. (2014). Methods for calculating confidence and credible intervals for the residual between-study variance in random effects meta-regression models. *BMC Medical Research Methodology*, *14*(1). https://doi.org/10.1186/1471-2288-14-103

Jacobs, P., & Viechtbauer, W. (2016). Estimation of the biserial correlation and its sampling variance for use in meta-analysis. *Research Synthesis Methods*, 8(2), 161–180. https://doi.org/10.1002/jrsm.1218

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E.., Devos, T., Eisner, M., Frankowska, N., Furrow,, D., Galliani, E. M., … Nosek, B. A. (2014). Investigating Variation in Replicability: A “Many Labs” Replication Project. *Social Psychology*, *45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, S., 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

Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine*, *22*(17), 2693–2710. https://doi.org/10.1002/sim.1482

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.., Bégue, L, Ben-Shakhar, G., Birt, A. R., Blatz, L, Charman, S. D., Claesen, A., Clay, S. L., ... & Yildiz, E. (2018). Registered replication report on Srull and Wyer (1979). *Advances in Methods and Practices in Psychological Science*, 1(3), 321-336.

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for Publication Bias in Meta-Analysis: An Evaluation of Selection Methods and Some Cautionary Notes.

Perspectives on Psychological Science, *11*(5), 730–749. https://doi.org/10.1177/1745691616662243

McShane, B. B., Tackett, J. L., Böckenholt, U., & Gelman, A. (2019). Large-Scale Replication Projects in Contemporary Psychological Research. *The American Statistician*, 73(sup1), 99–105. https://doi.org/10.1080/00031305.2018.1505655

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., Grahe, J. E., McCarth, R. J., Musser, E. D., Atfolk, J., Castille, C., M., Evans, R. 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.

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

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Https://www.R-project.org/

Rücker, G., Schwarzer, G., Carpenter, J. R., & Schumacher, M. (2008). Undue reliance on I 2 in assessing heterogeneity may mislead. *BMC medical research methodology, 8(1)*, 79.

Sánchez-Meca, J., Marín-Martínez, F., & Chacón-Moscoso, S. (2003). Effect-Size Indices for Dichotomized Outcomes in Meta-Analysis. *Psychological Methods*, *8*(4), 448–467. https://doi.org/10.1037/1082-989X.8.4.448

Schmidt, F. L., & Hunter, J. E. (2015). *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. SAGE Publications, Ltd. https://doi.org/10.4135/9781483398105

Simes, R. J. (1986). Publication bias: the case for an international registry of clinical trials. *Journal of clinical oncology*, *4*(10), 1529-1541.

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

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on Generality (COG): A Proposed Addition to All Empirical Papers. *Perspectives on Psychological Science*, *12*(6), 1123–1128. https://doi.org/10.1177/1745691617708630

Simonsohn, U. (2017, October 20). *[63] "Many Labs" Overestimated The Importance of Hidden Moderators*. *Data Colada*. Http://datacolada.org/63

Stanley, T. D. (2017). Limitations of PET-PEESE and other meta-analysis methods. *Social Psychological and Personality Science*, 8(5), 581-591.

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research*. Psychological Bulletin,* 144(12), 1325–1346. https://doi.org/10.1037/bul0000169

Steel, P. (2007). The nature of procrastination: A meta-analytic and theoretical review of quintessential self-regulatory failure. *Psychological bulletin*, *133*(1), 65.

Strack, F. (2016). Reflection on the Smiling Registered Replication Report. *Perspectives on Psychological Science*, *11*(6), 929–930. https://doi.org/10.1177/1745691616674460 van Aert, R. C. M. (2018). Meta-Analysis: shortcomings and potential. Tilburg University.

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 Aert R. C. M., Wicherts J. M., van Assen M. A. L. M. (2019). Publication bias examined in meta-analyses from psychology and medicine: A meta-meta-analysis. *PLOS ONE 14*(4): e0215052. doi: https://doi.org/10.1371/journal.pone.0215052

van Assen, M. A. L. M., van Aert, R. C. M., & Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods*, *20*(3), 293–309. https://doi.org/10.1037/met0000025

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

Van Erp, S., Verhagen, J., Grasman, R. P. P. P., & Wagenmakers, E.-J. (2017). Estimates of Between-Study Heterogeneity for 705 Meta-Analyses Reported in *Psychological Bulletin* From 1990–2013. *Journal of Open Psychology Data*, *5*(1), 4. https://doi.org/10.5334/jopd.33

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., Bégue, L., Ben-Shakhar, G., Birt, A. R., Blatz, L., Charman, S. D.,Claesen, A.., Clay, S. L.,... & Yildiz, E., (2018). Registered replication report on Mazar, Amir, and Ariely (2008). Advances in Methods and Practices in Psychological Science, 1(3), 299-317.

Viechtbauer, W. (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., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, Jr. 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., Connel, L., DeCicco, J. M., ... & Zwaan, A. (2016). Registered Replication Report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, 11(6), 917-928.

von Hippel, P. T. (2015). The heterogeneity statistic I2 can be biased in small meta-analyses. *BMC Medical Research Methodology*, *15*(1). https://doi.org/10.1186/s12874-015-0024-z

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2017). MAKING REPLICATION MAINSTREAM. *Behavioral and Brain Sciences*, 1–50. https://doi.org/10.1017/S0140525X17001972