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

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) and explored the association between heterogeneity and effect size in a sample of 37 effect sizes from ten pre-registered multi-lab direct replication projects in psychology. We found limited heterogeneity; only 7/37 (19%) effects had significant heterogeneity, and most effects (32/37; 86%) were most likely to have zero to small heterogeneity. Power to detect small heterogeneity was low for all projects (mean 36%), but good to excellent for medium and large heterogeneity. Our findings thus show little evidence of widespread heterogeneity in direct replication studies in psychology, implying that citing heterogeneity as a reason for non-replication of an effect is unwarranted unless predicted *a priori*. We also found a strong correlation between observed effect size and heterogeneity in our sample (r = .78, suggesting that heterogeneity and moderation of effects is implausible for a zero average true effect size, but increasingly plausible for larger average true effect size.

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

Word count: 156

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, Shoda, & Lindsay, 2017). In the first stage, inferential statistics 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, Etz, Lucas, & Donnellan, 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.

Heterogeneity is of concern for several reasons. First and foremost, 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, but heterogeneity can also have major consequences for topics such as child development, education, and business performance, where research often impacts policy recommendations. Thus, heterogeneity should be no less of a concern for psychologists than for medical practitioners.

Second, unaccounted for heterogeneity is an indication of incomplete theory, since it suggests that a theory is unable to predict all contextual factors of importance to its claims. As such, heterogeneity might imply previously unknown predictors, so called 'hidden moderators' (Van Bavel, 2016), the discovery of which can be seen as an opportunity for theoretical advancement (Simons et al., 2017; Tackett, McShane, Bockenholt, & Gelman, 2017).

Third, the possibility of heterogeneity can create controversy in the interpretation of replication results. The proclamation of a 'failure' to replicate an effect is sometimes taken to suggest that the original finding was merely a false positive, due to 'p-hacking' (Simmons, Nelson, & Simonsohn, 2011) or publication bias (Inzlicht, Gervais, & Berkman, 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, Szymkow, & Parzuchowski, 2015; Strack, 2016), is that the effect is more heterogeneous than (perhaps implicitly) claimed originally. Such explanations may be valid or not, but even if valid, an effect is typically of less general interest the more specific circumstances it requires to appear. To attenuate the risk of heated discussions on the (non)existence of an effect, original authors are recommended to consider pre-specifying the degree of heterogeneity that would make even them lose interest in the effect (e.g., by declaring 'constraints on generality' (Simons et al., 2017). 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.

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, West, & Reno, 1991). In addition, techniques that attempt to correct for publication bias in their estimate tend to fail in the presence of heterogeneity (McShane, Böckenholt, & Hansen, 2016; van Aert, Wicherts, & van Assen, 2016; van Assen, van Aert, & Wicherts, 2015), which is problematic considering the supposedly widespread publication bias in psychology (Cooper, DeNeve, & Charlton, 1997; Franco, Malhotra, & Simonovits, 2014, 2016)).

It is a commonly believed that heterogeneity is the norm in psychology. In support of this notion, recent large scale reviews of meta-analyses in psychology (Stanley, Carter, & Doucouliagos, 2017; Van Erp, Verhagen, Grasman, & Wagenmakers, 2017) report median heterogeneity levels that can best be described as 'large' (see next paragraph; Higgins, 2003). In comparison, the median heterogeneity estimate in medicine (Ioannidis, Patsopoulos, & Evangelou, 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. The median number of studies (effects) per meta-analysis in the psychology sample of Van Erp et al. (2017) was 12, whereas in medicine it is only 3 (Davey, Turner, Clarke, & Higgins, 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. Moreover, by only including pre-registered multi-lab studies we avoid the issue of publication bias, which can have a large and unpredictable effect on the assessment of heterogeneity (Augusteijn, van Aert, & van Assen, 2018).

In reference to meta-analyses of direct replications, several authors (McShane et al., 2016; Tackett et al., 2017) 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. However, even in the case of multi-lab direct replication projects, studies will 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.

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), which allows comparison of estimates across meta-analyses and has an intuitive interpretation. 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. Higgins (2003) tentatively defined I^2-values of 25, 50, and 75% as small/medium/large heterogeneity respectively, labels we also use in this paper. An alternative index of heterogeneity, though lacking the intuitive interpretation of I^2, is H^2 (Higgins & Thompson, 2002). This index ranges from zero to infinity, with higher values signaling more heterogeneity, with a value of 1 indicating homogeneity.

Tests of heterogeneity typically have low statistical power in many practical situations (Huedo-Medina, Sánchez-Meca, Marín-Martínez, & Botella, 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 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 I^2 has a central chi-square distribution (von Hippel, 2015), a distribution that is right-skewed with more than 40% of observations falling above the expected value (for all $k > 4$). In other words, even in the absence of true heterogeneity, a meta-analysis of 5 or more studies will have an I^2 point estimate above zero in more than 40% of cases. Heterogeneity estimates may thus be congruent with a wide range of true heterogeneity levels. Despite exhortations to the contrary (Ioannidis et al., 2007), it remains common to omit confidence intervals in the reporting of I^2. In consideration of such uncertainty and the prevalent belief that heterogeneity is the norm in psychology, we examine the existing evidence for heterogeneity in psychology using a sample of pre-registered multi-lab direct replication projects.

Association between effect size and heterogeneity

Effect size is likely associated with heterogeneity. Intuitively, it makes sense to believe that if there is no meta-analytic effect there is nothing to moderate (i.e., no heterogeneity). However, a null or near null 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. A large meta-analytic effect size on the other hand, can be expected to be associated with more 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 is the same, the observed correlation between the two variables will differ between studies (see also Schmidt & Hunter, 2015). Keeping measurement reliabilities constant, differences in observed 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, 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} increases the observed differences between studies, thereby increasing heterogeneity of observed effect sizes (see Table 1). We therefore explore with a meta-meta-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.

Effect size ρ_{xy} *and its heterogeneity as a function of true effect size and measurement reliability.*

	$\rho_{xy} = 0$	$\rho_{xy}= .3$	$\rho_{xy} = .5$
$\sqrt{R_{xx'}} \times \sqrt{R_{yy\prime}}. = .6$	0	0.18	0.30
$\sqrt{R_{xx'}} \times \sqrt{R_{yy\prime}}. = .7$	0	0.21	0.35
$\sqrt{R_{xx'}} \times \sqrt{R_{yy\prime}}. = .8$	0	0.24	0.40

Note. Values in cells are observed effect sizes arising from the true effect size ρ_{xy} and measurement reliabilities $\sqrt{R_{xx'}} \times \sqrt{R_{yy\prime}}$.Code to reproduce table: osf.io/kf6pt/

The pre-registered multi-lab replication projects

Table 2 lists the ten replication projects, with a total of 37 primary outcome variables, we used to examine heterogeneity and the correlation between effect size and heterogeneity in psychology. As all ten projects were (relatively) large-scale and pre-registered, our dataset arguably represents the best, least biased, 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. 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	k	Countries	Effects	Participants
ML1	Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, S., Bernstein, M. J., . . . Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project.	36	10	16	5975
ML3	Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., … & Brown, E. R. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication.	21	2	10	2845
RRR1	Alogna, V. K., Attaya, M. K., Aucoin, P., Bahník, S., Birch, S., Birt, A. R., … & Buswell, K. (2014). Registered replication report: Schooler and engstler-schooler (1990).	32	10	1	4117
RRR2	Alogna, V. K., Attaya, M. K., Aucoin, P., Bahník, S., Birch, S., Birt, A. R., … & Buswell, K. (2014). Registered replication report: Schooler and engstler-schooler (1990).	23	8	1	2442
RRR3	Eerland, A., Sherrill, A. M., Magliano, J. P., Zwaan, R. A., Arnal, J. D., Aucoin, P., … & Crocker, C. (2016). Registered replication report: Hart & Albarracín (2011).	12	2	3	1187
RRR4	Hagger, M. S., Chatzisarantis, N. L., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., … & Calvillo, D. P. (2016). A multilab preregistered replication of the ego-depletion effect.	23	10	1	2872
RRR5	Cheung, I., Campbell, L., LeBel, E. P., Ackerman, R. A., Aykutoglu, B., Bahník, S., … & Carcedo, R. J. (2016). Registered Replication Report: Study 1 from Finkel, Rusbult, Kumashiro, & Hannon (2002).	16	5	2	2071
RRR6	Wagenmakers, E. J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams Jr, R. B., … & Bulnes, L. C. (2016). Registered Replication Report: Strack, Martin, & Stepper (1988).	17	8	1	1894
RRR7	Bouwmeester, S., Verkoeijen, P. P., Aczel, B., Barbosa, F., Bègue, L., Brañas-Garza, P., … & Evans, A. M. (2017). Registered Replication Report: Rand, Greene, and Nowak (2012).	21	12	1	3596
RRR8	O'Donnell, M., Nelson, L., McLatchie, N. M., & Lynott, D. J. (2017). Registered Replication Report: Dijksterhuis & van Knippenberg (1998)	23	13	1	4493

Note. For studies with several effects the number of participants is the average across effects, rounded to the closest whole number. Participant numbers are those used for primary analyses by original authors (i.e., after exclusions). RP = Replication Project, k = no. primary studies, ML = Many Labs, RRR = Registered Replication Report. Code to reproduce table: osf.io/kf6pt/

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).

We downloaded and collated summary data from the ten pre-registered multi-lab replication projects in psychology (Table 2). The ten projects constitute all multi-lab direct replication projects (according to ww.curatescience.org) with public data available at the time of collection. Data from all ten projects were available on the Open Science Framework (osf.io) and downloaded between 2018/02/01 and 2018/03/31. Although some projects (e.g. RRR4) reported results from several outcome variables, we only included primary outcome variables as stated in accompanying publications, resulting in a total of 37 effects. For each effect we extracted (osf.io/3bmvc/) 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 effect (see codebook; osf.io/uhq4r/). 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, RRR8). In two cases (RRR5 and RRR6) it was necessary to download the raw data to extract summary data. 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 effects into one dataset for analysis (osf.io/456fz/).

To examine heterogeneity of each of the 37 effects, we computed meta-analytic estimates for all 37 effects in our dataset (Table 3). We ran all analyses as specified by the replication authors (osf.io/kf6pt/). The effect size of the original study, which was the focus of the replication effort, was not included in these meta-analyses. All effects 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: correlations (*r*), standardized mean differences (SMD), raw mean differences (MD), and risk differences (RD). ML1 transformed effect sizes measured as odd ratios into standardized mean differences when meta-analyzing under the assumption that responses followed logistic distributions (Sánchez-Meca, Marín-Martínez, & Chacón-Moscoso, 2003; Viechtbauer, 2010). Two projects (RRR5 and RRR7) used the Knapp and Hartung adjustment of the standard errors (Knapp & Hartung, 2003) and ML3 correlations were corrected for bias (Hedges, 1989; Viechtbauer, 2010). After estimating effect sizes, ML3 transformed correlations into eta-squared for reporting, which we did not.

For each effect we estimated I^2 and its 95% confidence interval. Both were estimated with metafor, which uses a general expression for I^2 (equation 9 in Higgins & Thompson 2002) and estimates its confidence intervals using the *Q*-profile method (Jackson, Turner, Rhodes, & Viechtbauer, 2014).

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 37 effects 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 effects at each of these four heterogeneity levels and compared them with the observed frequency distribution of the observed I^2 estimates of the 37 effects. Hence, five distributions of I^2 were obtained; four simulated and one observed. To do so we simulated results of I^2 for each effect given its number of studies (*K*), sample sizes of those studies (vector N_k), and each of the four heterogeneity levels (osf.io/gbf4u/). We directly simulated the distribution of I^2 for correlation, 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 ML1 which were measured as odds ratios (and then transformed into standardized mean differences) as standardized mean differences.

As our concern was heterogeneity, for convenience we set the 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 analyses assuming 'medium' effect sizes (Cohen, 1988) and indeed found the same average power at the different heterogeneity levels, see Appendix A (osf.io/bsjhu/).

In case the observed effect size was a correlation, one run of a simulation proceeded as follows. First, we randomly sampled K= 37 Fisher-transformed true correlations ρ_i from a normal distribution with mean 0 and heterogeneity (variance) τ^2. Fisher's *z*-transformation is a normalizing correlation transformation that ranges from negative infinity to positive infinity, though except for extreme correlations it stays within the -1 to 1 range (Fisher, 1915; 1921). Second, for each of the *K* true Fisher-transformed correlations we sampled one Fisher-transformed 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 mean differences (and hence also for risk ratios) 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 $N(0, \tau^2)$. Subsequently, K sample means for both control and treatment conditions were sampled, with $\overline{x}_c \sim N(0,1/n_c)$ and $\overline{x}_t \sim N(\mu_k, 1/n_t)$, where n_t and n_c were the observed treatment and control sample sizes for each study. 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 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 37 effects at zero, small, medium, and large heterogeneity we continued as follows. For each of the 37 effects 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 again ran 10,000 simulations, 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 37 effects. 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 37 effect sizes, was generated by pooling the 37 distributions of

10,000 I^2 values in each category of heterogeneity. Hence these I^2 distributions can be considered a mixture distribution of 37 distributions, using equal weights across all 37 effects.

To examine the correlation between effect size and I^2 across all 37 effects we converted all effect sizes to a common metric. We first converted all effect sizes into correlations (osf.io/h9pft/) and used the R-package metafor to estimate I^2 and meta-analytic effects expressed as Fisher-transformed correlations (osf.io/zuwpg/). In doing so, we fitted random-effects models with metafor's default REML estimator. For mean differences we calculated the pooled standard deviation, (Borenstein, 2009, p. 226), standardized the effect size and converted it to a correlation (p. 234) with a correction factor for unequal sample sizes (p. 234). In one case (RRR8), we first had to convert reported standard errors into pooled standard deviations (p.224). For risk differences and odds ratios we first added 1/2 to a cell if it was empty to avoid dividing by zero, next calculated the logarithmic odds ratio (p. 266), converted this to Cohen's d (p. 232) and then finally to a correlation. All formulas used are presented in Appendix B (osf.io/h4vfx/). Since I^2 is set to zero for the majority of cases under homogeneity (i.e., truncated), we also correlated effect size with the closely related heterogeneity estimate H^2 (Higgins & Thompson, 2002) as a sensitivity analysis (osf.io/zuwpg/). To avoid truncation of H^2 we computed it as $Q/(K-1)$, where Q is the Q-test statistic and K is the number of studies, although this expression is most appropriate when using the DerSimonian-Laird estimator of between-study variance rather than REML as we do (Higgins & Thompson, 2002). To describe the association between effect size and heterogeneity we report both Pearson’s product moment correlation and, as the association may be nonlinear, Spearman’s rank order correlation. For these statistics we also report 95% bootstrap confidence intervals using the percentile method (osf.io/zuwpg/).

Results

Table 3 presents the meta-analytic effect size estimates and I^2 with confidence intervals for each of the 37 included effects, as well as simulated type I error and statistical power for zero, small, medium, and large true heterogeneity.

Table 3.

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

							Statistical power			
							Level of heterogeneity			
RP	Effect	k	Effect type	Effect size estimate	I^2 (%)	I^2 95% CI	Zero	Small	Medium	Large
ML1	Anchoring 3 – Everest	36	SMD	2.41	91.29	[86.61, 95.23]	0.04	0.46	0.91	1.00
ML1	Allowed vs. forbidden	36	SMD	1.93	75.56	[60.32, 85.46]	0.05[a]	0.47[a]	0.91[a]	1.00[a]
ML1	Anchoring 2 – Chicago	36	SMD	2.00	75.36	[61.11, 87.15]	0.05	0.44	0.92	1.00
ML1	Anchoring 4 – Babies	36	SMD	2.53	64.67	[45.67, 83.33]	0.05	0.47	0.92	1.00
ML1	Quote Attribution	36	SMD	0.31	52.05	[24.63, 76.25]	0.04	0.43	0.91	1.00
ML1	Anchoring 1 – NYC	36	SMD	1.21	40.23	[10.62, 73.94]	0.05	0.45	0.92	1.00
ML1	IAT correlation math	35	R	0.39	40.05	[3.93, 64.97]	0.05	0.40	0.91	1.00
RRR3	Grammar on intentionality	12	MD	-0.25	38.06	[0.00, 85.72]	0.06	0.22	0.68	0.97
ML3	Subjective Distance interaction	21	R	0.02	33.51	[0.00, 76.78]	0.05	0.33	0.83	0.99
ML1	Gender math attitude	35	SMD	0.57	28.06	[0.00, 67.34]	0.05	0.44	0.90	1.00
ML3	Credentials interaction	21	R	0.02	24.03	[0.00, 73.82]	0.05	0.30	0.81	1.00
ML1	Gambler's Fallacy	36	SMD	0.61	22.85	[0.00, 69.16]	0.05	0.44	0.91	1.00
ML1	Imagined Contact	36	SMD	0.12	20.60	[0.00, 62.50]	0.05	0.44	0.91	1.00
ML1	Low vs. high category scales	36	SMD	0.88	19.20	[0.00, 49.95]	0.04	0.46	0.92	1.00
RRR8	Professor priming	23	MD	0.14	17.32	[0.00, 64.77]	0.05	0.34	0.83	1.00
ML1	Norm of reciprocity	36	SMD	-0.36	17.21	[0.00, 47.51]	0.05	0.43	0.91	1.00
ML3	Metaphor	20	R	0.14	13.03	[0.00, 57.02]	0.05	0.32	0.80	0.99
RRR1	Verbal overshadowing 1	32	RD	-0.03	12.23	[0.00, 46.51]	0.06[a]	0.38[a]	0.90[a]	1.00[a]
ML1	Sunk Costs	36	SMD	0.29	9.18	[0.00, 45.93]	0.05	0.44	0.91	1.00
RRR7	Intuitive-cooperation	21	MD	-0.39	2.80	[0.00, 39.28]	0.05	0.32	0.83	1.00
ML3	Availability	21	R	0.04	0.51	[0.00, 56.09]	0.05	0.34	0.83	1.00

Table 3 continued.

RP	Effect	k	Effect type	Effect size estimate	I^2 (%)	I^2 95% CI	Zero	Small	Medium	Large
ML1	Gain vs. loss framing	36	SMD	-0.66	0.01	[0.00, 55.57]	0.05[a]	0.43[a]	0.91[a]	1.00 [a]
ML3	Power and Perspective	21	SMD	0.03	0.01	[0.00, 57.17]	0.05	0.32	0.81	0.99
RRR3	Grammar on intention attribution	12	MD	0.00	0.00	[0.00, 70.62]	0.06	0.24	0.70	0.96
ML3	Conscientiousness and persistence	21	R	0.02	0.00	[0.00, 61.42]	0.05	0.29	0.79	1.00
RRR3	Grammar on detailed processing	12	MD	-0.10	0.00	[0.00, 54.49]	0.06	0.24	0.70	0.97
RRR5	Commitment on neglect	16	MD	-0.05	0.00	[0.00, 53.18]	0.06	0.28	0.74	0.99
ML3	Warmth Perceptions	21	SMD	0.01	0.00	[0.00, 47.10]	0.04	0.37	0.91	1.00
RRR4	Ego depletion	23	SMD	0.00	0.00	[0.00, 46.91]	0.05	0.32	0.85	1.00
ML1	Flag Priming	36	SMD	0.02	0.00	[0.00, 36.23]	0.05	0.43	0.90	1.00
ML1	Money Priming	36	SMD	-0.02	0.00	[0.00, 33.18]	0.05	0.44	0.91	1.00
RRR2	Verbal overshadowing 2	23	RD	-0.15	0.00	[0.00, 32.36]	0.06[a]	0.31[a]	0.83[a]	1.00[a]
ML3	Weight Embodiment	20	SMD	0.03	0.00	[0.00, 29.97]	0.05	0.35	0.84	1.00
RRR6	Facial Feedback hypothesis	17	MD	0.03	0.00	[0.00, 25.13]	0.06	0.27	0.77	0.99
ML3	Elaboration likelihood interaction	20	R	0.00	0.00	[0.00, 18.62]	0.05	0.31	0.83	0.99
RRR5	Commitment on exit	16	MD	-0.06	0.00	[0.00, 17.44]	0.06	0.27	0.77	0.99
ML3	Stroop effect	21	R	0.41	0.00	[0.00, 13.61]	0.05	0.29	0.80	0.99

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’. RP = Replication Project, k = no. primary studies, Estimate = Point estimates of effect sizes, I^2 95% CI = I^2 95% confidence interval. 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. SMD = Standardized Mean difference (Hedge’s g), MD = Mean Difference, RD = Risk Difference, r = correlation. Code to reproduce table: osf.io/kf6pt/

[a] Odds ratio or risk difference simulated as (standardized) mean difference

***I^2* estimates and confidence intervals**

There is limited evidence for widespread heterogeneity across the examined effects. Of the 37 effects, 4/37 (11%) have I^2 estimates that best correspond to large heterogeneity (I^2 = 75%), 4/37 (11%) to medium heterogeneity (I^2 = 50%), 9/37 (24%) to small heterogeneity (I^2 = 25%) and 20/37 (54%) to zero heterogeneity (I^2 = 0%). However, despite a relatively large number of studies and total sample size for most projects (see Table 2), Table 3 shows very wide confidence intervals (spanning 50% or more) for many effects. The lower bound I^2 confidence interval excludes zero for only 7/37 effects (19%; Table 3), all part of the ML1 project. The percentage of heterogeneity estimates larger than 0 (25/37; 68%, two effects had $I^2 < .005$ and are rounded down in Table 3) suggests heterogeneity for at least some effects, as this percentage is higher than the expected frequency of non-zero estimates under homogeneity (46%, or about 17/37), based on the chi-square distribution and average *k* across projects.

***I^2* and power**

Figure 1 shows how estimated I^2 varies across all 37 effects 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). Since the between studies variance is not measured on the same scale when using different effect size measures, estimates are not directly comparable across effect types.

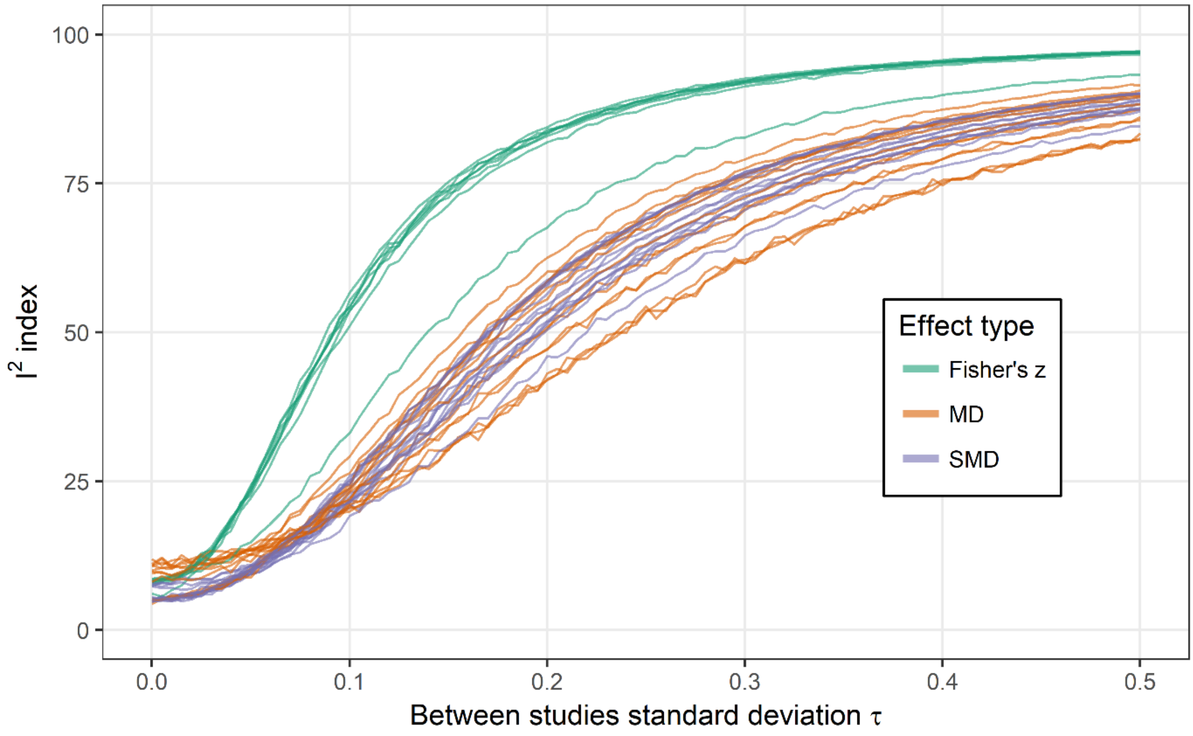

Figure 1. Result of simulation relating I^2-values to between studies standard deviation. Each line represent one of 37 effects. Tau is not directly comparable across effect size measures. Code to reproduce figure: osf.io/zuwpg/

Estimated type I error and power for zero/small/medium/large heterogeneity are shown for each effect 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 24% to 47%, with an average of 36%. Power to detect medium heterogeneity was generally very good, with an average of 85% power, but goes down to as low as 68 - 69% for several effects with low *K* (i.e., effects 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 23) and number of participants (median 102 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.

Figure 2 shows the distribution of I^2 at different heterogeneity levels and a histogram of the observed effects. The shortest bars in the histogram correspond to the heterogeneity estimate of one effect and taller bars correspond to more than one effect. The considerable overlap of the theoretical (simulated) probability density functions illustrates why sufficient power can be difficult to achieve, and why confidence intervals for I^2 are often wide. In particular, the dispersion of the distribution under small heterogeneity is illustrative of the low power to be expected under such circumstances. Given the histogram of observed effect sizes and densities of the distributions (height of the curves), the majority of observed effects are most likely to have zero or zero to small heterogeneity. Only for five effects there seems to be substantial evidence that they originate from medium (two) or large true effect size heterogeneity, as they fall outside the dominant densities of lower true effect size heterogeneity.

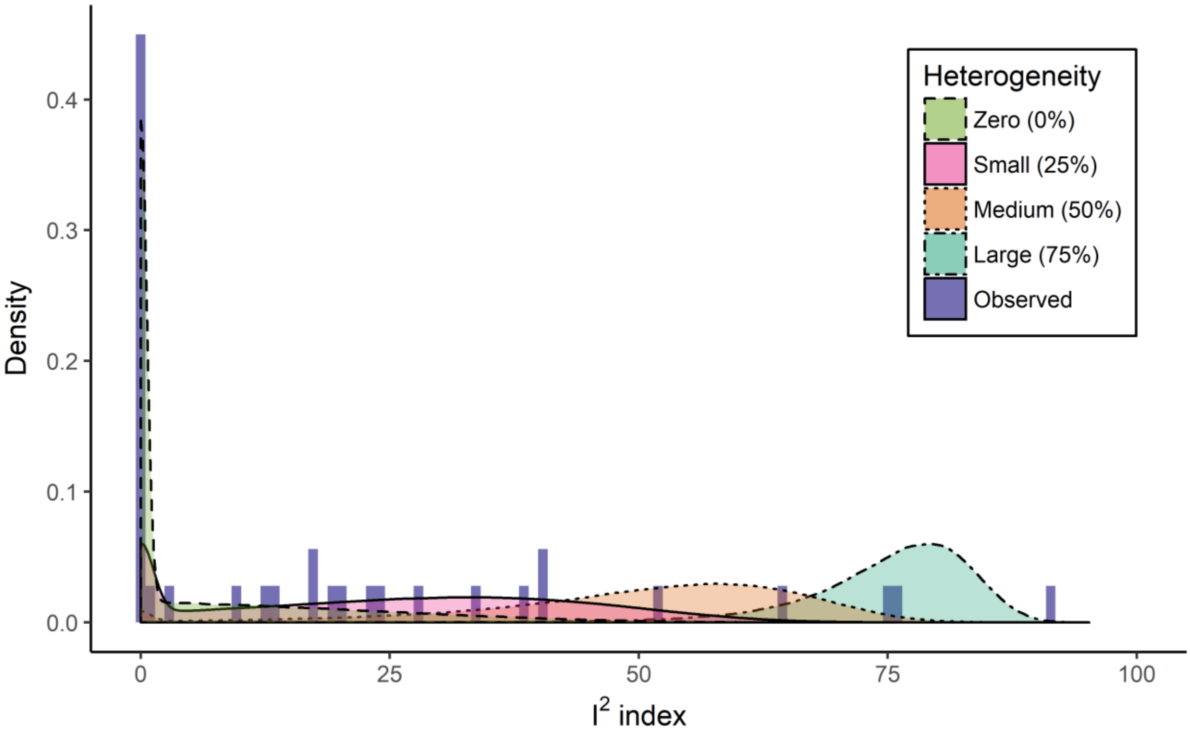

Figure 2. *Simulated I^2 densities across 37 effects for zero, small, medium, and large heterogeneity according to the definitions of Higgins (2003), and a histogram of the observed*

I^2 estimates for the 37 effects. Each simulated density consists of approximately 370,000 estimates. Code to reproduce figure: osf.io/zuwpg/

I^2 and effect sizes

Larger estimated effect sizes appear to be associated with higher heterogeneity estimates. In Table 3, the four effects with highest I^2 estimates also have the largest effect size estimates. Our data show a strong correlation between I^2 and (absolute) effect size ($r = .78$ $p < .001$, 95% bootstrap CI [.59, .90]), see Figure 3). The closely related, but unbounded, heterogeneity estimate H^2 provides a similar result ($r = .72$; $p < .001$, 95% bootstrap CI [.47, .89]). Excluding Anchoring effects (the 1st, 3rd, 4th, and 6th largest effect sizes) as robustness check results in an only slightly lower correlation for I^2 ($r = .71$ p < .001, 95% bootstrap CI [.21, .89]). Spearman's rank-order correlation is also strong for I^2; $r = .70$, $p < .001$, 95% bootstrap CI [.49, .85].

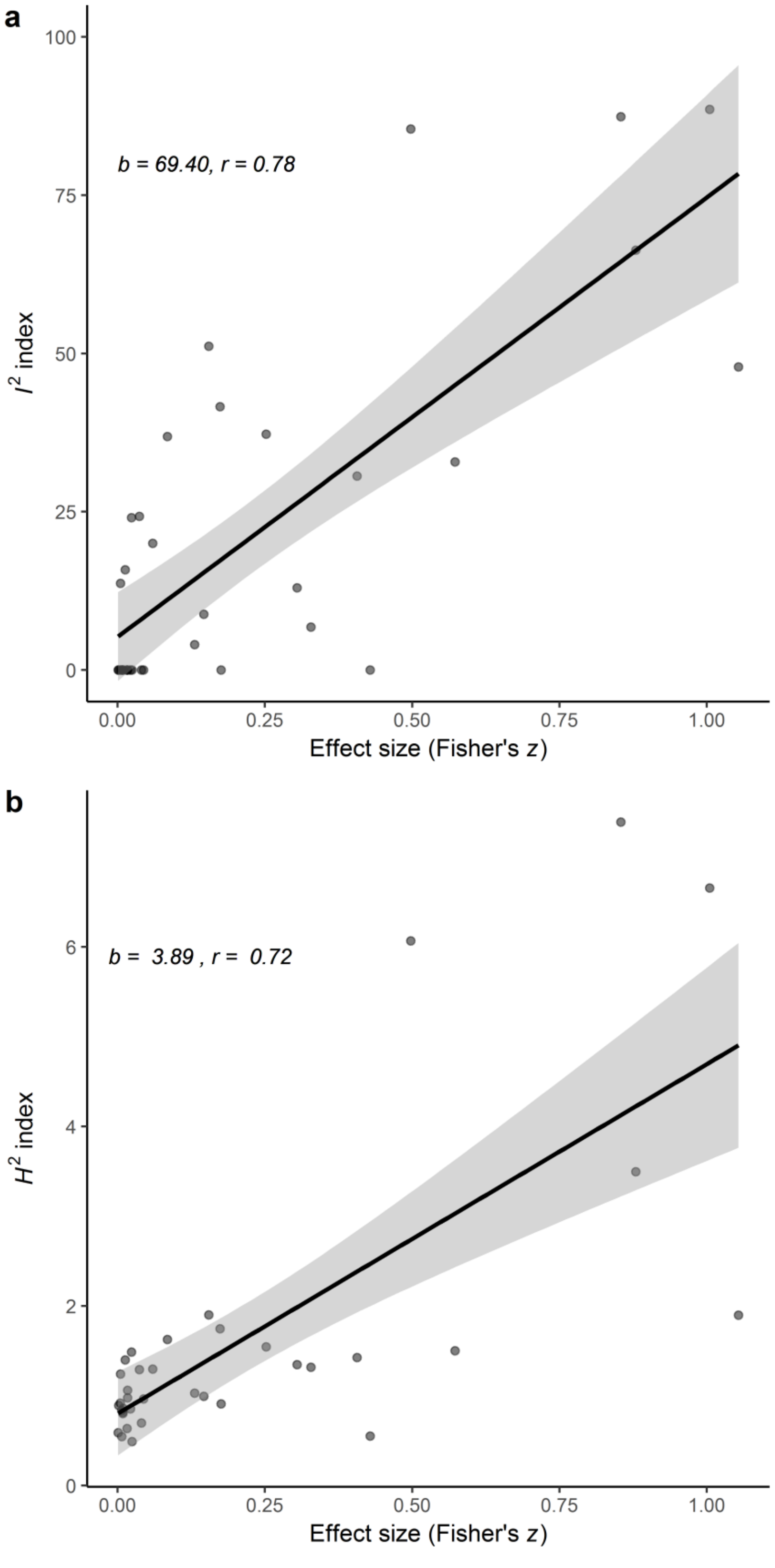

Figure 3. The correlation between a) I^2 and effect size and b) H^2 and effect size for 37 effects from ten pre-registered multi-lab replication projects. Shaded bands represent 95% confidence intervals. Code to reproduce figure: osf.io/zuwpg/

Discussion

We examined the evidence for widespread heterogeneity in psychology and the correlation between effect size and heterogeneity, in a sample of ten pre-registered multi-lab replication projects in psychology. These ten projects examined a total of 37 primary outcome variables and arguably represent the best, least biased, 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 by far most effects in our sample likely had zero (81% of confidence intervals included zero heterogeneity) to small heterogeneity, that power to distinguish between zero and small heterogeneity was low for all projects, and that heterogeneity was strongly correlated with effect size in our sample.

In addition to most effects showing no or small heterogeneity, the effects that showed evidence for medium to large heterogeneity were primarily effects that might have been expected to be sensitive to changes in sampling population. That is, save two effects (Anchoring - Everest and IAT correlation math), all other effects that demonstrated heterogeneity were related to 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). Although ML1 tested US vs. non-US as a moderator of heterogeneity and found very small effect sizes, these are all effects for which heterogeneous responses also within the US would be unsurprising (e.g., someone living close to Chicago is more likely to know the population of Chicago). We must note, however, that this observation is based on our ad hoc reasoning, and exploratory analyses.

Our finding that heterogeneity appears to be small or non-existent except where it might have been expected, 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 estimates. That is, our results imply that citing heterogeneity as a reason for non-replication of an effect is unwarranted unless predicted *a priori* (Simons et al., 2017). 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. The good news is that sufficient power to detect large and medium heterogeneity is realistically achievable for many meta-analyses. We therefore conclude that large (preferably preregistered) multi-lab studies are very valuable for increasing understanding of psychological phenomena.

Heterogeneity amongst the studied effects was strongly associated with effect size. There are thus both good 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. This creates challenges in disentangling the roots of heterogeneity. Consequently, for researchers who wish to examine heterogeneity to further theory development, it may be desirable to endeavor to account for measurement reliability in the effects that are aggregated in meta-analyses. We recognize that not all measurements (e.g., behavioral) admit adjusting for reliability as easily as questionnaires, meaning a correlation between effect size and heterogeneity could at times be difficult to control for. Nonetheless, the extensive use of different scales in psychology means that for many meta-analysts there should be little reason not to control for measurement reliability when aggregating results. Controlling for measurement error, however, introduces the problem into the analysis that estimates of measurement reliability are themselves imprecise, which will affect the study effect size estimates and thereby the estimated overall effect size and heterogeneity.

There are some limits to the generalizability of claims based on the data in our study. For one, 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. More particularly, 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. Relatedly, the relatively small number of effects in our sample means the association between heterogeneity and effect size might be an artifact of the data, although exclusion of the rather extreme anchoring effects from our analysis only slightly reduced the correlation between effect size and heterogeneity. Finally, we should stress that while our results point towards most effects 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. To fully establish whether zero to small heterogeneity is the standard for direct replications in psychology, as suggested by our results, it would be desirable to examine heterogeneity in a larger sample of meta-analyses of direct replications than the 37 examined here. 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, Many Labs 2, ManyBabies, the Psych Science Accelerator). Relatedly, a larger sample of effects would enable testing whether the correlation between heterogeneity and effect size is generally as strong as what we found in our sample. Moreover, it may be worthwhile to attempt to disentangle the contribution of reliability to this correlation from other aspects of measurement that are likely to contribute, such as range restrictions (Schmidt & Hunter, 2015).

To conclude, in the arguably best meta-analytic data currently available in psychology, most effects likely had zero to small heterogeneity, and heterogeneity was strongly correlated with effect size. 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 little reason to believe heterogeneity is widespread in psychology.

References

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

Augusteijn, H., van Aert, R. C. M., & van Assen, M. A. L. M. (2018, September 13). The Effect of Publication Bias on the Assessment of Heterogeneity. https://doi.org/10.31219/osf.io/gv25c

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. doi: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). New York: Russel Sage Foundation.

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

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences. 2nd*. Hillsdale, NJ: erlbaum.

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.

Coyne, J. C. (2016). Replication initiatives will not salvage the trustworthiness of psychology. *BMC Psychology*, *4*(1). doi: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). doi:10.1186/1471-2288-11-160

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*(3), 891–904. doi: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. doi: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.

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

Higgins, J. P. T. (2003). Measuring inconsistency in meta-analyses. *BMJ*, *327*(7414), 557–560. doi: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. doi: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. doi: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*. doi: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*. doi: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. doi:10.1136/bmj.39343.408449.80

Iso-Ahola, S. E. (2017). Reproducibility in Psychological Science: When Do Psychological Phenomena Exist? *Frontiers in Psychology*, *8*. doi: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. doi: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). doi:10.1186/1471-2288-14-103

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., … Nosek, B. A. (2014). Investigating Variation in Replicability: A "Many Labs" Replication Project. *Social Psychology*, *45*(3), 142–152. doi:10.1027/1864-9335/a000178

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

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. doi:10.1177/1745691616662243

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

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. doi: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*. 1 Oliver's Yard, 55 City Road London EC1Y 1SP: SAGE Publications, Ltd. doi:10.4135/9781483398105

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. doi: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. doi:10.1177/1745691617708630

Simonsohn, U. (2017, October 20T11:00:52+00:00). *[63] "Many Labs" Overestimated The Importance of Hidden Moderators*. *Data Colada*. Retrieved June 27, 2018, from http://datacolada.org/63

Stanley, T., Carter, E. C., & Doucouliagos, H. (2017). What Meta-Analyses Reveal about the Replicability of Psychological Research. *Deakin Laboratory for the Meta-Analysis* *of Research*, *Working Paper*. Retrieved from http://www.deakin.edu.au/__data/assets/pdf_file/0007/1198456/WhatMeta-AnalysesReveal_WP.pdf

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. doi:10.1177/1745691616674460

Tackett, J. L., McShane, B. B., Bockenholt, U., & Gelman, A. (2017). Large Scale Replication Projects in Contemporary Psychological Research. *ArXiv Preprint ArXiv:1710.06031*.

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. doi:10.1177/1745691616650874

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. doi:10.1037/met0000025

Van Bavel, J. J. (2016). Contextual Sensitivity Helps Explain the Reproducibility Gap between Social and Cognitive Psychology. *SSRN Electronic Journal*. doi: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. doi:10.5334/jopd.33

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

von Hippel, P. T. (2015). The heterogeneity statistic I2 can be biased in small meta-analyses. *BMC Medical Research Methodology*, *15*(1). doi: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. doi:10.1017/S0140525X17001972