Heterogeneity of Research Results as a New Perspective to Assess and Promote Progress in Psychological Science

Audrey Helen Linden and Johannes Hönekopp

Will appear in *Perspectives on Psychological Science*

Author note

We would like to thank Frank Renkewitz and Peter Sedlmeier for critical comments on an earlier version of this paper.

Correspondence concerning this article should be addressed to Johannes Hönekopp, Department of Psychology, Northumbria University, Newcastle upon Tyne, NE1 8ST, United Kingdom.

Contact: johannes.honekopp@unn.ac.uk

Abstract

Heterogeneity emerges when multiple close or conceptual replications on the same subject produce results that vary more than expected from sampling error. Here we argue that unexplained heterogeneity reflects a lack of coherence between the concepts applied and data observed, and therefore a lack of understanding of the subject matter. Consequently, typical levels of heterogeneity offer a useful but neglected perspective on levels of understanding achieved in psychological science. Focusing on continuous outcome variables, we surveyed heterogeneity in 150 meta-analyses from cognitive, organizational, and social psychology and 57 multiple close replications. Heterogeneity proved very high in meta-analyses, with powerful moderators being conspicuously absent. Population effects in the average meta-analysis vary from small to very large, for reasons that are typically not understood. In contrast, heterogeneity was moderate in close replications. A newly identified relationship between heterogeneity and effect size allowed us to make predictions about expected heterogeneity levels. We discuss important implications for the formulation and evaluation of theories in psychology. Based on insights from the history and philosophy of science, we argue that reduction of heterogeneity is important for progress in psychology and its practical applications, and we suggest changes to our collective research practice towards this end.

keywords: meta-analysis; heterogeneity; replication; statistical power; philosophy of science; psychological research.

Heterogeneity of Research Results as a New Perspective to Assess and Promote Progress in Psychological Science

Meta-analysis, which seeks to summarize results of multiple studies into the same phenomenon, has become an indispensable tool in contemporary research. In pioneering work, Smith and Glass (1977) showed that psychotherapy has a strong positive effect on the average patient studied, and Schmidt and Hunter (1977) demonstrated that the validity of employment tests generalize more readily across different job types than previously believed. Influential surveys of meta-analyses demonstrated the effectiveness of psychological interventions (Lipsey & Wilson, 1993), provided effect size benchmarks for social psychology (Richard, Bond, & Stokes-Zoota, 2003), and summarized findings on psychological gender similarities (Hyde, 2014). Here, we provide a survey of meta-analyses that shifts the perspective from the mean effect size in a population of studies (i.e. how large is the average effect in a particular domain?) to the heterogeneity of results (i.e. how much do results differ across studies into the same issue?). In any meta-analysis, heterogeneity indicates to what extent the summarized studies tap into the same population effect size. If they investigate the same population effect size, heterogeneity will be zero. Even in this case, sampling error will create differences in observed effects across studies. Zero heterogeneity is inferred where these observed differences do not exceed the level expected due to sampling error. Consider the effectiveness of psychotherapy as an example. If heterogeneity was zero, the effectiveness of psychotherapy would be the same across all studies, regardless of the issue patients present with (e.g. anorexia, depression, specific phobia), the type of therapy they receive (e.g. cognitive-behavioral, psychoanalytic), and other differences. Obviously, this is unrealistic; for example, some conditions are treated more successfully than others (Huhn et al., 2014). Heterogeneity then reflects how much the population effect sizes differ across studies. We provide a formal treatment of heterogeneity later, but Figure 1 provides examples with high and low heterogeneity.

Heterogeneity tends to receive little attention from researchers (Aytug, Rothstein, Zhou, & Kern, 2012; Dieckmann, Malle, & Bodner, 2009; Ioannidis, 2008), but we argue here that much is to be gained from its study because a) heterogeneity reflects the degree of understanding of the subject matter being investigated, and b) its study offers useful suggestions regarding the improvement of our collective research practice.

Why Heterogeneity Matters

Low (as opposed to high) heterogeneity reflects a more advanced understanding of the subject matter being studied. This is because high heterogeneity, at least as long as it remains unexplained, suggests the lack of a strong coherence between the concepts applied and the data observed. Take visuo-spatial skills in people with autism spectrum conditions (ASC) as an example (Muth, Hönekopp, & Falter, 2014). In line with current theorizing, the average study found (moderately) better visuo-spatial performance in people with ASC than in IQ-matched controls on a number of standardized tasks. At the same time, heterogeneity of results proved high even for the same task. Not accounted for by any theory, this random variation in study results (which might be down to unrecognized variability in ASC, unreliability in diagnosis, or other factors) points to a shortcoming in our understanding. It also implies that the result of the next study into the same question is highly unpredictable (i.e. over and above the uncertainty arising from sampling error).

Moreover, low heterogeneity should facilitate future progress for two reasons. First, a clear structure in observable data can in itself guide understanding – a point stressed by 17th century luminaries Francis Bacon and Newton as well as modern philosophers of science such as Reichenbach, Hanson, and Simon (Schickore, 2018; Simon, 1973). For example, the 19th century astronomer William Huggins observed that the light of different stars, when seen through a prism, shows the same set of spectral lines; but these are collectively shifted to varying degrees. The observation of this systematic redshift pattern led to the discovery that stars move away from us and at different speeds (Schneider, 2014). Sixty years later, Edwin Hubble observed that the degree of stars' redshift is linearly related to their distance from us, which led to the discovery that the universe is expanding (Schneider, 2014). Skinner (1956) and Stevens (1957) provide prominent examples for a guiding role of orderly observation data in psychology. Second, the systematic violation of expectations has often proved crucial for scientific discovery (Kuhn, 1970). Thus, the failure of an increasingly convoluted Ptolemaic system to further improve the predictions of astronomic events motivated Copernicus to devise a new, heliocentric model of the cosmos; and the failure to detect expected changes in the speed of light – derived from the idea that light propagates through a medium – led Einstein to abandon the idea of a luminiferous ether and to fundamentally rethink physics. As captured in Bacon's dictum that "truth emerges more readily from error than from confusion" (Kuhn, 1970, p. 57), such anomalies cannot emerge where theoretical concepts and observed data lack a clear connection in the first place.

We therefore propose heterogeneity as a useful perspective from which to judge the success of psychological science, alongside other yardsticks such as the generation of good theories (Wallis, 2015), the design of successful interventions (Lipsey & Wilson, 1993), and beneficial contributions to policy design (Fischhoff, 1990). Thus, heterogeneity is of considerable intrinsic value, which is why we seek to systematically measure this in psychological research results here. What are typical levels? Do they differ across domains, and if so, can we make sense of these differences? Apart of its intrinsic value, knowledge of actual levels of heterogeneity has immediate practical implications: Recent work demonstrated that heterogeneity typically decreases the statistical power of studies[1], i.e. any real effect under investigation is less likely to produce a statistically significant result (Kenny & Judd, 2019; McShane & Böckenholt, 2014; Shrout & Rodgers, 2018). For sample size planning to take this into account, reliable estimates of heterogeneity are needed, which we supply here. Finally, and perhaps most importantly, our findings have clear implications for improving our collective research practice, as we will discuss at the end of our paper.

Before we can address the details of our study, it is necessary to deal with a number of critical points, which we address in the next sections.

Moderators

Importantly, heterogeneity reflects a lack of understanding only where it remains unaccounted for. Let us reconsider our example of psychotherapy effectiveness. A meta-analysis that summarizes all sensible studies should find large heterogeneity because these studies will differ in key variables like the issue being treated, the therapy being used, etc. If this heterogeneity can be explained by moderators (e.g. that effectiveness differs strongly across treated disorders or across types of psychotherapy), obviously this no longer indicates a lack of knowledge. (On the contrary, it might be argued that explained heterogeneity reflects a progress in understanding.) To date, we are not aware of any study which systematically investigates to what extent heterogeneity that is observed in a set of studies is accounted for by moderators. We therefore investigate this here.

Conceptual versus Close Replications

Heterogeneity as a concept only makes sense if the set of studies for which it is computed can, in some sense, be conceived as replications of each other. In this context, the

[1] This is because the gain in power for larger-than-expected effects is less than the loss in power for smaller-than-expected effects.

differentiation of close and conceptual replications has become fruitful (Schmidt, 2009; Zwaan, Etz, Lucas, & Donnellan, 2018). The former seek to replicate an earlier study as faithfully as possible. The Open Science Collaboration (OSC, 2015) project is a famous example. In a massive collaborative effort, the authors sought to replicate 100 studies recently published in high profile psychology journals. The replications sought to copy study materials, data analyses, and other key aspects of the original studies as closely as possible and can therefore be considered close replications. In contrast, the studies summarized in a meta-analysis can typically be considered to be conceptual replications (Schmidt & Oh, 2016), i.e. whilst they address the same topic or mechanism, they often differ markedly regarding their design, study materials, participants, data analysis, and other key aspects. Consequently, heterogeneity should tend to be larger in conceptual replications than in close replications.

A systematic comparison of heterogeneity in close and conceptual replications should be instructive. For example, Stanley, Carter and Doucouliagos (2018) argued that the low replicability observed in the OSC (2015) project might reflect low power caused by high heterogeneity. However, the heterogeneity data that they presented in support of this argument stemmed almost exclusively from conceptual replications. Their assumption that heterogeneity in close replications attempts might be similar rested on only two examples for the latter.

Note that the heterogeneity for each of the 100 twin studies in the OSC (original and replication) cannot be estimated reliably. Instead of a single replication, this would require multiple close replications of the same effect (e.g., Klein et al., 2014). Consequently, we use such Many-Labs type replications to study heterogeneity in close replications.

Measuring Heterogeneity

So far, we have not addressed how heterogeneity can be quantified. In psychology the idea of heterogeneity is usually discussed in the context of standardized effect sizes (e.g. Cohen's *d* instead of a difference between group means in raw scores), and we stick to this perspective here. Two established approaches to quantify heterogeneity are I^2, and τ. Before we deal with them, it is helpful to consider what we should expect to see in the absence of heterogeneity. Even if all primary studies tap into the same population effect size, we expect to see differences in the observed effect sizes due to sampling error. Thus, observed differences between effect sizes do not necessarily point to heterogeneity.

The first approach, I^2, estimates the percentage of observed effect size variability that reflects real differences in effect sizes. Where I^2 is near zero, the observed variability is mostly down to sampling error; where I^2 is near 100, most of the observed variability reflects differences in population effect sizes. However, I^2 does not discriminate well when heterogeneity is large (Huedo-Medina, Sánchez-Meca, Marín-Martínez, & Botella, 2006). Moreover, I^2 depends on the sample size in the primary studies (Borenstein, Higgins, Hedges, & Rothstein, 2017; IntHout, Ioannidis, Borm, & Goeman, 2015). Imagine that all studies used small samples. Individual effect sizes will scatter widely around their population effect size. Consequently, a large percentage of observed variability reflects sampling error and I^2 will be low. Now imagine that all studies used very large samples. Each study will provide a highly accurate estimate of its population effect size. Consequently, only a small percentage of observed variability reflects sampling error and I^2 will be high. Finally, for the current study, our approach to heterogeneity involves summarizing heterogeneity estimates across multiple meta-analyses. Using I^2 in this way strikes us as questionable unless average sample sizes are similar.

The second approach directly estimates the variability in population effect sizes. It is generally assumed that population effect sizes relating to a given phenomenon follow a normal distribution; tau (τ) refers to their standard deviation (Borenstein, Hedges, Higgins, & Rothstein, 2009), and can be calculated when individual study effect sizes and standard errors are available. As an example, consider the meta-analysis in the right panel of Figure 1. The standard deviation of the observed effect sizes in the primary studies is 0.36. (For the sake of consistency, we use Cohen's *d* as a measure of effect size in this example and throughout our paper.) Some of the observed effect size variability must be down to sampling error. When this is removed, heterogeneity is estimated as only 0.11. To better understand heterogeneity, reconsider Figure 2. Here, the mean for the population of true effect sizes is $\delta = 0.45$ and their standard deviation is 0.33, therefore $\tau = 0.33$. The standard deviation roughly reflects how far data points are, on average, away from the mean. Consequently, any study's population effect size will typically deviate from 0.45 by approximately 0.33; also, the 95% credibility interval ranges from -0.20 to 1.10, estimating that 95% of all population effect sizes fall into this bracket (Hunter & Schmidt, 2004). As τ is independent from *N*s in primary studies (which differ markedly between meta-analyses) and τ, unlike I^2, is on an equal interval scale which allows meaningful computations of means, we use it here[2].

Being an unknown population parameter, τ has to be estimated. Its estimator T often comes with considerable uncertainty, especially when a meta-analysis is based on few studies (e.g. Huedo-Medina et al., 2006). For individual meta-analyses this can be a serious issue, especially when heterogeneity is wrongly estimated to be zero (Chung, Rabe-Hesketh, & Choi, 2013). However, this is less of a concern for us here, because our focus is not on individual meta-analyses but on aggregates of 50 or more, and we do not believe there is

[2] Indeed, across 141 meta-analyses based on Cohen's *d*, I^2 and T correlated only moderately at $r = .39$ (van Erp, Verhagen, Grasman, & Wagenmakers, 2017).

reason to suggest that heterogeneity estimates will be consistently biased in one direction (see Hönekopp and Linden, 2019).

A Sensible Sampling Frame

What is a sensible sampling frame for a survey of heterogeneity? One potential strategy would be to use a representative sample of meta-analyses across all of psychology. However, our heterogeneity measure *T* is not suitable for odds-ratios and similar effect sizes, which are frequently used in clinical psychology. Consequently, a sample of meta-analyses amenable to our heterogeneity measure would fail to be representative. We therefore decided to focus on a number of sub-fields instead in which effect size measures *d* and *r*, for which our heterogeneity measure works, predominate. We chose cognitive, social, and organizational psychology because they differ in their relative emphasis on fundamental versus applied research and because they were the focus of recent meta-scientific inquiry (Mitchell, 2012; Open Science Collaboration, 2015). Mitchell (2012) compared effect sizes from laboratory-based and field-based studies in organizational and social psychology. The correlation between lab- and field-based effect sizes was higher in the former ($r = 0.89$) than in the latter ($r = 0.53$). Open Science Collaboration found that cognitive psychology findings replicated substantially better than social psychology findings. These observations could point to greater heterogeneity within social psychology in general. Indeed, Stanley et al. (2018) argued that the low replication rate observed in the Open Science Collaboration might reflect low power caused by high heterogeneity. From this perspective, the difference in observed Open Science Collaboration replication rates would then suggest higher heterogeneity in social psychology as compared to cognitive psychology. We test this idea here.

Aims

To summarize, our aims are as follows. Given its intrinsic value as an indicator of a lack of understanding, we seek to establish typical levels of heterogeneity in conceptual replications. We compare these across the sub-fields of cognitive, organizational, and social psychology, and against heterogeneity observed in close replications. We also investigate to what extent heterogeneity in a set of studies can typically be accounted for by moderators. We further explore if any characteristics explain differences in heterogeneity.

To foreshadow our key results, we find that heterogeneity tends to be very large in conceptual replications but moderate in close replications. Our investigations regarding the drivers of heterogeneity show that moderators do little to account for heterogeneity. We also find a previously unexplored strong relationship between heterogeneity and effect size, which allows us, for the first time, to make predictions about expected levels of heterogeneity for a given phenomenon. These findings have clear implications for the improvement of our collective research practice, as we discuss at the end of our paper.

Method

Study Search and Selection Strategy

We intended to investigate all available Many-Labs type replications. We searched Curate Science (2017) for relevant reports in April 2017 and added studies from Many Labs 2 (Klein et al., 2018) at a later stage. Further, we thought to investigate 50 meta-analyses each from cognitive, organizational, and social psychology. Feasibility, rather than power considerations, determined this choice. Our pre-registered study protocol is available at: http://aspredicted.org/blind.php?x=bf46k8.

We searched PsycINFO, journals only, for “meta-analy*” in the abstract field, in November 2016. We restricted searches to PsycINFO classifications “3000 Social Psychology”, “3600 Industrial and Organizational Psychology”, and for cognitive psychology

"2340 Cognitive Processes", "2343 Learning and Memory", and "2346 Attention". Not enough eligible meta-analyses could be obtained in this way (see below for inclusion criteria). We therefore searched Web of Science, articles only, for "meta-analy*" in the categories "Psychology Social", "Psychology Applied", and "Psychology", excluding meta-analyses that fell outside our target sub-disciplines (see Figure 3). These were inspected in random order until we reached the desired number of 50 meta-analyses.

Inclusion criteria. For the three sub-disciplines, meta-analyses were included if they met all of the following criteria. i) It addressed a substantive psychological effect (rather than, for example, the psychometric properties of a questionnaire). ii) The analyzed effects were described as standardized mean differences (Cohen's d, Hedge's g) or correlations (Pearson's r, or Fisher's Z). Standardized differences and correlations are closely related concepts, and one can easily be converted into the other. Similar conversions are less sensible were categorical dependent variables were used (Ferguson, 2009), and our heterogeneity measure T is also not suitable for these types of effect sizes. For this reason, we excluded meta-analyses that used odds ratios, risk ratios, and similar measures. iii) Effect size and sample size information was provided for the original studies. This was necessary to calculate heterogeneity. Where only the sample sizes or effect sizes were available, an attempt was made to obtain missing data from the corresponding author. iv) For practical reasons, the full article had to be available in English. All close replication reports that met the same criteria (Many Labs 1, 2 and 3, and Registered Replication Reports 3 to 6) were included (Cheung et al., 2016; Ebersole et al., 2016; Eerland et al., 2016; Hagger et al., 2016; Klein et al., 2014; Klein et al., 2018; Wagenmakers et al., 2016)[3].

---

[3] For Many Labs 2 (Klein et al., 2018), effect sizes for individual studies were not reported, but we could compute them from the published raw data.

In this way, we identified 50 meta-analyses for cognitive, organizational, and social psychology each, and 57 for close replications (see Supplementary Table 1).

Data extraction and analysis. Where the results of more than one meta-analysis were reported, the one including the largest number of studies was extracted. If multiple meta-analyses included the same number of studies, the first was used.

Heterogeneity for each meta-analysis was computed using the DerSimonian-Laird estimator in Metafor in R (Viechtbauer, 2010)[4]. In order to keep effect sizes and levels of heterogeneity consistent across studies, all effect sizes were input as Cohen's *d*. All other effect sizes were converted accordingly.

It turned out that the frequency distributions for some of our observed outcome variables were right skewed. For example, among the 150 meta-analyses, T had a skewness 0.99 (the largest Z score being 3.57). We therefore report winsorized means (M_{win}), and respective standard deviations (SD_{win}). Winsorizing replaces the smallest and largest values in a distribution (in this case the smallest 10% and largest 10%) with the observations that are closest to them. Where frequency distributions are normal in nature, this will not alter the results. M_{win} therefore removes the undue effect of outliers, but retains much more information than the median, which trims all scores but the one in the middle of the distribution (Erceg-Hurn & Mirosevich, 2008). Specifically for T, M_{win} should also counteract the likely overestimation resulting from setting negative heterogeneity estimates to $T = 0$. We use the Yuen-Welch method (Wilcox, 2005), which is similar to the t-test but based on M_{win}, for group comparisons of T. Similarly, we use winsorized correlations, r_{win} (Wilcox, 2005).

[4] We also computed analyses using the widely used Hunter-Schmidt, Paule-Mandel, and REML estimators. These led to similar results (see Supplementary Table 2) and the same conclusions.

These limit the effect of outliers, but retain more information than Spearman's rank-based correlation, r_s. All data and further materials can be found at https://osf.io/yr3xd/.

Results

How Meta-Analyses Address Heterogeneity

Out of 150 meta-analyses, 123 tested moderators, but only 83 (55%) reported a measure of heterogeneity. In 2009, the influential PRISMA guidelines (Preferred Reporting Items for Systematic Reviews and Meta-Analyses, Moher, Liberati, Tetzlaff, & Altman, 2009) recommended that meta-analyses should address heterogeneity. Even in post-2009 meta-analyses, heterogeneity was only reported in 60% of cases. Interestingly, statistical significance of heterogeneity, e.g. Q, was widely reported (77 times); of those, 45% did not report quantifying information. This focus on statistical significance and neglect of quantifying information runs counter to the meta-analysis estimation perspective (Hunter, 1997).

Overall, heterogeneity was quantified in less than a third of cases (43 times out of 150): I^2 was reported in 33 cases, T^2 in 9, and another measure was reported once. In addition to the observed neglect of quantification, it is interesting that authors unanimously reported T^2 (the heterogeneity variance) instead of T (the standard deviation). Whereas standard deviation has a meaning that is comparatively easy to grasp (it approximates the average difference from the mean), variance does not have a similarly accessible interpretation. (This is why researchers most commonly report standard deviations, and not variances, in their descriptive statistics. Similarly, we can picture standard deviations in Figure 2, but would be lost with variances.)

Heterogeneity Observed in Close Replications and Meta-Analyses

Table 1 shows descriptive statistics for close replications and meta-analyses. As expected, heterogeneity was much lower (M_{win} = 0.09) in close replications than in the meta-analyses (M_{win} = 0.33), t(94.9) = 10.43, p < .001 (see Figure 4).

Contrary to our hypothesis, levels of heterogeneity were very similar across all three sub-disciplines (cognitive versus social psychology: t(57.3) = 0.33, p = .370 (1-tailed); social versus organizational: t(57.0) = 1.17, p = .125 (1-tailed); organizational versus cognitive: t(54.9 = 0.74, p = .463 (2-tailed).

Moderators

In order to investigate to what extent moderators account for heterogeneity, we looked at all 36 meta-analyses with $k \geq 60$, because moderators will be most reliably identified in large meta-analyses. We looked only at meta-analyses where moderators were reported by the original authors. Where possible, we used the strongest moderator for which sufficient information was reported for further analyses. All moderators thus identified were grouping variables (i.e. none was continuous). We used these moderators to partition studies into appropriate subsets, which left us with 22 meta-analyses. We then excluded broad subsets. For example, Baker, Peterson, Pulos, and Kirkland, R. A (2014) looked at intelligence and “Reading the Mind in the Eyes”. The strongest moderator they examined was the type of intelligence test used. Based on this, studies were split into two subsets: IQ measured using the Wechsler IQ test, and IQ measured using any other test. We excluded the broad ‘other’ subset and compared this T against the T in the initial overall meta-analysis. Average T in the 22 subsets (M_{win} = 0.33) was very similar to average T in the corresponding 22 overall meta-analyses (M_{win} = 0.37), t(13) = 1.39, p = .187. Powerful moderators might only emerge when they are based on theoretical considerations (Tipton, Pustejovsky, & Ahmadi, 2019a, 2019b). We therefore looked at those 10 out of the 22 meta-analyses that presented a theoretical rationale for the moderator. Again, we used these moderators to split the studies from each meta-analysis into appropriate subsets and repeated the previous analysis. We found that average T in the 10 moderator-based subsets (M_{win} = 0.37) was again very similar to average T in the 10 corresponding overall meta-analyses (M_{win} = 0.37; $t(5)$ = 0.73, p = .499).

This moderator analysis does not suggest that the large heterogeneity in our meta-analysis sample is readily explained by mixing apples and oranges. Still, the possibility remains that authors (potentially unwisely) combine highly diverse studies and then fail to address relevant moderators. In order to address this point, we rated (on a single, global five-point scale, from ‘low’ to ‘high’) for each meta-analysis how broad or narrow its inclusion criteria were. Ratings considered to what extent the addressed question was broad (e.g. ‘How effective is psychotherapy?’) versus narrow (e.g. ‘How effective is cognitive behavioral therapy to treat simple phobias?’); to what extent the manipulation of the IV and the measurement of the DV followed a standard protocol; and the similarity of the samples included. Ratings were conducted by the second author without knowledge of the actual levels of heterogeneity; to establish reliability, a random sample of 30 meta-analyses were independently re-rated by the first author. We computed inter-rater agreement as Cohen’s kappa using quadratic weights and observed κ_w = 0.73, which is typically interpreted as good (Jakobsson & Westergren, 2005). For the 58 meta-analyses whose broadness of inclusion criteria was rated ‘low’ or ‘low-to-medium’, average heterogeneity was still very high (M_{win} = 0.29). In other words, if authors generally avoided meta-analyses that integrate fairly diverse studies, levels of observed heterogeneity would probably not be much lower. In sum, our analyses do not support the view that unwise mixing of apples and oranges is a strong driver of observed heterogeneity in meta-analyses.

Exploratory Analyses on what Drives Heterogeneity

Heterogeneity differed substantially between meta-analyses (SD_{win} = 0.11). This begs the question, why? A number of ideas have been proposed that we can test here. Kenny and

Judd (2019) suggested that research areas with larger average effect sizes should have greater levels of heterogeneity. We therefore correlated mean *d* with *T*. In support of this idea, we found a strong correlation $r_{\text{win}} = .49$ ($p < .001$) for the set of 150 meta-analyses (see Figure 5). This replicated across all three sub-disciplines (cognitive: $r_{\text{win}} = .34$, $p = .019$; social: $r_{\text{win}} = .52$, $p < .001$; organizational: $r_{\text{win}} = .62$, $p < .001$). The relationship between mean *d* with *T* also held for the set of 57 close replications, $r_{\text{win}} = .48$ ($p < .001$), see Figure 5. The observed relationship between *d* and *T* might, at least partly, arise from differences in participants' motivation across studies. When the treatment has a strong effect, it should typically make a big difference whether participants engage well with the study or not; when the treatment has only a small effect, such differences should be less consequential (Weiss et al., 2017).

In light of the link between mean *d* and *T*, the direct comparison of heterogeneity in close replications versus meta-analyses might appear doubtful. This is because mean *d* proved considerably lower in close replications (0.31) than in meta-analyses (0.47). Correction via the relevant regression equation (Figure 5) suggests an average *T* of 0.27 for meta-analyses at $d = 0.31$ (the average for close replications). This is still much larger than that observed for close replications ($M_{\text{win}} = 0.09$).

Looking at systematic reviews in healthcare, IntHout et al. (2015) found that small studies are more heterogeneous (measured as T^2) than large ones. We therefore worked out the median sample size for each meta-analysis and correlated this with *T*. This resulted in $r_{\text{win}} = -.10$ ($p = .108$, 1-tailed), not suggesting an important role for average sample size.

Richard et al. (2003) proposed that as a research field matures, the focus shifts from establishing an effect to exploring its boundaries, and this should increase heterogeneity in findings. If we accept the number of studies (*k*) as a proxy for the maturity of a research field, we should expect a positive correlation between *k* and *T*. In line with this idea, they found a correlation of $r = .11$ in a large survey of meta-analyses in social psychology. For our 150 meta-analyses, we found r_{win} = .23 (p =.005) in support of this idea. An obvious alternative interpretation is that meta-analyses with broader inclusion criteria cast a wider net and will therefore include more studies than those that use narrow inclusion criteria (Murphy, 2017).

We thought to test these two competing explanations (exploring boundaries versus broader inclusion criteria). If later research into an effect tends to explore its boundaries, we would expect to see higher heterogeneity in studies conducted late than in those conducted early. We therefore looked at all meta-analyses that seemed to capture a sufficiently mature research area, and included those 82 with k ≥ 30 and a time range for included studies of ≥ 10 years. For each of these, we then determined T separately for the earlier and the latter half of the included studies. Although the difference was in the expected direction (early: M_{win} = 0.336, SD_{win} = 0.097; late: M_{win} = 0.342, SD_{win} = 0.110), it was small and not statistically significant ($t(49) = 1.01$, p = .319), therefore not supporting the idea of boundary exploration.

If the observed correlation between k and T is down to broader inclusion criteria, we would expect to see that meta-analyses with broader inclusion criteria show higher T. We therefore correlated our ratings of the broadness of inclusion criteria with T and found r_{win} = .12 (p = .142). This does not offer strong evidence that the observed correlation between k and T reflects broadness of inclusion criteria. In sum, our data do not offer a clear explanation why T tends to increase with k.

Discussion

We found that quantification of heterogeneity in meta-analyses is uncommon. Where it is undertaken, authors rarely rely on the measure that we argue is most informative. Average heterogeneity proved T = 0.33 for meta-analyses, with powerful (or even decent) moderators being conspicuously absent, and T = 0.09 for close replications. Heterogeneity showed a strong positive association with average effect size. Although based on exploratory analyses, strong consistency across all three sub-fields and close replications lends credence to this finding.

Recently, the effect of heterogeneity on statistical power and its implications for the interpretation of low replicability rates in the OSC (2015) project received considerable attention (Shrout & Rodgers, 2018; Stanley et al., 2018). We address these more specific issues here. We discuss more general implications for the progress of psychological science in the General Discussion.

The Meaning of Observed Heterogeneity Levels

It is helpful to first consider the meaning of average heterogeneity levels (T = 0.09 for close replications and T = 0.33 for meta-analyses). As we discussed earlier, this partly depends on the effect size, which averaged d = 0.31 and d = 0.47 (winsorized means for close replications and meta-analyses, respectively)[5]. The average result for meta-analyses is depicted in Figure 2 (solid line). Remember that Cohen's d of 0.2/0.5/0.8 are often considered as benchmarks for small/medium/large effects. All of these occur frequently in the distribution of population effect sizes for the average meta-analysis. Therefore, observed heterogeneity in conceptual replications appears large. In contrast, the average close replication showed moderate variability in population effect sizes.

We further illustrate the meaning of heterogeneity with two examples from cognitive psychology. With the importance of working memory for second language proficiency development and processing being debated, Linck, Osthus, Koeth, and Bunting (2014) investigated the strength of this link in a meta-analysis. Included studies used a range of

[5] We note that our value for meta-analyses corresponds closely to the average effect size d = 0.39 observed in a sample of meta-analyses from *Psychological Bulletin* (Stanley et al., 2018).

working memory tasks and second language comprehension measures in diverse samples. The strength of the relationship proved medium in size (d = 0.51), and heterogeneity was estimated to be low (T = 0.11) (see Figure 1, left panel). The latter implies high consistency of the relationship and ready generalizability across paradigms. In line with this, most population effect sizes (mean ± 1SD) should fall in a narrow range of medium sized effects (d = 0.40 to 0.62).

Baker, Peterson, Pulos, and Kirkland (2014) used meta-analysis to investigate the degree of independence between general intelligence and mental state understanding. Included studies used a range of established intelligence tests in diverse samples, however, all used the same widely used test of mental state understanding (Reading the Mind in the Eyes Test). As in the previous example, the strength of the relationship proved medium in size (d = 0.49), however, heterogeneity was estimated to be much higher (T = 0.35) (Figure 1, right panel), even though the same test of mental state understanding was used throughout. Even though the observed level of heterogeneity was medium and not large, it already implies low consistency of the studied relationship and a lack of generalizability across paradigms. Most population effect sizes (mean ± 1SD) should fall in a wide range of very small to large effects (d = 0.14 to 0.84). The authors reported a statistically significant moderator, but given its a-theoretical nature, and the number of moderators tested, it remains debatable if this reflects progress in understanding, or successful capitalization on chance (Ioannidis, 2008).

In sum, it appears that the relationship between working memory and second language proficiency is better understood than that between intelligence and performance on the Reading the Mind in the Eyes Test. More generally, everything else being equal, meta-analyses with lower heterogeneity will be more informative.

Potential Biases in Heterogeneity Estimates

Before we address implications of these findings in greater detail, it is necessary to highlight a number of points regarding the trustworthiness of our estimates.

Representativeness of our samples. Our sampling of meta-analyses in cognitive, organizational, and social psychology was rigorous, and perusal of the topics (see Supplement Table 1) confirms a broad coverage of topics typical for these sub-disciplines. We did not find evidence for heterogeneity differences across these sub-fields, which might indicate that our results generalize more broadly across psychology. This is supported by the fact that Stanley et al. (2018), in a broader sample of meta-analyses from *Psychological Bulletin*, found heterogeneity levels (median T = 0.35) very similar to ours (T = 0.33). In contrast, representativeness cannot be claimed for our sample of close replications. Owing to the novelty of the concept and the enormous effort involved, such Many-Labs type studies are relatively rare, focus on effects that are relatively easy to study, and consequently the set of original studies that motivated these replications cannot be considered to be representative of the three sub-disciplines we studied or psychological research in general. The observed difference in average effect size (close replications: d = 0.31, meta-analyses: d = 0.47) reinforces this point. It therefore remains unknown how readily the low heterogeneity observed in close replications would generalize to psychological research findings in general.

Publication bias and questionable research practices. Publication bias (Sterling, 1959) and questionable research practices (QRPs, Simmons et al., 2011) are problems in the psychological research literature (John, Loewenstein, & Prelec, 2012; McShane, Böckenholt, & Hansen, 2016; Simmons, Nelson, & Simonsohn, 2011; Sterling, 1959). Under publication bias, only a biased sample of all conducted studies appears in the published literature; ‘unsuccessful’ studies typically remain invisible. Given that larger as compared to smaller observed effects are more likely to be statistically significant (and thus ‘successful’), publication bias leads to upwardly biased effect sizes in published studies and meta-analyses (e.g., McShane et al., 2016). In order to achieve statistically significant, and therefore publishable, results, researchers might resort to QRPs (e.g., collect a number of similar dependent variables, but only report findings from the most ‘successful’ one). QRPs can dramatically increase the rate of false-positive results (Simmons et al., 2011) and thus lead again to inflated effect sizes in published studies and meta-analyses.

Mathematical modelling and computer simulations suggest that publication bias can lead to under- or overestimation of heterogeneity, however underestimation tends to be more prevalent than overestimation (Augusteijn, van Aert, & van Assen, 2019; Jackson, 2006). Also, the overestimation of heterogeneity due to publication bias and QRPs tended to be much smaller than the levels of heterogeneity observed in conceptual replications here (Hönekopp and Linden, 2019). From this viewpoint, the very large T in conceptual replications cannot be attributed to bias, but instead seems to represent real heterogeneity that is not well understood. Our heterogeneity estimates for close replications are not affected in this way, because the protocol for Many-Labs type replications precludes publication bias and QRPs (e.g., Shrout & Rodgers, 2018).

Overreliance on WEIRD samples. Studies in psychology, even if they seek to address human nature in general, rely almost exclusively on samples from Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies. Henrich, Heine and Norenzayan (2010) argued that WEIRD samples are among the least suitable to make general inferences about human nature, and that many phenomena that are well-established in WEIRD populations fail to generalize to other populations. Obviously, this is only a concern where studies are aimed at human nature, however this is frequently the case. For these cases, Henrich et al.’s findings imply that observed heterogeneity would often be higher if researchers did not rely almost entirely on WEIRD samples. This should hold equally for conceptual and close replications.

Accuracy of meta-analyses. For feasibility reasons, we had to rely on reported effect sizes for the underlying primary studies. One systematic investigation of meta-analyses in medicine found that about 1 in 5 effect size computations for primary studies were erroneous (Gøtzsche, Hróbjartsson, Marić, & Tendal, 2007). This should add (error) variance to the meta-analysis, and consequently inflate observed heterogeneity. Given the strict protocols and high degree of transparency for Many-Labs type studies (e.g. Klein et al., 2014; Klein et al., 2018), erroneous effect sizes should be less of a concern for close replications.

Summary. In sum, our data for cognitive, organizational, and social psychology should be fairly representative for these disciplines, and results might generalize fairly well beyond. Publication bias, QRPs, and overreliance on WEIRD samples should artificially lower heterogeneity estimates; meta-analytic errors regarding the extraction of effect sizes from primary studies should have the opposite effect. On balance then, there is no strong evidence to suggest that our very high heterogeneity estimates grossly overestimate actual levels of heterogeneity. If anything, heterogeneity-deflating biases appear more numerous than heterogeneity-inflating biases. Consequently, our results suggest that actual heterogeneity is typically very high in sets of conceptual replications. While the representativeness of our close-replications sample is unclear, resulting heterogeneity estimates should, overall, be less error prone than those for conceptual replications.

Implications for the Replicability of Close Replications

OSC (2015) famously attempted close replications of 100 recent studies. Although using larger samples than the original studies, statistical significance was achieved in only 36% of replications (25% in social psychology and 50% in cognitive psychology). This finding has become a catalyst of the controversial debate about the health of psychology research, which is still ongoing (e.g., Earp & Trafimow, 2015; Pashler & Harris, 2012; Schmidt & Oh, 2016; Simons, 2014; Stroebe & Strack, 2014). This is not the place to review this debate (see Zwaan et al., 2018, for a comprehensive summary), but one of its strands is of particular interest here. Stanley et al. (2018) suggested that heterogeneity accounts for OSC's low replication rates. Based on only two Many-labs type close replications, the authors estimated heterogeneity to be high (T = 0.25). Subsequent power calculations demonstrated that heterogeneity should therefore decrease power in the typical psychological study to levels that are in line with the low replication rate observed in OSC (2015). However, heterogeneity of the order we observed in a much larger sample for close replications (mean T = 0.09) reduces statistical power only marginally[6]. If we stick to sample sizes that generate 80% power at zero heterogeneity, power does not drop at all for large effects, drops to 78% for medium effects, and to 70% for small effects, as McShane and Böckenholt (2014) showed. The mean effect size for the original studies included in the OSC (2015) was large (d = 0.87)[7]. Therefore replication power should not be greatly affected, provided that the differences between the OSC replication studies and their original counterparts is comparable to the differences in multiple close replications.

Moreover, if replication failure reflects heterogeneity-driven low power, as Stanley et al. (2018) claimed, the large difference in replication rates between cognitive and social psychology (Open Science Collaboration, 2015) should be reflected in larger heterogeneity in the latter. Our finding of virtually identical heterogeneity levels across cognitive and social

[6] Stanley et al. (2018) based their heterogeneity estimate for close replications on Eerland et al. (2016) and Hagger et al. (2016), both of which were part of our much larger sample.

[7] The mean effect size of the original studies underlying the 57 Many-labs type close replications was d = 0.75 (SD = 0.37). Regarding the effect size of the underlying original studies, Many-labs type close replications and OSC replications are therefore comparable. This matters because of the observed link between T and mean effect size (see Figure 5).

psychology does not support this view. In conjunction with the low heterogeneity observed in close replications, it strengthens the interpretation that the low replication rate demonstrated in OSC might be attributable to publication bias and QRPs. In our eyes, this is good news because promising strategies to combat these biases have been developed (Munafò et al., 2017). On a more general level, one may note that the central issue with the OSC results is less about the percentage of original results that are true; more importantly, they suggest that a key plank in our common standards to accept evidence as valid ($p < .05$) has little utility (Sedlmeier & Renkewitz, 2018, p. 621).

How Should Heterogeneity be Estimated for Power Calculations?

Average levels of heterogeneity ($T = 0.33$) have quite dramatic effects on power: it drops from 80% to 69% for large effects, from 80% to 63% for medium effects, and from 80% to 56% for small effects (McShane & Böckenholt, 2014). What level of heterogeneity should we expect for a new study that is not a close replication? This is an important question for proper sample size planning. The researcher's informed judgment will always be necessary, however the following suggestions might appear sensible. Where a relevant meta-analysis reports *T*, use this. Where such a meta-analysis only reports the effect size, use $T = 0.18 + 0.30d$ (see Figure 5) to estimate heterogeneity. Where there is no meta-analysis, the heterogeneity can still be estimated (although with lower precision) from the effect size of a single study. When we used all effect sizes from all 150 meta-analyses in our sample to predict heterogeneity, $T = 0.28 + 0.11d$ was the resulting regression ($R = .38$). Finally, when an effect size estimate is not available, use the mean ($T = 0.33$).

General Discussion

We suggested that heterogeneity is a useful perspective to reflect the degree of understanding psychology achieves. Science can be described as a quest to explain the apparent complexity of the natural world through simpler, fundamental principles. Empirical cumulativeness reflects the extent to which empirical findings fit such a simple or explicable pattern. Ceteris paribus, high levels of (unexplained) heterogeneity indicate lower empirical cumulativeness (Asendorpf et al., 2013; Hedges, 1987; Murphy, 2017; Richard et al., 2003; Sells, 1963). For conceptual replications in three of psychology's core disciplines (and plausibly beyond, cf. Stanley et al., 2018; van Erp et al. 2018), we found that heterogeneity is typically large (see Figure 2) and unexplained, with little reason to believe that our estimates are inflated. To add some perspective, we can compare typical levels of heterogeneity (variability within a specific topic) with the variability in mean effect sizes across meta-analyses (variability between topics). Whereas we found T = 0.33 for the former, for the latter we observed SD = 0.42 across all 150 meta-analyses. In other words, variability within phenomena measured in this way is only about 20% less than variability between phenomena. This large heterogeneity is sobering, as it reflects low empirical cumulativeness and therefore low coherence between the concepts researchers employ and the data observed. On a brighter note, our findings also showed that large heterogeneity is not inevitable – in close replications, it was typically of moderate magnitude (T = 0.09) – and even hard sciences face some heterogeneity in their measurements (Hedges, 1987).

Before we explore important implications of this twin finding and possible improvements for our collective research practice, we address a likely objection to our argument that heterogeneity meaningfully reflects the degree of understanding psychology achieves.

Reply to an Objection

A likely objection is that progress is driven by theories, and that effect sizes tend to be irrelevant for most psychological theories (e.g. Baumeister, 2016; Strack, 2017); where effect sizes are largely irrelevant, their variability (i.e. heterogeneity) is likewise of little consequence. We think that such a perspective is mistaken for a number of reasons. First, even if effect sizes were largely irrelevant, the direction of effects remains of importance: in the face of large heterogeneity, the direction of an effect might be difficult to predict. Second, effect sizes are by no means irrelevant for progress of understanding, and therefore their degree of variability is important, too. Although some psychological theories are not rooted in quantitative concepts (e.g. Piaget's stages in cognitive development), most psychological research is rooted in measurement. Given that measurement is regarded as a practically indispensable tool for investigation, it seems inconsistent to be disinterested in its result. In general, strong theories tend to be specific in the sense that they declare a large range of potential observations to be contrary to theory, thereby creating ample scope for the theory to be empirically challenged (Kuhn, 1970). Similarly, the ability to make precise predictions is often a hallmark of more mature science (Schickore, 2018). Effect sizes are obviously not the only route to achieve such specificity, but often they might provide a viable way forward. Where heterogeneity is high, such specificity is difficult to achieve.

Finally, effect sizes are highly relevant for both explanations and practical applications. Psychological explanations typically rely on probabilistic relationships – for example in mate choice, men *tend* to put more emphasis on a partner's physical attractiveness than women do (Feingold, 1990) – and, ceteris paribus, stronger effects convey better explanations (Woodward, 2003). For example, the sex difference in height (approximately d = 2, (cf. Gustafsson & Lindenfors, 2009)) is much stronger than the sex difference in relevance of attractiveness for mate choice (approximately d = 0.5). Consequently, "because she is a woman" is more suitable to explain why Aminah is shorter than Muhammad than to explain why physical attractiveness matters less in her mate search than in his. Where heterogeneity is large, it becomes unclear how powerful particular explanations are, which is obviously undesirable. Similarly, effect sizes are also highly relevant for practical applications. For example, sleep quality is a particularly strong predictor of adolescents' well-being, and for this reason it is a particularly promising lever for improving young people's wellbeing (Gireesh, Das, & Viner, 2018). Again, where heterogeneity is large, the effect of any intervention, which might be thought of as another conceptual replication, becomes more difficult to predict; and unless the average effect size is large, even the direction of the effect of the intervention could be uncertain (Figure 2). In line with this, successful interventions can rarely be delineated from research findings but need to be tested (Cowen, Virk, Mascarenhas-Keyes, & Cartwright, 2017).

Implications for Testing Theories

Our twin finding of large heterogeneity in conceptual replications and moderate heterogeneity in close replications has important implications for the testing of theories.

Knowledge as a tool. One relates to the use of knowledge as a tool. Imagine a situation where the test of a psychological theory X requires induction of a particular mood. Where this is based on a general principle that shows large heterogeneity, a negative finding of the test can be blamed on (unreliable) methods and theory X is protected from failure. Where heterogeneity thus precludes the meaningful empirical scrutiny of theories, theoretical progress will be limited (Ferguson & Heene, 2012; Greenwald, 2012; Kerr, 1998; LeBel & Peters, 2011; Meehl, 1978). In this context, the moderate heterogeneity observed in close replications (T = 0.09) is encouraging, and it has a clear implication: a test of theory X should not rely on a general principal of mood induction but closely stick to a particular, successful protocol. This should typically bring about the expected change in mood[8]. In this context, we

[8] Note, however, that our heterogeneity estimate for close replications stems from pre-registered studies published irrespective of their results. This precludes distortions of their effect sizes through publication bias and QRPs, which might often affect other published

note that any systematic investigation about which psychological effects are particularly reliable (i.e. strong and with low heterogeneity) is curiously absent.

Theories' boundaries. Another implication of our findings is that evaluation of theories also requires a broad exploration of the 'research space' (Asendorpf et al., 2013), i.e. the space defined by the combination of different manipulations of the independent variable, different dependent variables, different study populations, etc. As an example, consider the set of stimuli employed. If only a single standard set is used in a research domain to evoke the expected effect, some theory-irrelevant feature of that set might drive the observed effect (Fiedler, 2011). This problem can only be detected by using diverse (but theory-conforming) sets of stimuli. Also consider the case where a theory offers a narrow explanation to account for an observation (e.g. memory for a word list is improved when the survival value of its items is to be judged). If a more general and thus more parsimonious explanation holds (e.g. memory for a word list is improved by any judgments that trigger self-referent encoding), this can only be discovered by testing instances of the research space that violate the overly narrow theory while still holding for the more general account (Fiedler, Kutzner, & Krueger, 2012; Shrout & Rodgers, 2018).

Meta-analysis and the testing of theories. A good theory should specify its scope. To evaluate the theory, meta-analysts must move beyond a narrow focus on the mean effect size and its statistical significance and take heterogeneity into account. Obviously, this is not a new insight (e.g. Higgins & Thompson, 2002; Hunter & Schmidt, 1990). However, our results regarding the reporting of heterogeneity in meta-analyses suggest this is rarely

---

results (Ferguson & Brannick, 2012; Kühberger et al., 2014; LeBel & Peters, 2011; Levine et al., 2009; Pashler & Harris, 2012; Sterling, 1959). Consequently, close replications based on a single published result might be less reliable (OSC, 2015).

implemented in practice. One reason might be that frequently used approaches to heterogeneity fail to appeal to researchers' imagination: as we saw, quantification of heterogeneity is often missing, or else it is expressed in ways that might elude intuitive understanding (I^2, T^2). An increased focus on T might facilitate thinking about heterogeneity.

Reducing Unexplained Heterogeneity as a Sensible Heuristic to Advance Understanding

Given that unexplained heterogeneity tends to be both large and undesirable, its reduction should become an important goal. Among other advantages, this will increase coherence between the concepts we use and our observational data; facilitate empirical scrutiny of our theories; provide greater clarity regarding the power of the explanations we can offer; and facilitate the design of practical applications. Weiss, Bloom, and Brock (2014) offered a conceptual framework for heterogeneity in experiments, which is useful to discuss measures to either explain or reduce it.

A conceptual framework for heterogeneity. From this (perspective Weiss et al., 2014), heterogeneity in a set of experiments arises from three sources. First, studies can differ in their treatment contrasts, i.e. the experimentally induced difference between experimental and control group. The second source of heterogeneity are moderators that reside in the participants. Thus, if an effect is age-dependent, differences in participants' age across studies will induce heterogeneity. Finally, studies might differ on relevant context moderators, e.g. an effect might vary across cultures or situations. Fruitful applications of this framework can be found in Weiss et al. (2017).

Treatment contrasts. Differences in studies' treatment contrasts will typically be driven by the strength of experimental manipulations. Often, stronger manipulations will bring about stronger effects than weaker manipulations. Consequently, variability in the strength of manipulations across studies will induce heterogeneity in results. When the strength of manipulations cannot be (or is not) properly expressed, it will be difficult to explain this heterogeneity. Un- or underspecified strength of experimental manipulations strikes us to be a frequent issue across psychology, which could often be avoided. We take the effect of bilateral symmetry on facial attractiveness as an arbitrary example. Correlational studies and experiments alike suggest that symmetry increases facial attractiveness (Rhodes, 2006). If the strength of experimental symmetry manipulations were described in relation to the natural variation in facial symmetry on which correlational studies rely, the variability of symmetry could be described on a common scale across all studies. These between-study differences in variability of symmetry (whether naturally occurring or experimentally induced) should be able to explain differences in results across studies and thus reduce heterogeneity. We are not aware of such attempts.

Our suggestion that systematic specification of the strength of manipulations of the independent variable will prove helpful is underpinned by the observation that many seminal insights in behavioral science relied on description of the independent variable on a ratio scale. This is true for probabilities in classical conditioning (Rescorla & Wagner, 1972), operant conditioning (Herrnstein, 1961), perception under uncertainty (Tanner & Swets, 1954), and judgements and decision making under uncertainty (Gigerenzer, Hoffrage, & Kleinbölting, 1991; Kahneman & Tversky, 1979); for the temporal relationship of stimuli or events and their effects on visual perception (Marcel, 1983), memory (Peterson & Peterson, 1959), and the discounting of future outcomes (e.g. Frederick, 2002); the physical stimulus intensity and its relationship with perceived stimulus intensity (Stevens, 1957); and for degrees of genetic similarity, which underpin all estimates of the heritability of psychological traits (Plomin, 1990).

Importantly, differences in studies' treatment contrasts can also be affected by differences in the control groups, particularly in the case of real-world interventions. For these, "business as usual" (i.e., what it means *not* to be assigned to the intervention) will often differ in important ways between studies (Weiss et al., 2014; Weiss et al., 2017). For example, an intervention promoting healthy behavior by providing information about health risks might create only a small treatment contrast in an environment in which ample information on health risks is at hand, but a large treatment contrast in an environment in which such information is scarce.

Person and context moderators. The experimental test of a motivational intervention conducted by Yeager et al. (2019) provides an excellent illustration for both, person and context moderators. Their short online intervention taught a nationally representative sample of US students in secondary education that they can train their intellectual abilities similar to a muscle, which proved to have a small positive effect on students' grades. Importantly, the authors hypothesized and confirmed that low achieving students would benefit more from the intervention than high achieving students (person moderator) and that the intervention would be most effective in schools with supportive peer norms (context moderator).

Meta-analytic search for moderators is most promising when theory-driven (Tipton et al., 2019a, 2019b). In this context it is noteworthy that psychologists have devoted great energy to describing individual differences in systematic ways (e.g., McCrae & Costa Jr, 1997) but that comparable approaches to classify situations are, to the best of our knowledge, missing.

Multi-site experiments. Meta-analyses are often limited in their ability to explain heterogeneity because relevant information on moderators or other sources of heterogeneity is unavailable for some or all of its primary studies. Multi-site experiments, which directly address potential moderators in their design, are a promising alternative (e.g., Yeager et al., 2019). Such experiments are naturally arduous, but collaboration between many researchers through crowdsourcing holds great potential for such projects (Uhlmann et al., 2019).

Standardized versus original-units effect sizes. Finally, we want to draw attention to a points outside Weiss et al.'s (2014) heterogeneity framework. Our treatment of heterogeneity was based on description of individual study results using standardized effect sizes. This is the norm for meta-analyses and conveys the obvious advantage that studies can be sensibly integrated even when they use different dependent variables. Nonetheless, standardized effect sizes might not be the best way to capture study results (Baguley, 2009; Bond, Wiitala, & Richard, 2003; Tukey, 1969). Table 2 provides an example, in which differences in sample means might be said to provide a more accurate description of individual results and of their differences across studies. This increase in accuracy might lead to reduced heterogeneity estimates and to the clearer emergence of informative moderators. The wealth of available data from Many-Labs type close replication studies (in which sets of close replications share the same dependent variable) provides rich opportunities to develop heterogeneity analyses based on mean differences instead of standardized effect sizes and establish if this reduces heterogeneity estimates. If that is the case, we should investigate to what extent this can be fruitfully used for the analysis of conceptual replications, too.

Critics might argue that the portrayed shortcoming in standardized effect sizes (see Table 2) undermines our survey of heterogeneity. However, heterogeneity on the scale observed by us in conceptual replications cannot result from moderate inaccuracies in standardized effect sizes. Large heterogeneity is real and its reduction should therefore become an important aim. To judge whether we make progress on this issue and in order to learn which strategies are best suited to reduce unexplained heterogeneity, its measurement is necessary. The approach we presented here strikes us as the most appropriate currently available.

Outlook

Chemists in the 18th century, who did not yet understand the difference between compounds and mixtures, realized that substances often combine in fixed proportions (e.g. you need 61.5g magnesia to neutralize 100g of sulphuric acid; Leicester, 1965). Although useful for their daily practice, they did not attach much importance to this regularity because it appeared to lack universality (after all, you can mix one or three spoons of sugar in a cup of tea). Early in the 19th century, John Dalton parsed the seemingly incongruous observational data in a new way and realized the significance of fixed proportions, thus paving the way for measurement of relative atomic weights and atomic theory, a major breakthrough in the history of chemistry (Kuhn, 1970). The linear relationship between stars' distance from Earth and the speed at which they move away from us was probably more obvious to perceive: within a short timespan, Georges Lemaître and Edwin Hubble independently discovered this law and, consequently, the expansion of the universe (Schneider, 2014). These examples illustrate two points: (i) regularity in observational data often acts as a lodestar for discovery (Simon, 1973); and (ii) even identification of pockets of regularity might be greatly beneficial. Reduction of heterogeneity should make it easier for psychologists to perceive such regularities, and the prospect of new discoveries might be the strongest incentive to do so. We suggested some means to this end. We are sure that, once heterogeneity and its reduction receives more of the attention it deserves, the ingenuity of our colleagues will greatly add to our own ideas.

Augusteijn, H. E., van Aert, R., & van Assen, M. A. (2019). The effect of publication bias on the Q test and assessment of heterogeneity. *Psychological Methods*, *24*(1), 116-134.

Aytug, Z. G., Rothstein, H. R., Zhou, W., & Kern, M. C. (2012). Revealed or concealed? Transparency of procedures, decisions, and judgment calls in meta-analyses. *Organizational Research Methods*, *15*(1), 103-133.

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology, 100*(3), 603-617.

Baker, C. A., Peterson, E., Pulos, S., & Kirkland, R. A. (2014). Eyes and IQ: A meta-analysis of the relationship between intelligence and “Reading the Mind in the Eyes”. Intelligence, 44, 78-92.

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.

Bond Jr, C. F., Wiitala, W. L., & Richard, F. D. (2003). Meta-analysis of raw mean differences. *Psychological Methods, 8*(4), 406-418.

Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester: Wiley Online Library.

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.

Cheung, I., Campbell, L., LeBel, E. P., Ackerman, R. A., Aykutoğlu, B., Bahník, Š., . . . Caprariello, P. (2016). Registered Replication Report: Study 1 from Finkel, Rusbult, Kumashiro, & Hannon (2002). *Perspectives on Psychological Science, 11*(5), 750-764.

Chung, Y., Rabe-Hesketh, S., & Choi, I. H. (2013). Avoiding zero between-study variance estimates in random-effects meta-analysis. *Statistics in Medicine, 32*(23), 4071-4089.

Cowen, N., Virk, B., Mascarenhas-Keyes, S., & Cartwright, N. (2017). Randomized controlled trials: how can we know “what works”? *Critical Review, 29*(3), 265-292.

Curate Science. (2017). Quantifying the Trustworthiness of Empirical Research. Retrieved from http://www.curatescience.org/

Dieckmann, N. F., Malle, B. F., & Bodner, T. E. (2009). An empirical assessment of meta-analytic practice. *Review of General Psychology, 13*(2), 101-115.

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology, 6*. doi:10.3389/fpsyg.2015.00621

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., . . . Boucher, L. (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., Aucoin, P., . . . Carlucci, M. (2016). Registered replication report: Hart & Albarracín (2011). *Perspectives on Psychological Science, 11*(1), 158-171.

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: an easy way to maximize the accuracy and power of your research. *American Psychologist, 63*(7), 591-601.

Feingold, A. (1990). Gender differences in effects of physical attractiveness on romantic attraction: A comparison across five research paradigms. *Journal of Personality and Social Psychology, 59*(5), 981-993.

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice, 40*(5), 532-538.

Fischhoff, B. (1990). Psychology and public policy: tool or toolmaker? *American Psychologist*, *45*(5), 647-653.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic metnal models: A Brunswikian theory of confidence. *Psychological Review, 98*(4), 506-528.

Gireesh, A., Das, S., & Viner, R. M. (2018). Impact of health behaviours and deprivation on well-being in a national sample of English young people. *BMJ Paediatrics Open, 2*(1).

Gøtzsche, P. C., Hróbjartsson, A., Marić, K., & Tendal, B. (2007). Data extraction errors in meta-analyses that use standardized mean differences. *JAMA, 298*(4), 430-437.

Gustafsson, A., & Lindenfors, P. (2009). Latitudinal patterns in human stature and sexual stature dimorphism. *Annals of Human Biology, 36*(1), 74-87.

Hagger, M. S., Chatzisarantis, N. L., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., . . . Bruyneel, S. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science, 11*(4), 546-573.

Hedges, L. V. (1987). How hard is hard science, how soft is soft science? The empirical cumulativeness of research. *American Psychologist, 42*(5), 443-455.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*, 61-83.

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*(11), 1539-1558.

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ, 327*(7414), 557-560.

Hönekopp, J., & Linden, A. H. (2019). Heterogeneity estimates in a biased world. https://osf.io/zx96p/

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I^2 index? *Psychological Methods, 11*(2), 193-206.

Huhn, M., Tardy, M., Spineli, L. M., Kissling, W., Förstl, H., Pitschel-Walz, G., . . . Davis, J. M. (2014). Efficacy of pharmacotherapy and psychotherapy for adult psychiatric disorders: a systematic overview of meta-analyses. *JAMA Psychiatry, 71*(6), 706-715.

Hunter, J., & Schmidt, F. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological science, 8*(1), 3-7.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Thousand Oaks, CA: Sage Publications.

Hyde, J. S. (2014). Gender similarities and differences. *Annual Review of Psychology, 65,* 373-398.

IntHout, J., Ioannidis, J. P., Borm, G. F., & Goeman, J. J. (2015). Small studies are more heterogeneous than large ones: a meta-meta-analysis. *Journal of Clinical Epidemiology, 68*(8), 860-869.

Ioannidis, J. (2008). Interpretation of tests of heterogeneity and bias in meta-analysis. *Journal of Evaluation in Clinical Practice, 14*(5), 951-957.

Jackson, D. (2006). The implications of publication bias for meta-analysis' other parameter. *Statistics in Medicine, 25*(17), 2911-2921.

Jakobsson, U., & Westergren, A. (2005). Statistical methods for assessing agreement for ordinal data. *Scandinavian Journal of Caring Sciences, 19*(4), 427-431.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524-532.

Kenny, D. A., & Judd, C. M. (2019). The unappreciated heterogeneity of effect sizes: Implications for power, precision, planning of research, and replication. *Psychological Methods*. doi:10.1037/met0000209

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., . . . Brumbaugh, C. C. (2014). Investigating variation in replicability. *Social Psychology, 45*(3), 142-152.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., . . . Bahník, Š. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science, 1*(4), 443-490.

Kuhn, T. S. (1970). *The structure of scientific revolutions.* (2nd ed.). Chicago, IL: University of Chicago Press.

Leicester, H. M. (1965). *The historical background of chemistry*: Wiley.

Linck, J. A., Osthus, P., Koeth, J. T., & Bunting, M. F. (2014). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic Bulletin & Review, 21*(4), 861-883.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48*(12), 1181-1209.

McCrae, R. R., & Costa Jr, P. T. (1997). Personality trait structure as a human universal. *American Psychologist, 52*(5), 509-516.

McShane, B. B., & Böckenholt, U. (2014). You cannot step into the same river twice: When power analyses are optimistic. *Perspectives on Psychological Science, 9*(6), 612-625.

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.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine, 151*(4), 264-269.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., du Sert, N. P., . . . Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*(1).

Murphy, K. R. (2017). What inferences can and cannot be made on the basis of meta-analysis? *Human Resource Management Review, 27*(1), 193-200.

Muth, A., Hönekopp, J., & Falter, C. M. (2014). Visuo-spatial performance in autism: a meta-analysis. *Journal of Autism and Developmental Disorders*, *44*(12), 3245-3263.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), 943-951.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7*(6), 531-536.

Peterson, L., & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology*, *58*(3), 193-198.

Richard, F. D., Bond Jr, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7*(4), 331-363.

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology*, *52*(1), 59-82.

Schickore, J. (2018). Scientific discovery. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Summer 2018 ed.). Retrieved from https://plato.stanford.edu/archives/sum2018/entries/scientific-discovery/.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62*(5), 529-540.

Schmidt, F. L., & Oh, I.-S. (2016). The crisis of confidence in research findings in psychology: Is lack of replication the real problem? Or is it something else? *Archives of Scientific Psychology, 4*(1), 32-37.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13*(2), 90-100.

Schneider, P. (2014). *Extragalactic astronomy and cosmology: An introduction*. Berlin: Springer.

Sedlmeier, P., & Renkewitz, F. (2018). *Forschungsmethoden und Statistik für Psychologen und Sozialwissenschaftler*: Pearson Studium München.

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology, 69*(1), 487-510.

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.

Simon, H. A. (1973). Does scientific discovery have a logic? *Philosophy of Science, 40*(4), 471-480.

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science, 9*(1), 76-80.

Skinner, B. F. (1956). A case history in scientific method. *American Psychologist, 11*(5), 221-233.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32*(9), 752-760.

Stanley, T., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin, 144(12)*, 1325-1346.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance - or vice versa. *Journal of the American Statistical Association*, *54*(285), 30-34.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review, 64*(3), 153-181.

Strack, F. (2017). From data to truth in psychological science. A personal perspective. *Frontiers in Psychology, 8*. doi:10.3389/fpsyg.2017.00702

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science, 9*(1), 59-71.

Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2019a). Current practices in meta-regression in psychology, education, and medicine. *Research synthesis methods, 10*(2), 180-194.

Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2019b). A history of meta-regression: Technical, conceptual, and practical developments between 1974 and 2018. *Research synthesis methods, 10*(2), 161-179.

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist, 24*(2), 83-91.

Uhlmann, E. L., Ebersole, C. R., Chartier, C. R., Errington, T. M., Kidwell, M. C., Lai, C. K., . . . Nosek, B. A. (2019). Scientific utopia III: Crowdsourcing science. *Perspectives on Psychological Science*, *14*(5), 711-733.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1-48.

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams Jr, R., . . . Blouin-Hudon, E.-M. (2016). Registered replication report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science, 11*(6), 917-928.

Wallis, S. E. (2015). Integrative propositional analysis: a new quantitative method for evaluating theories in psychology. *Review of General Psychology, 19*(3), 365-380.

Weiss, M. J., Bloom, H. S., & Brock, T. (2014). A conceptual framework for studying the sources of variation in program effects. *Journal of Policy Analysis and Management, 33*(3), 778-808.

Weiss, M. J., Bloom, H. S., Verbitsky-Savitz, N., Gupta, H., Vigil, A. E., & Cullinan, D. N. (2017). How much do the effects of education and training programs vary across sites? Evidence from past multisite randomized trials. *Journal of Research on Educational Effectiveness, 10*(4), 843-876.

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing* (Second ed.). Burlington, MA: Elsevier Academic press.

Woodward, J. (2003). Scientific explanation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2017 ed.).

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., . . . Hinojosa, C. P. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature, 573*(7774), 364-369.

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences, 41*.

Table 1. Descriptive Statistics for Study 1

Sub-discipline	No. meta-analyses	Mean *k* per meta-analysis	Mean *T* (*SD*)
Close replications	57	35.2	0.09 (0.07)
Social	50	35.7	0.31 (0.11)
Cognitive	50	36.5	0.32 (0.13)
Organizational	50	38.3	0.35 (0.10)
Total	*150*	*36.9*	*0.33 (0.11)*

Note. All means are winsorized.

Table 2. Irrelevant differences in standard deviations across studies negatively affect the suitability of standardized effect sizes.

	Control group (N = 200)	Experimental group (N = 200)	Difference	*d*
Study 1	M = 50.0 SD = 10.0	M = 60.0 SD = 10.0	10.0	1.00
Study 2	M = 50.0 SD = 15.0	M = 60.0 SD = 15.0	10.0	0.67
Study 3	M = 50.0 SD = 10.0	M = 56.7 SD = 10.0	6.7	0.67

Note: Three similar, fictitious studies into the same phenomenon employ the same dependent variable. Due to Study 2 employing a more diverse sample, the standardized effect size *d* misleadingly suggests that Studies 2 and 3 obtained the same results, whereas the difference in means shows that Studies 1 and 2 obtained the same results.

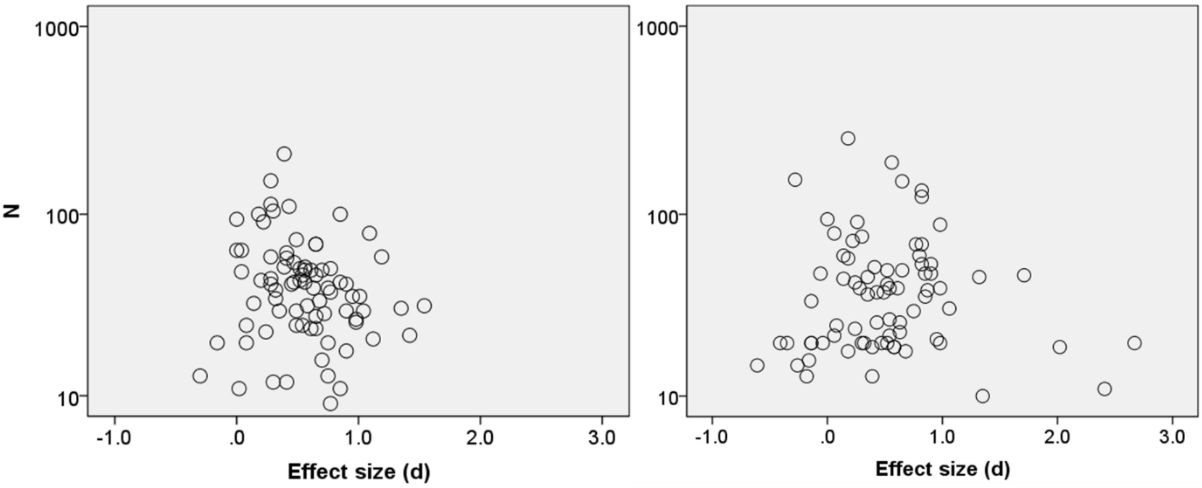

Figure 1. Funnel plots for two meta-analyses

Note. **Left-hand panel**: Linck, Osthus, Koeth, and Bunting (2014), investigated the link between working memory and second language comprehension. Estimated mean of the population of effect sizes *d* = 0.51, standard deviation of observed effect sizes is 0.36, estimated heterogeneity of true effect sizes *T* = 0.11, I^2 = 11. **Right-hand panel**: Baker, Peterson, Pulos, and Kirkland (2014), investigated the link between intelligence and performance on the Reading the Mind in the Eyes test. *d* = 0.49, *SD* = 0.59, *T* = 0.35, I^2 = 53.

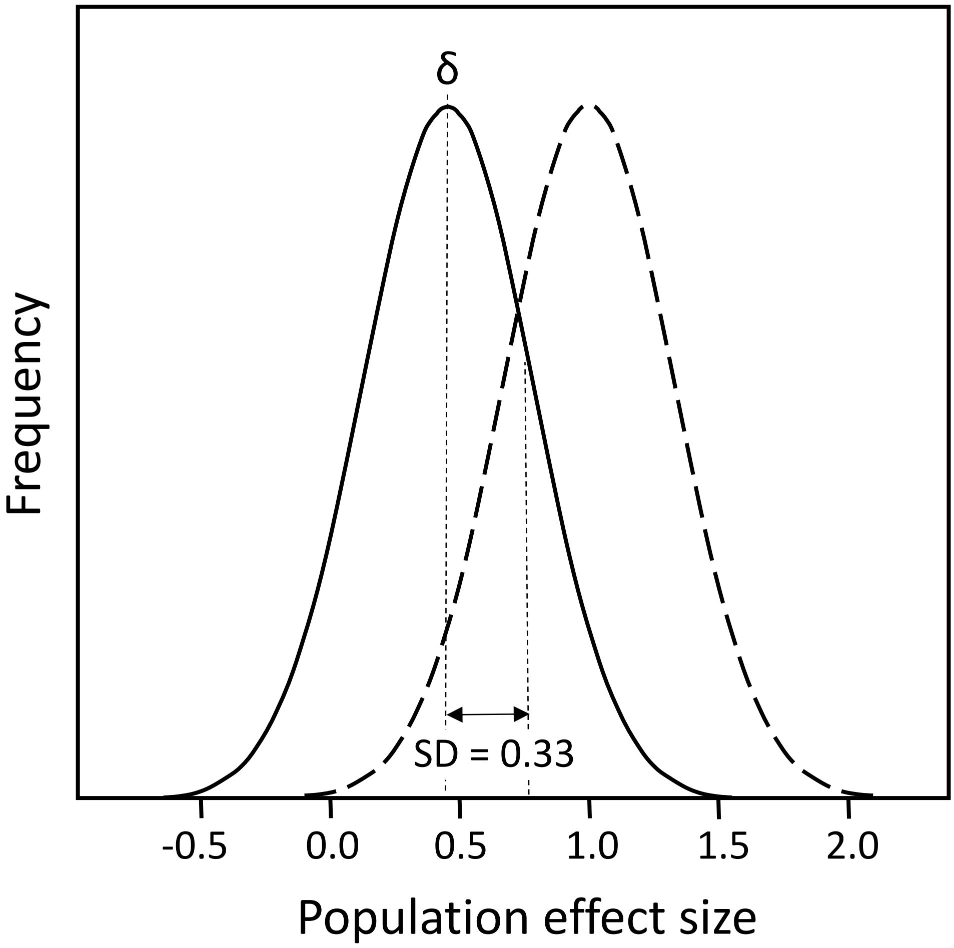

Figure 2. Two distributions of population effect size (standardized mean differences).

Note. The distribution on the left (solid line) shows a population effect size with a mean (δ) of 0.45 and a standard deviation (*T*) of 0.33. The distribution on the right (dashed line) shows a population effect size with a mean (δ) of 1 and a standard deviation (*T*) of 0.33.

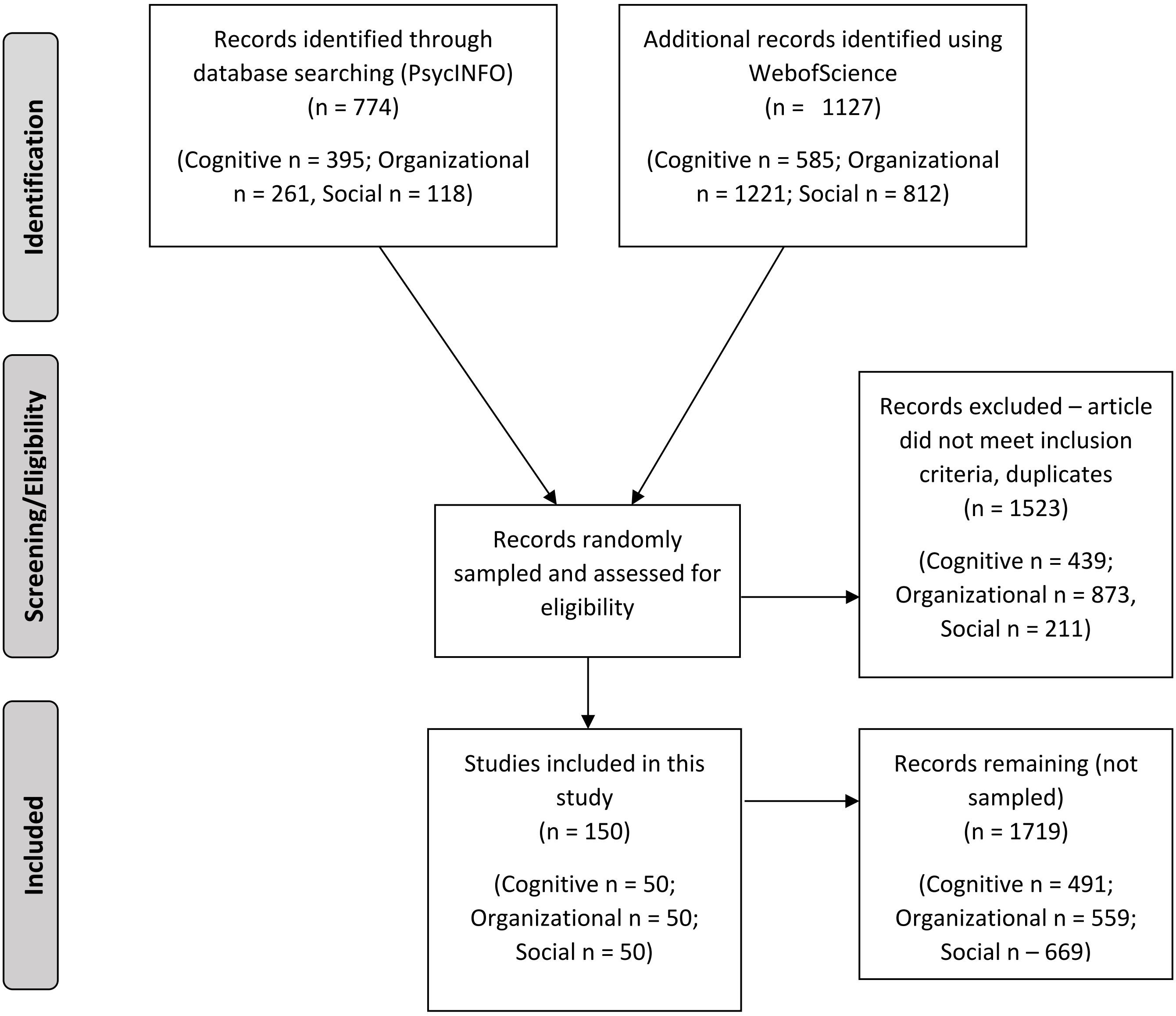

Figure 3. Sampling of meta-analyses.

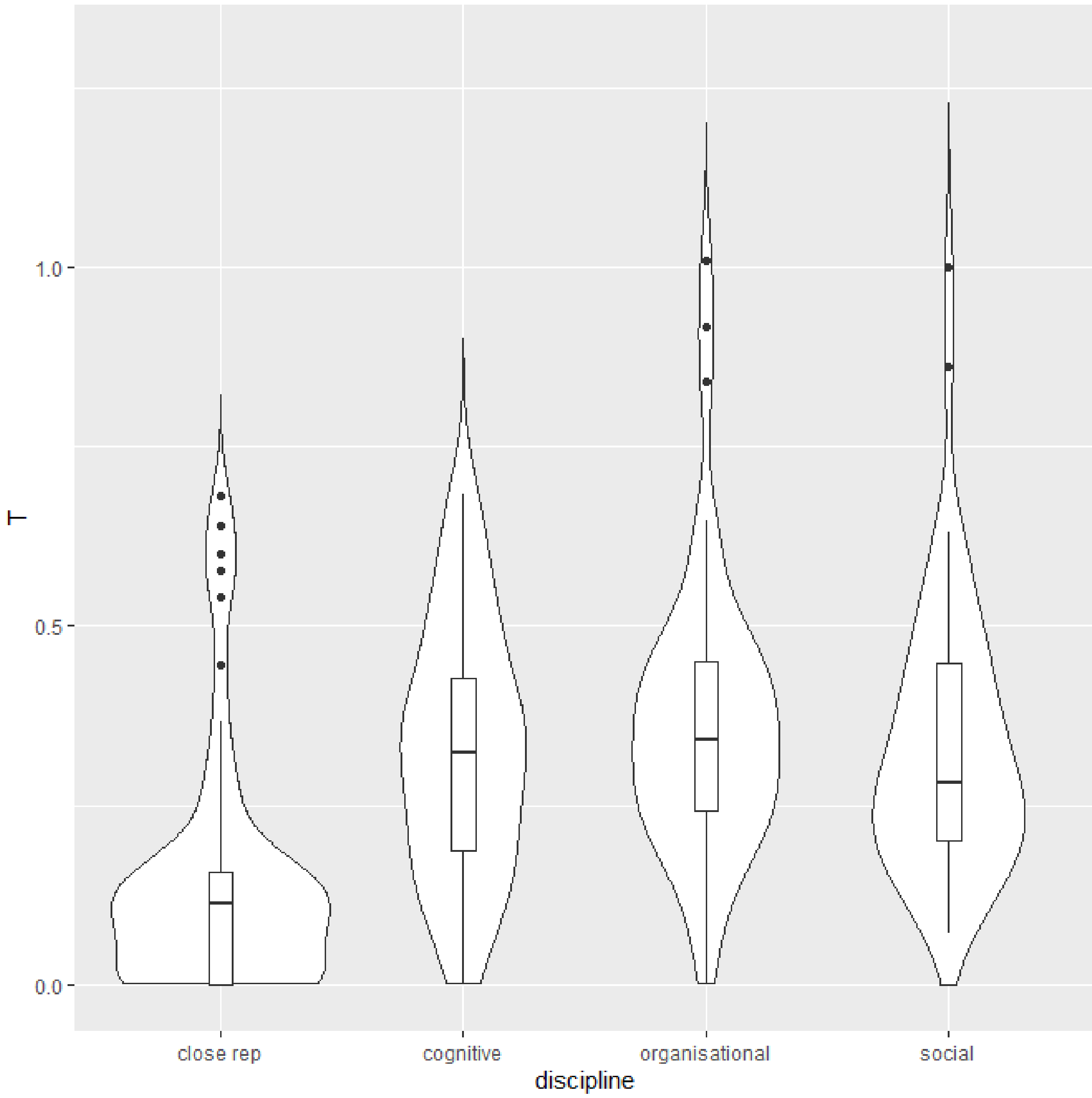

Figure 4. Observed levels of heterogeneity for 57 close replications and for 50 meta-analyses (each) in cognitive, organizational and social psychology, with box-plots for M_{win}.

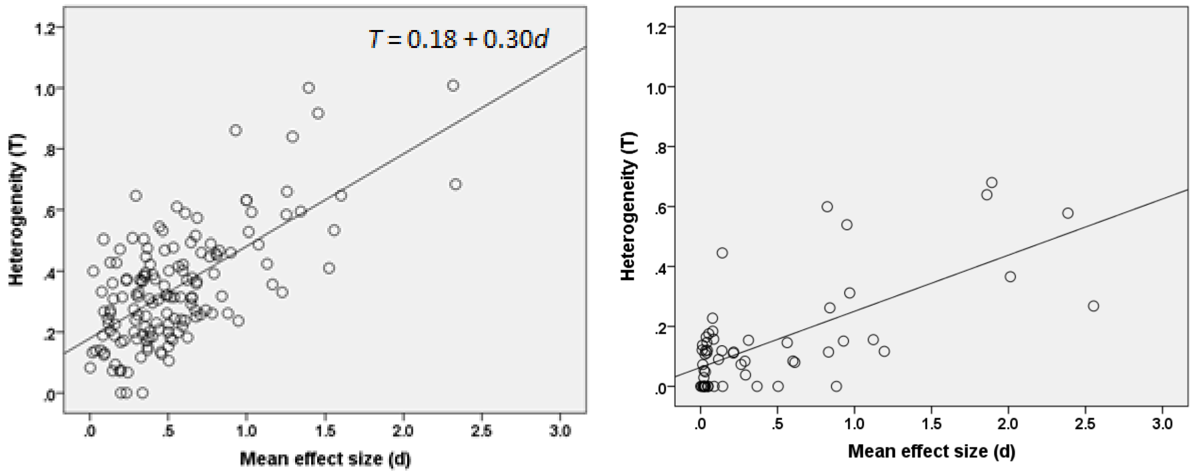

Figure 5. Heterogeneity as a function of meta-analyses' mean effect size. Left hand panel: 150 meta-analyses from cognitive, social, and organizational psychology; right hand panel: 57 meta-analyses for close replications.