Holzmeister, Felix et al.

Working Paper

Heterogeneity in Effect Size Estimates: Empirical Evidence and Practical Implications

I4R Discussion Paper Series, No. 102

Provided in Cooperation with:
The Institute for Replication (I4R)

Suggested Citation: Holzmeister, Felix et al. (2024) : Heterogeneity in Effect Size Estimates: Empirical Evidence and Practical Implications, I4R Discussion Paper Series, No. 102, Institute for Replication (I4R), s.l.

This Version is available at:
https://hdl.handle.net/10419/283318

INSTITUTE for REPLICATION

No. 102
I4R DISCUSSION PAPER SERIES

Heterogeneity in Effect Size Estimates: Empirical Evidence and Practical Implications

Felix Holzmeister
Magnus Johannesson
Robert Böhm
Anna Dreber
Jürgen Huber
Michael Kirchler

February 2024

INSTITUTE for REPLICATION

I4R DISCUSSION PAPER SERIES

I4R DP No. 102

Heterogeneity in Effect Size Estimates: Empirical Evidence and Practical Implications

Felix Holzmeister[1], Magnus Johannesson[2], Robert Böhm[3,4], Anna Dreber[1,2], Jürgen Huber[1], Michael Kirchler[1]

[1]University of Innsbruck/Austria
[2]Stockholm School of Economics, Stockholm/Sweden
[3]University of Vienna/Austria
[4]University of Copenhagen/Denmark

FEBRUARY 2024

E-Mail: joerg.peters@rwi-essen.de
RWI – Leibniz Institute for Economic Research
Hohenzollernstraße 1-3
45128 Essen/Germany
www.i4replication.org

ISSN: 2752-1931

Heterogeneity in effect size estimates: Empirical evidence and practical implications

Felix Holzmeister[1,*], Magnus Johannesson[2], Robert Böhm[3,4], Anna Dreber[1,2], Jürgen Huber[5], Michael Kirchler[5]

[1] Department of Economics, University of Innsbruck, Innsbruck, Austria. [2] Department of Economics, Stockholm School of Economics, Stockholm, Sweden. [3] Faculty of Psychology, University of Vienna, Vienna, Austria. [4] Department of Psychology and Center for Social Data Science, University of Copenhagen, Copenhagen, Denmark [5] Department of Banking and Finance, University of Innsbruck, Innsbruck, Austria.

* Correspondence should be addressed to:

Felix Holzmeister
University of Innsbruck, Department of Economics
Universitätsstraße 15, 6020 Innsbruck Austria
felix.holzmeister@uibk.ac.at

Abstract

A typical empirical study involves choosing a sample, a research design, and an analysis path. Variation in such choices across studies leads to heterogeneity in results that introduce an additional layer of uncertainty not accounted for in reported standard errors and confidence intervals. We provide a framework for studying heterogeneity in the social sciences and divide heterogeneity into population heterogeneity, design heterogeneity, and analytical heterogeneity. We estimate each type's heterogeneity from multi-lab replication studies, prospective meta-analyses of studies varying experimental designs, and multi-analyst studies. Our results suggest that population heterogeneity tends to be relatively small, whereas design and analytical heterogeneity are large. A conservative interpretation of the estimates suggests that incorporating the uncertainty due to heterogeneity would approximately double sample standard errors and confidence intervals. We illustrate that heterogeneity of this magnitude—unless properly accounted for—has severe implications for statistical inference with strongly increased rates of false scientific claims.

Introduction

Designing an empirical study, collecting or sourcing data, and analyzing data calls for making decisions in heaps, many of which are up to the researcher's discretion[1]. This flexibility, dubbed researcher degrees of freedom, opens the door to a "garden of forking paths"[2] involving many branches (choices) and countless designations (empirical results). Yet, empirical research methods in the social sciences typically involve relying on one particular sample, choosing one out of many possible study designs, and reporting the results for one of many possible analysis pipelines. In light of a "publish or perish" culture in academia[3–5], scholars have a strong incentive to exploit researcher degrees of freedom to obtain statistically significant results and selectively report empirical estimates that maximize the publication potential[6–9]. It is now acknowledged that the opportunistic misuse of researcher degrees of freedom—commonly referred to as selective reporting and *p*-hacking—implicates increased false-positive rates[10–12] and inflated effect sizes[13–15]. Alongside publication bias[16–18], low statistical power[19–22], and HARKing[23,24], *p*-hacking has been argued to be one of the "four horsemen of the reproducibility apocalypse"[25]. The overall impact of questionable research practices has been empirically demonstrated in several large-scale direct replication projects[26–28], which suggest that, on average, replication effect sizes are only about 50% of the published effect sizes in empirical social science research. Scientific reforms such as open, transparent, and confirmatory research practices have been advocated—and implemented to a greater or lesser extent—to reduce systematic bias in the published literature and "rein in the four horsemen"[29–33].

Even if researchers and journals adopt a culture of confirmatory research practices[34,35] to remedy systematic bias in the scientific knowledge accumulation, the scientific community faces another major obstacle on its way toward reliable empirical evidence: the doubt about the generalizability and robustness of the reported results to alternative populations, research designs, and analytical decisions[36–40]. Typically, empirical studies only capture tiny snapshots of the range of possible results, and common estimates of the uncertainty about these snapshots do not account for the uncertainty due to the flexibility in choosing a sample, a research design, and an analysis path during a research project. The magnitude of this unaccounted-for uncertainty—commonly referred to as heterogeneity—depends on how much results vary across populations, alternative research designs, and alternative analysis paths. Failing to account for heterogeneity undermines the generalizability of empirical findings and can result in unwarranted claims.

In this paper, we delve into the various sources of heterogeneity in the empirical social sciences, categorizing heterogeneity into three distinct types: population heterogeneity, design heterogeneity, and analytical heterogeneity. We review the evidence on different types of heterogeneity in the social sciences based on research settings where each type is isolated and systematic bias in effect sizes (due to *p*-hacking and publication bias) has been ruled out by design. We illustrate the implications of the observed levels of heterogeneity for statistical inference and show that it can drastically increase the fraction of false scientific claims and severely limit the informativeness and generalizability of individual scientific studies. We discuss the implications of our findings for scientific practice and shed light on potential pathways to improve the knowledge generation process of empirical studies in the social sciences to avoid getting stuck in a generalizability crisis[36–39]. We argue for moving away from the common "one population–one design–one analysis" approach toward large-scale preregistered prospective meta-analyses systematically varying populations, designs, and analyses.

Framework

While the term *heterogeneity* may be used with slightly different meanings across various contexts, we adhere to the definition that is specific to random-effects meta-analyses, where a distinction is made between the within-study variance (σ^2; i.e., the sampling error) and the between-study variance (τ^2; heterogeneity)[41]. In this realm, heterogeneity is uniformly defined as the variation in effect size estimates over and above sampling variation, i.e., observing study outcomes being more different from one another than would be expected due to chance alone. The square root of the between-study variance (τ) has the intuitive interpretation of the standard deviation of the distribution of true effect sizes across the studies included in the meta-analysis.

Heterogeneity can be quantified in terms of τ and can be expressed in both absolute and relative terms. While the absolute magnitude of heterogeneity is important, it is difficult to compare estimates across studies utilizing different effect size measures. Estimates of τ can only be reasonably compared across meta-analyses if they utilize the same standardized effect size measure. As the effect size measurement varies across the empirical studies reviewed below, we focus on quantifying heterogeneity in relative terms to facilitate comparability but also report the heterogeneity estimates in absolute terms. A common way to quantify heterogeneity in relative terms is to express the overall variability in effect sizes, i.e., $\tau^2 + \sigma^2$, in within-study variance (σ^2) units. This ratio is commonly denoted as H^2 and can be thought of as a variance inflation factor due to heterogeneity. In what follows, we favor the square root-transformed version of H^2 to

facilitate interpretability (i.e., to characterize heterogeneity in standard deviation units rather than in variance units), and refer to it as the *heterogeneity factor* (*H*), which is defined as

$$H = \sqrt{\frac{\sigma^2 + \tau^2}{\sigma^2}}.$$

This expression has previously been proposed as a heterogeneity measure in the context of random-effects meta-analyses[42,43], and is related to the commonly reported heterogeneity measure I^2, defined as the percentage of the total variability in effect size estimates attributable to heterogeneity, i.e.,

$$I^2 = \frac{\tau^2}{\sigma^2 + \tau^2} \quad \longrightarrow \quad H = \sqrt{\frac{1}{1 - I^2}}.$$

The commonly referenced cut-off values of 25%, 50%, and 75% for I^2 are used to indicate small, medium, and large heterogeneity[44,45] and translate into cutoff values for *H* of 1.15, 1.41, and 2.00, respectively. *H* is the factor that the sampling standard error needs to be multiplied by to account for heterogeneity; $H = 1$ implies homogeneity, and $H = 2$ implies that incorporating uncertainty due to between-study variation will double the sample standard error of an individual study.

Heterogeneity in effect sizes may stem from various sources: study outcomes might be heterogeneous across samples drawn from different populations (*population heterogeneity*), estimates can vary depending on the study design used to address a particular hypothesis (*design heterogeneity*), and effect sizes may differ depending on the analysis path implemented (*analytical heterogeneity*). For studies that rely on prospective data collections, such as experiments, the three types of heterogeneity relate to degrees of freedom in different layers of the research process. These include deciding *(i)* which population(s) to use to draw a sample from, *(ii)* which research design to implement, and *(iii)* how to analyze the sampled data. For empirical studies that rely on observational data, it may not be clear where to draw the line between design and analytic decisions. For ease of exposition, we consider all researcher decisions made after choosing which raw data to use as part of the analytical domain in empirical studies.

Each type of heterogeneity can be isolated and quantified by implementing proper research designs. By allowing for variation in only one dimension (e.g., the study designs) while holding the other dimensions (e.g., the population and the analysis) constant, the magnitude of heterogeneity due to the different sources of variation can be examined systematically. In our empirical review, we report estimates of heterogeneity expressed in terms of the heterogeneity factor (*H*) separately for

population, design, and analytical heterogeneity (based on meta-analyses isolating either of the three sources by design). Note that heterogeneity estimates due to different sources of variability are linearly additive in variances, such that the overall heterogeneity (in absolute terms) is given by $\tau^2 = \tau_P^2 + \tau_D^2 + \tau_A^2$, where τ_s denotes the between-study variance due to variability across populations ($s = P$), designs ($s = D$), or analysis paths ($s = A$), respectively. The total variance in the effect size estimate of study j, v_j^2, is thus given by $v_j^2 = \sigma_j^2 + \tau_P^2 + \tau_D^2 + \tau_A^2$, and its square root, v_j, can be thought of as study j's total standard error. Importantly, the way H is constructed remains unchanged irrespective of the source of variability, i.e., whether H pertains to one of the three sources of heterogeneity or any combination thereof is solely governed by the research design. In all empirical analyses reported below, the heterogeneity factor H is determined based on the within-study variance (σ^2) estimated as part of the meta-analytic random-effects model. If the sample variance of some study j differs from the average within-study variance (σ^2) due to, for instance, using a larger sample size or due to more precise measurement, the heterogeneity factor H for study j can be derived by replacing σ^2 by σ_j^2.

Empirical estimates of population, design, and analytical heterogeneity

To gauge the extent of heterogeneity in empirical social science research, we provide a review of heterogeneity estimates re-estimated using random-effects meta-analysis based on published crowd science projects. We outline the inclusion criteria and the estimation procedures in the Methods section; details about the individual studies are provided in sections 1–3 in the Supplementary Methods. The results are illustrated in Figure 1, which also indicates benchmarks for low, medium, and high heterogeneity based on an I^2 of 25% (H = 1.15), 50% (H = 1.41), and 75% (H = 2.00); these benchmarks are commonly used in meta-analysis to classify the magnitude of heterogeneity[44,45]. Estimates of the heterogeneity measures τ, I^2, and H (together with their corresponding 95% CIs) and the results of Cochran's Q-test for each meta-analysis reviewed in our empirical analysis are tabulated in Supplementary Table 1.

Population heterogeneity. Population heterogeneity can be measured by implementing the same research design and analysis in separate samples from different populations and estimating the standard deviation in true effect sizes across samples (τ) in a random-effects meta-analysis. This is what has been pioneered in the *ManyLabs* (ML) replication studies and various Registered Replication Reports (RRRs) in psychology, which are ideal for measuring population heterogeneity. Our analysis involves four ML studies[46–49] and nine RRRs[50–58]. As some of the included studies report

results for multiple effects, our sample of studies isolating population heterogeneity comprises 70 meta-analyses.

The estimated population heterogeneity varies substantially across the meta-analyses in our sample, with a large number of estimates (19/70 = 27.1%) indicating homogeneity (H = 1.00), but also some estimates unveiling substantial heterogeneity of up to H = 3.91 (with 4/70 estimates exceeding the threshold value of H = 2.00, indicative of large heterogeneity). The median H across the 70 meta-analyses is 1.08, and a large fraction of the estimates are in the small to moderate heterogeneity range. Cochran's Q-test rejects the null hypothesis of homogeneity at the 5% level for 21 (30%) of the sampled meta-analyses and at the 0.5% level for 14 (20%) of the sampled meta-analyses. Some meta-analyses (46/70 in 4/13 papers) are based on effect sizes measured in terms of Cohen's d; heterogeneity can be reasonably compared across studies in absolute terms (τ) for this subsample. The estimated τ varies between 0.00 and 0.69 for these estimates, with a median of 0.06. Note that the distributions of H and τ estimates of population heterogeneity are subject to some upward bias as a consequence of following the convention to truncate τ^2 estimates at zero (i.e., to prevent the identification of excess homogeneity)[42,43]. For genuinely homogeneous effect sizes, randomness would lead to both negative and positive estimates of τ^2. Whenever the fraction of meta-analyses with zero estimated heterogeneity is large—as is the case for our sample of studies isolating population heterogeneity—, this upward bias can be substantial.

Design heterogeneity. Design heterogeneity can be measured by randomly allocating experimental participants sampled from the same population to different research designs while holding the analysis constant, and estimating the standard deviation in true effect sizes across research designs (τ) in a random-effects meta-analysis. We identified two studies that implemented such a research design, reporting the results of 11 meta-analytic estimates for six empirical claims: Landy et al.[59] tested five hypotheses on moral judgments, negotiations, and implicit cognition in 12 to 13 experimental designs each (once in a "main study" and once in a replication). Huber et al.[60] examined the effect of competition on moral behavior across 45 crowd-sourced experimental protocols. The estimates of H for the 11 meta-analyses reported in the sampled studies (see Figure 1) suggest that the extent of design heterogeneity is substantial. The estimates of H vary between 1.92 and 10.44, with a median of 3.36, and Cochran's Q-test rejects the null hypothesis of homogeneity ($p < 0.005$) for each of the 11 meta-analyses. These results suggest that design heterogeneity is substantially larger than population heterogeneity and adds substantial uncertainty to studies based on individual designs. All estimates of design heterogeneity are in Cohen's d units, and the estimated τ varies between 0.14 and 0.78, with a median of 0.23.

Analytical heterogeneity. An effective means to estimate analytical heterogeneity involves randomly allocating independent analysts to test the same hypothesis on mutually exclusive random sub-samples of a dataset and estimate the standard deviation in true effect sizes across analysts (τ) in a random-effects meta-analysis. To the best of our knowledge, no studies have employed this method yet. The most similar approach to this ideal comprises studies that rely on the multi-analyst approach, where different analysts independently test the same hypothesis on the same data. Our review involves three papers that fulfill our inclusion criteria (see Methods for details), examining the variability in effect sizes due to analytical flexibility for five hypotheses[61–63].

As the analysts in multi-analyst studies are required to estimate the effect in question using the same data, the individual estimates generated by analysts are not independent. Despite the violation of the model assumptions, we use random-effects meta-analyses to estimate τ and H as an approximation of the analytical heterogeneity for the sampled multi-analyst studies. This method was also recently used by a multi-analyst study in biology[64] to estimate the heterogeneity of results across analysts. The estimates reported in Figure 1 should be interpreted cautiously since relying on the random-effects meta-analytic model will underestimate heterogeneity for correlated observations (as the within-study variation will be lower in the case of dependent observations). The estimated analytical heterogeneity is large, with H estimates ranging from 1.72 to 12.69, with a median of 4.08. Cochran's Q-test is statistically significant ($p < 0.005$) for each of the five meta-analyses. In the Methods section we provide a robustness test on the estimates of analytical heterogeneity, showing that the high estimate of 12.69 for Silberzahn et al.[61] is driven by two outliers in terms of sampling variance; removing these outliers reduced the estimated heterogeneity factor H (and the median estimate for the multi-analyst studies in our sample) to 2.72. Our results suggest that analytical heterogeneity is substantial, in the same ballpark as the estimates for design heterogeneity, and substantially larger than population heterogeneity. None of the estimates of analytical heterogeneity are in standardized effect size units, and the analytical heterogeneity estimates cannot be reasonably compared in absolute terms across the studies.

Limitations. It is worth emphasizing that the estimated heterogeneity factor for all three types of heterogeneity carries a significant amount of uncertainty: Heterogeneity estimates can be highly sensitive to outliers in effect sizes and sample variances of the individual studies included in the meta-analyses. The individual estimates of H in Figure 1 should therefore be interpreted with caution. Multi-analyst studies have also been criticized for overestimating the analytical variation, e.g., due to ambiguity about the studied research question[65,66]. Notwithstanding, it is important to note that even the

smallest estimates obtained from our review of the literature point toward considerable heterogeneity due to the variability in designs (H = 1.92) and analyses (H = 1.72). It is also worth emphasizing that a typical empirical study in the social sciences will involve heterogeneity due to all three sources investigated above, which implies that the overall level of heterogeneity is likely to be even higher than the estimates reported in Figure 1.

Illustrating the Impact of Heterogeneity on Statistical Inference

The above results illustrate that the variation in results across studies testing the same hypothesis can be large, especially due to the variation in research designs and analytical decisions. In this section, we illustrate the implications of unaccounted-for heterogeneity for statistical inference (see the Methods section for more details). We use the following scenario as a starting point: A researcher sources data to test a hypothesis with 90% statistical power to detect the hypothesized effect size at the 5% significance threshold (in a two-sided z-test) only taking into consideration sampling uncertainty but not heterogeneity (i.e., the standard way to test hypotheses and to do power calculations). What are the consequences of ignoring heterogeneity if the true effect size is genuinely heterogeneous?

We distinguish between the nominal and effective error rates below, where the effective error rates are the observed error rates after taking into account heterogeneity. Fig. 2 demonstrates the effective type-I and type-II error rates for this scenario, assuming that the true effect size is heterogeneous with H = 2, which is close to the smallest estimates of design heterogeneity and analytical heterogeneity reported in Fig. 1. As illustrated in Fig. 2, heterogeneity increases the dispersion of an effect's probability density under both the null and the alternative hypothesis as the sample standard error will double when heterogeneity is incorporated. Since the researcher in our scenario ignores heterogeneity, she will not adapt the test's critical value but will decide on whether or not to reject the hypothesis based on the critical value based on the *nominal* significance threshold. As a consequence, both the effective false positive rate and the effective false negative rate are inflated (or, put differently, both the specificity and the sensitivity of the test are depleted). As exemplified in Fig. 2, the implications of heterogeneity can be severe: a study that is supposed to involve a type-I error rate of 5% and a type-II error rate of 10% actually entails a 33% risk of a false positive and a 26% risk of a false negative result under the assumption of H = 2.

Fig. 3a plots the relationship between the nominal and effective type-I error rates for various levels of heterogeneity. The effective type-I error rate increases strongly with heterogeneity, implying that unaccounted-for heterogeneity has considerable adverse effects even for comparably low levels of between-study variance. The corresponding

relationship between nominal and effective statistical power is illustrated in Fig 3b. As we already saw in Fig. 2, large heterogeneity substantially decreases effective power when the nominal power is high, but the effect of heterogeneity goes in the opposite direction for low nominal power. This may at first seem counterintuitive, but the effective power will go toward 50% when heterogeneity increases, irrespective of the nominal power, as high levels of heterogeneity decrease the chance of detecting a statistically significant effect when nominal power is high (>50%) but increase the chance of observing a statistically significant effect when nominal power is low (<50%).

By taking into account the prior likelihood of a tested hypothesis being true (ϕ), the effective false discovery rate can be estimated. Fig. 4a shows how the false discovery rate—i.e., the fraction of statistically significant findings that are false—varies with heterogeneity for different priors ϕ (assuming 90% nominal statistical power and a 5% nominal significance threshold as in Fig. 2). The effective false discovery rate (*FDR'*) increases strongly with the H unless heterogeneity is incorporated into statistical testing; e.g., for a prior of ϕ = 30%, *FDR'* goes from 11.5% for H = 1 to 50.8% for H = 2. For lower priors, the impact of heterogeneity on the false discovery rate is even more severe.

Instead of directly incorporating heterogeneity into the standard errors of reported effect sizes (by multiplying sampling errors with an appropriate heterogeneity factor H), the adverse effects of heterogeneity could be tamed through applying a stricter nominal significance threshold in statistical testing. Benjamin et al.[67] recently suggested lowering the p-value threshold from 5% to 0.5%. In Fig. 4b, we show the false discovery rate for nominal p-value thresholds of 5%, 0.5%, and 0.05% for different priors as a function of the heterogeneity factor H (based on our example with a nominal statistical power of 90%). For H = 2, lowering the p-value threshold from 5% to 0.5% (0.05%) reduces *FDR'* from 50.8% to 33.6% (20.5%) for a prior of 30%. While adopting lower p-value thresholds curbs the detrimental impact of heterogeneity on false discoveries, the nominal significance level needs to be lowered more drastically than proposed by Benjamin et al.[67] to cope with the magnitude of design and analytical heterogeneity identified in our empirical exercise summarized in Fig. 2. To counteract the impact of heterogeneity of H = 2 given a prior of 30%, the (nominal) α-threshold needs to be reduced to 0.005% (to uphold the FDR' for H = 1 and α = 0.05)). For settings with lower statistical power, the patterns highlighted in Fig. 4a and 4b are similar, but the *FDR'* will converge even faster toward its limit of $1 - \phi/2$ for increasing magnitudes of H.

The considerations sketched above draw an unmistakable picture of why the scientific enterprise ought to start taking action to parse and cope with heterogeneity[36–40]. As illustrated, disregarding heterogeneity can have a substantial impact on statistical inference, which in turn implies that a priori power analyses can be misleading and the planning of original and replication studies might be misguided[68]. Ignoring the

implications of heterogeneity will leave us with substantially inflated numbers of false scientific claims, and compromises the informativeness and conclusiveness of individual scientific contributions. Consequently, unaccounted-for heterogeneity bears the risk of generating research waste, potentially slowing down the process of scientific discovery, and generating a poor return on the invested funding[69–71].

Discussion

Aside from the literature reviewed above, there is a wealth of meta-analytic studies in the social sciences reporting heterogeneity estimates as part of random-effects meta-analyses. For the sake of comparison, the results of two studies are worthwhile to mention: van Erp et al.[72] sourced more than 700 meta-analyses published in the *Psychological Bulletin* and reported a median I^2 estimate of 71%; Stanley et al.[73] reviewed a convenience sample of 200 meta-analyses published in the same journal and reported a median I^2 of 74%. These I^2 estimates—pooling all potential sources of heterogeneity—translate into heterogeneity factors (H) of 1.86 and 1.96, respectively. However, these estimates are difficult to draw on for our purpose and the comparability with our estimates is limited since heterogeneity estimates in meta-analyses based on the published literature will be impacted by publication bias and *p*-hacking[15–18,74]. In our review of results, we only draw on studies that are, by design, free from publication bias and obvious incentives for *p*-hacking. This literature is still at an early stage, and our results should be interpreted with care; yet, drawing some preliminary conclusions appears tenable. Our results suggest that population heterogeneity is typically small, which is consistent with two other recent studies estimating population heterogeneity based on multi-lab replication studies[75,76]. Our results furthermore suggest that both design heterogeneity and analytical heterogeneity are large: even the lowest estimates in the reviewed literature imply that design and analytical heterogeneity almost double standard errors and confidence intervals if accounted for in statistical testing.

A typical empirical study will be associated with all three sources of heterogeneity, implying even higher levels of uncertainty not captured by standard errors. However, we would be reluctant to simply add up our three estimates for different sources of heterogeneity as they are based on different types of studies. The estimates of population and design heterogeneity are based on experimental studies, whereas the estimates of analytical heterogeneity are based on observational data research. We would expect less analytical heterogeneity for the typical experiment than for the typical observational study due to fewer analytical choice points encountered on average[7,8]. Conversely, for observational data studies, it is more difficult to cleanly separate the research design from analytical decisions; analytical heterogeneity may

incorporate (part of) the variability of "design elements," whereas the remaining design heterogeneity may be lower than for experiments. More research is needed to gauge the relative importance of various types of heterogeneity. The tentative insights gained from our review suggest that total heterogeneity in social science research can be expected to be substantial, with substantial implications for statistical inference.

The sizeable analytical heterogeneity identified in our review also implies that the scope to selectively report favorable results is wide. While *p*-hacking is often thought of as marginally affecting results around the significance thresholds, the extent of observed analytical heterogeneity suggests that there is potential for much larger systematic bias in published effect sizes. Similarly, design heterogeneity implies that researchers may be able to selectively report results for research designs that deliver the desired results. "Design hacking" could manifest itself in opportunistically choosing the experimental design that is expected to maximize the chances of finding statistically significant results based on, e.g., piloting different protocols and parameterizations. In reported research, all pilot studies and related tests that have been used to inform the eventual research design should be explicitly reported; ideally, studies should be preregistered before conducting any pilot tests such that the piloting choices are explicitly incorporated into the overall research design.

For our estimates of population heterogeneity, an important caveat is that the reviewed multi-lab replication studies are typically based on university student samples from different western countries, which may involve lower population heterogeneity than in other settings. Put differently, our comparatively low estimates of population heterogeneity might be subject to population heterogeneity itself. To what extent one should incorporate population heterogeneity into the reported uncertainty of individual studies also depends on which population the researcher wants to generalize the results to[77,78]. When conducting an experiment on university students, it seems fair to expect that results are generalizable to similar student populations. However, it may not be justifiable to generalize the findings to other populations, such as students in different countries, or the general population. To avoid overgeneralization, empirical investigations should start with representative samples of the population for which the results ought to be informative, in which case the population heterogeneity will be "absorbed" by the sampling standard error of the study. For population heterogeneity, it may also be important to study whether and to what extent effect sizes systematically vary across populations rather than generalizing results beyond the population studied in a specific study. Gauging the variability in effect sizes across populations is of direct interest, can inform future research agendas, and may be policy-relevant.

For analytical heterogeneity, there is a strong case for adding the analytical uncertainty to the sampling variance uncertainty of individual studies *per se*. The same applies to

design heterogeneity, except for normative studies that aim to identify the most effective design to achieve some specific goal. An example would be randomized control trials testing various nudging interventions, where different experimental designs compete in a horserace to achieve a particular goal most efficiently[79–81]. In such a setting, we are not interested in generalizing the results of a specific design to all the feasible designs, and the heterogeneity in effect sizes across all feasible research designs is not part of the uncertainty of the effect of the most efficient design.

Incorporating the additional uncertainty due to heterogeneity into statistical testing is difficult due to the uncertainty about the magnitude of heterogeneity that should be expected for different settings. An alternative could be to report the level of heterogeneity an individual study would be robust to in generalizing findings to other populations, designs, and analysis paths. The heterogeneity factor *H,* at which an individual result would turn insignificant, could be reported alongside the *p*-value for studies reporting statistically significant findings. For *z*- and *t*-tests, this "heterogeneity buffer" can be straightforwardly determined as the ratio of the test statistic and the critical value. The buffer can be interpreted as an indicator of the generalizability of an individual study's result, and the cutoff values for small ($H = 1.15$), medium ($H = 1.41$), and large ($H = 2.00$) heterogeneity could be used as a pointer as to whether an empirical claim is robust to low, medium, or large heterogeneity.

The estimated levels of design and analytical heterogeneity imply that the informativeness and generalizability of individual studies are typically low, and we believe that the common "one population – one design – one analysis" approach is outdated. Besides the low generalizability of such studies, another major issue is that the sequential production of studies implies that the scientific knowledge-generation process is delayed.[39] The publication of one random study might inspire follow-up studies, which in turn trigger follow-up studies, etc. With scientific evidence pertaining to a narrowly defined set of hypotheses being published sequentially, it could take years to reach a broader perspective on heterogeneity and generalizability. The process of sequential publication further involves the threat that flawed initial results could steer an entire sub-discipline in the wrong direction, which, in turn, could lead to loads of research waste and impede efficient knowledge accumulation.

We thus argue that it is time to initiate a paradigm shift, both in how to conduct (empirical) scientific research and in how to communicate the evidential value of scientific contributions to various stakeholders. We advocate moving towards fewer and much larger empirical studies in which conclusive research designs, justifiable analysis paths, and relevant populations are systematically varied as part of an encompassing research design. When analyzing such studies using a random-effects meta-analytic model, heterogeneity is incorporated into the standard errors of the meta-analytic effect

size, and heterogeneity can be leveraged to build more complete and more nuanced theories[38]. Importantly, such studies should be preregistered and can involve elements of crowd science to reach the necessary scale[71,82]. We think of such studies as preregistered prospective meta-analyses to distinguish them from classical meta-analyses based on the published literature that are hampered by publication bias and *p*-hacking (not only in primary studies but also in meta-analyses, which involve many degrees of freedom such as, e.g., defining inclusion criteria)[74]. The *ManyLabs* studies[46–49] in psychology can be thought of as examples of pioneering work in this direction, although primarily concerned with studying population heterogeneity.

There is also a case for more use of multi-analyst studies and multiverse analysis[83–87]. These approaches can be used to unveil the scope of analytical variation and incorporate the implied uncertainty into eventual conclusions as to the hypothesis in question. While a wider adoption of these methodologies is desirable *per se*, integrating these approaches into pre-registered prospective meta-analyses will unleash their full potential. Furthermore, sharper theoretical predictions, methodological standardization, and clearer alignment of theoretical conceptualizations and empirical instrumentalizations facilitate narrowing down the set of plausible research designs and analytical choices, ultimately reducing heterogeneity[36–38].

Methods

Included studies

Population heterogeneity. We reviewed the literature for *ManyLabs* (ML) replication studies and Registered Replication Reports (RRRs) in psychology, which are ideal for measuring population heterogeneity, and included all ML and RRRs using random-effects meta-analysis and with available data on effect sizes and standard errors for each included lab. We included ML1-4[46–49] and nine RRRs published in *Perspectives on Psychological Science* and *Advances in Methods and Practices in Psychological Science*[50–58]. We did not include ML5[88] due to a lack of data availability. As ML1–3 and several RRRs report results for multiple effects, our sample comprises 70 separate meta-analyses for which we estimated population heterogeneity. See Supplementary Methods, Section 1, for details of the included studies.

Design heterogeneity. To the best of our knowledge, there are only two studies[59,60] that vary the experimental design to test the same hypothesis in random subsamples to isolate design heterogeneity in a random-effects meta-analysis. Both studies, reporting results on six different hypotheses, are included in our analysis of design heterogeneity.

The five hypotheses examined in the study by Landy et al.[59] were tested in a main study and an independent replication study each, implying that the number of estimates on design heterogeneity from that study is 10, and the total number of estimates is 11). See Supplementary Methods, Section 2, for more details of the included studies.

Analytical heterogeneity. We reviewed all multi-analyst studies in the social sciences for which data on effect sizes and standard errors are available for each analyst, and effect sizes are measured in the same units across analysts. We found three papers that meet our criteria: Silberzahn et al.[61], Huntington-Klein et al.[62], and Hoogeveen et al.[63]. In total, these papers examine analytical variability for five different hypotheses. We identified five more published multi-analyst studies in the social sciences, but these did not meet our inclusion criteria: Bastiaansen et al.[89] detail the variation in analytic decisions across analysis teams but do not report estimates pertaining to each of the proposed analysis pipelines; Botvinik-Nezer et al.[90] was excluded as the primary outcome reported by analysis teams is a binary classification of whether the hypotheses are supported by the data, but no effect size measure is reported; Schweinsberg et al.[91] was excluded since the individual results by analysts are not available in standardized effect-size units but only in terms of z-scores; Menkveld et al.[92] was excluded as the data is yet embargoed; Breznau et al.[93] was excluded as the research teams reported various results for the same hypothesis and it is not clear which effect size estimate to include for each team. Note that the reported variation in results is very large across analysts also in the five excluded studies. See Supplementary Methods, Section 3, for more details of the included studies.

Estimation of results for included studies

For each included study, we re-estimated the random-effects meta-analytic models based on the original data. In Supplementary Table 1, we provide detailed results for each included meta-analysis, comprising the Q-test, whether effect sizes were measured in Cohen's d units, the between-study variation (τ and its 95% CI), the within-study variation (σ), the ratio between the between- and within-study variation (τ/σ; which we refer to as the heterogeneity ratio HR), I^2 and its 95% CI, and H and its 95% CI. If not indicated otherwise in Supplementary Methods, Sections 1–3, we were able to precisely (computationally) reproduce the results reported in the papers. As such, our study provides—as a "side product"—evidence on the computational reproducibility[94–98] of large-scale meta-scientific results.

To keep things simple and easily replicable, we created copies of the relevant input data for the meta-analyses (i.e., the effect size estimate and the corresponding standard error for each study included in the meta-analysis) for each paper based on the original data (all of which are publicly available under a CC-by license). These copies of the original

data constitute our raw data. All data and analysis scripts used to generate the results reported in the main text and the SI are available at our project's OSF repository: osf.io/yegsx.

Random effects meta-analyses were estimated using the `metafor` package (v-4.4.0)[99] in `R` (v-4.3.2)[100]. Estimates for the confidence intervals around the heterogeneity measures τ^2, I^2, and H^2 were based on the Q-profile method[101] implemented using the `confint()` function shipped with the `metafor` package. For all papers reporting the results of meta-analyses (i.e., all papers on population or design heterogeneity), we used the same estimator for τ^2 as used in the original paper. The majority of these papers relied on the restricted maximum likelihood estimator[102]; but one study used the DerSimonian-Laird[103] estimator, and one study used the Hartung-Knapp[104] estimator.

For multi-analyst studies, heterogeneity estimates were based on the restricted maximum likelihood estimator. Note that estimating a random-effects model on multi-analyst-style data is unconventional, as discussed in the main text. Hence it does not come as a surprise that none of the multi-analyst studies included in our review did report the results of a random-effects meta-analysis; however, a recent multi-analyst study in biology[64] used a meta-analytic random-effects model to estimate the heterogeneity of results across analysts. The estimated heterogeneity measures τ^2, I^2, and H^2 for multi-analyst studies can be interpreted as lower bound estimates as they are derived based on the within-study variance that would be observed if the effect size estimates reported by multiple analysts were independent observations; if the sampling variances of the multiple analysts are correlated (which is the case for multi-analyst studies, since analysts based their estimates on the same dataset), the actual within-study variance is lower, and the between-study variance is higher.

Robustness tests on Analytical Heterogeneity

Huntington-Klein et al.[62] introduced the ratio between the standard deviation of effect size estimates across analysts and the mean standard error as a measure of the analytical heterogeneity in multi-analyst studies. This measure can be interpreted as a proxy for the ratio of the between-study variation and the within-study variation (HR) and can be converted to a proxy measure of H by taking the square root of 1 plus the squared ratio; to distinguish the two measures from HR and H obtained from the estimates of random-effects meta-analyses, we denote them as HR_P and H_P. In Supplementary Table 2, we report both HR_P and H_P for the multi-analyst studies included in our review. HR_P varies between 1.48 and 3.98 for the multi-analyst studies, with a median of 3.07; H_P varies between 1.79 and 4.11 for the multi-analyst studies, with a median of 3.23. Note that while the between-study variation (τ) estimated in the

random-effects meta-analysis is a lower bound of the standard deviation in effect sizes across analysts, the proxy H_P may exceed H as estimated using a random-effects meta-analysis as the estimated within-study variation (σ) in the random-effects meta-analysis can differ from the average standard error of estimates generated by the analysts. The proxy H_P is quite similar to H based on the random-effects meta-analysis for four of the five multi-analyst estimates, but differs substantially for Silberzahn et al.[61]. This is due to two outliers in terms of low standard errors strongly affecting the estimated within-study variance in the random-effects meta-analysis (as the individual effects are weighted by the inverse of their variance). This is an indication that the estimate of H for of Silberzahn et al.[61] should be interpreted very cautiously; but also the proxy H_P indicates substantial analytical heterogeneity. Removing the two outliers for Silberzahn et al.[61] (implying $k = 27$ effect size estimates) results in the following heterogeneity estimates in a random-effect meta-analysis: $Q(26) = 130.3$, $p < 0.001$; $\tau = 0.107$, $I^2 = 86.5\%$, $H = 2.717$, $HR = 2.527$; the Huntington-Klein et al.[62]-based proxy measures remain qualitatively unchanged ($H_P = 1.721$, $HR_P = 1.401$).

Illustrating the impact of heterogeneity on statistical inference

Consider a generic two-tailed z-test with power π to detect an effect θ at a type-I error rate α. The effect size θ (measured in z-score units in the generic test) corresponds to the non-centrality parameter $\delta = |z_{\alpha/2}| + |z_{\beta}|$, where z_p denotes the p^{th} quantile of the inverse cumulative standard normal distribution and $\beta = 1 - \pi$ denotes the false negative rate. Assuming that the true effects to be estimated are homogeneous, $\theta_i \sim N(\mu_0, \sigma^2)$ under the null hypothesis $\mathcal{H}_0$ (with μ_0 indicating the test value and σ^2 denoting the test's sampling variance); under the alternative hypothesis $\mathcal{H}_A$, $\theta_i \sim N(\delta, \sigma^2)$.

Now suppose there is variation in the true effect size above and beyond the uncertainty that is accounted for by the test's sampling variance (σ_i^2). Put differently, effect size estimates are subject to an additional source of uncertainty—heterogeneity—such that the overall variance of study i's estimate θ_i is given by $v_i^2 = \sigma^2 + \tau^2$. The heterogeneity estimate τ^2 indicates the variance of the genuine effect, such that $\theta_i \sim N(\mu_0, v_i^2)$ under $\mathcal{H}_0$ and $\theta_i \sim N(\delta, v_i^2)$ under $\mathcal{H}_A$.

Instead of quantifying the extent of heterogeneity in absolute terms (i.e., in terms of τ^2 or τ, respectively), it is expedient to denote it relative to the test's sampling variance (σ_i^2). Following the notational conventions applicable to random-effects meta-analysis[41], we define a heterogeneity factor H as

$$H = \sqrt{\frac{\sigma^2 + \tau^2}{\sigma^2}},$$

which is equivalent to the square root of H^2, a commonly used measure of heterogeneity reported in meta-analyses. H^2 can be interpreted as a "variance inflation factor" due to heterogeneity[42,43], i.e., as the factor a test's sampling error σ_i needs to be multiplied with to incorporate the uncertainty due to heterogeneity into statistical inference. We prefer H over H^2 as thinking about heterogeneity in terms of standard deviation units appears more convenient than thinking about it in terms of variance units. H^2 is defined as the relative excess of the Q-statistic over its degrees of freedom, i.e., $H^2 = Q / (k - 1)$. Following conventions, we presume $H = \max(1, H)$, though we acknowledge that this prevents identifying excessive homogeneity—i.e., less variability than would be expected due to chance.[42]

The effective false positive rate α' in a two-tailed z-test in the presence of heterogeneity (expressed in terms of the heterogeneity factor H) is given by

$$\alpha' = 2 \cdot \Phi\left(\frac{z_{\alpha/2}}{H}\right),$$

where $\Phi(\cdot)$ indicates the cumulative standard normal distribution and $z_{\alpha/2}$ denotes the $\alpha/2$-percentile of the inverse cumulative standard normal density function $\Phi^{-1}(\cdot)$ (i.e., the critical value of a two-tailed z-test at a nominal significance threshold α). It follows that $\alpha' > \alpha$ for any $H > 1$. Correspondingly, the effective false negative rate β' is given by

$$\beta' = \Phi\left(\frac{z_{\beta}}{H}\right),$$

which implies that $\beta' > \beta$ for any $H > 1$ whenever $\beta < 0.5$ and $\beta' < \beta$ for any H whenever $\beta > 0.5$. The relationship between nominal and effective error rates is graphically illustrated in Fig. 2 and Fig. 3.

The false discovery rate (*FDR*) is defined as the ratio of false positive results to the total number of positive classifications, which implies that the *FDR* is a function of the prior probability ϕ for the alternative hypothesis being genuinely true, i.e.,

$$FDR = \frac{(1 - \phi) \cdot \alpha}{(1 - \phi) \cdot \alpha + \phi \cdot \beta}.$$

Since heterogeneity inflates the effective type-I error rate (for any nominal α-level in a two-tailed test) and the effective type-II error rate (for a nominal $\beta > 0.5$), it follows that the effective false discovery rate *FDR'* is given by

$$FDR' = \frac{(1 - \phi) \cdot 2 \cdot \Phi\left(\frac{z_{\alpha/2}}{H}\right)}{(1 - \phi) \cdot 2 \cdot \Phi\left(\frac{z_{\alpha/2}}{H}\right) + \phi \cdot \Phi\left(\frac{z_{\beta}}{H}\right)}.$$

Since the cumulative normal density $\Phi(\cdot)$ is convex in the domain $(-\infty, 0]$, it follows that $FDR' > FDR$ for any $H > 1$. Figure 4 illustrates the effective false discovery rate *FDR'* as a function of H for various levels of the prior ϕ and different significance thresholds α.

Data and Code Availability

The data used to estimate population, design, and analytical heterogeneity and the analysis scripts generating all results, figures, and tables reported in the main text and the Supplementary Information are available at the project's OSF repository (osf.io/yegsx).

Acknowledgments

For financial support, we thank the Austrian Science FWF (grant SFB F63), Jan Wallander and Tom Hedelius Foundation (grants P21-0091 and P23-0098), Knut and Alice Wallenberg Foundation and Marianne and Marcus Wallenberg Foundation (Wallenberg Scholar grant to A.D.), and Riksbankens Jubileumsfond (grant P21-0168).

Author Contributions

Conceptualization: R.B., A.D., F.H., J.H., M.J., M.K.; *Methodology:* F.H., M.J.; *Data Curation:* F.H.; *Investigation:* F.H., M.J.; *Formal Analysis:* F.H.; *Visualization:* F.H.; *Supervision:* M.J., M.K.; *Writing—Original Draft:* F.H., M.J., M.K.; *Writing—Review & Editing:* R.B., A.D., F.H., J.H., M.J., M.K.

Competing Interest Statement

The authors declare no competing interests.

References

1. Wicherts, J. M. *et al*. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Front. Psychol*. 7, (2016).
2. Gelman, A. & Loken, E. The statistical crisis in science. *Am. Sci*. 102, 460 (2014).
3. Martinson, B. C., Anderson, M. S. & de Vries, R. Scientists behaving badly. *Nature* 435, 737–738 (2005).
4. Breeding cheats. *Nature* 445, 242–243 (2007).
5. Bakker, M., van Dijk, A. & Wicherts, J. M. The rules of the game called psychological science. *Perspect. Psychol. Sci*. 7, 543–554 (2012).
6. Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci*. 22, 1359–1366 (2011).
7. Brodeur, A., Lé, M., Sangnier, M. & Zylberberg, Y. Star wars: The empirics strike back. *Am. Econ. J. Appl. Econ*. 8, 1–32 (2016).

8. Brodeur, A., Cook, N. & Heyes, A. Methods matter: p-hacking and publication bias in causal analysis in economics. *Am. Econ. Rev.* 110, 3634–3660 (2020).

9. Stefan, A. M. & Schönbrodt, F. D. Big little lies: A compendium and simulation of p-hacking strategies. *R. Soc. Open Sci.* 10, 220346 (2023).

10. Ioannidis, J. P. A. Why most published research findings are false. *PLoS Med.* 2, e124 (2005).

11. Nuzzo, R. Scientific method: Statistical errors. *Nature* 506, 150–152 (2014).

12. DeCoster, J., Sparks, E. A., Sparks, J. C., Sparks, G. G. & Sparks, C. W. Opportunistic biases: Their origins, effects, and an integrated solution. *Am. Psychol.* 70, 499–514 (2015).

13. Ioannidis, J. P. A. Why most discovered true associations are inflated. *Epidemiology* 19, 640 (2008).

14. Simonsohn, U., Nelson, L. D. & Simmons, J. P. p-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results. *Perspect. Psychol. Sci.* 9, 666–681 (2014).

15. van Aert, R. C. M., Wicherts, J. M. & van Assen, M. A. L. M. Conducting meta-analyses based on p-values: Reservations and recommendations for applying p-uniform and p-curve. *Perspect. Psychol. Sci.* 11, 713–729 (2016).

16. Hedges, L. V. Modeling Publication Selection Effects in Meta-analysis. *Stat. Sci.* 7, 246–255 (1992).

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

18. Franco, A., Malhotra, N. & Simonovits, G. Underreporting in psychology experiments: Evidence from a study registry. *Soc. Psychol. Personal. Sci.* 7, 8–12 (2016).

19. Button, K. S. *et al.* Power failure: Why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* 14, 365–376 (2013).

20. Fraley, R. C. & Vazire, S. The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLoS One* 9, e109019 (2014).

21. Dumas-Mallet, E., Smith, A., Boraud, T. & Gonon, F. Poor replication validity of biomedical association studies reported by newspapers. *PLoS One* 12, e0172650 (2017).

22. Ioannidis, J. P. A., Stanley, T. D. & Doucouliagos, H. The power of bias in economics research. *Econ. J.* 127, F236–F265 (2017).

23. Kerr, N. L. HARKing: Hypothesizing After the Results are Known. *Personal. Soc. Psychol. Rev.* 2, 196–217 (1998).

24. Rubin, M. When does HARKing hurt? Identifying when different types of undisclosed post hoc hypothesizing harm scientific progress. *Rev. Gen. Psychol.* 21, 308–320 (2017).

25. Bishop, D. Rein in the four horsemen of irreproducibility. *Nature* 568, 435–435 (2019).

26. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* 349, aac4716 (2015).

27. Camerer, C. F. *et al*. Evaluating replicability of laboratory experiments in economics. *Science* 351, 1433–1436 (2016).

28. Camerer, C. F. *et al*. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nat. Hum. Behav.* 2, 637–644 (2018).

29. Munafò, M. R. *et al*. A manifesto for reproducible science. *Nat. Hum. Behav.* 1, 0021 (2017).

30. Nosek, B. A., Ebersole, C. R., DeHaven, A. C. & Mellor, D. T. The preregistration revolution. *Proc. Natl. Acad. Sci.* 115, 2600–2606 (2018).

31. Aczel, B. *et al*. A consensus-based transparency checklist. *Nat. Hum. Behav.* 4, 4–6 (2020).

32. Armeni, K. *et al*. Towards wide-scale adoption of open science practices: The role of open science communities. *Sci. Public Policy* 48, 605–611 (2021).

33. Ferguson, J. *et al*. Survey of open science practices and attitudes in the social sciences. *Nat. Commun.* 14, 5401 (2023).

34. Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J. & Kievit, R. A. An agenda for purely confirmatory research. *Perspect. Psychol. Sci.* 7, 632–638 (2012).

35. Chambers, C. D. & Tzavella, L. The past, present and future of Registered Reports. *Nat. Hum. Behav.* 6, 29–42 (2021).

36. Baribault, B. *et al*. Metastudies for robust tests of theory. *Proc. Natl. Acad. Sci.* 115, 2607–2612 (2018).

37. Yarkoni, T. The generalizability crisis. *Behav. Brain Sci.* 45, e1 (2020).

38. Bryan, C. J., Tipton, E. & Yeager, D. S. Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nat. Hum. Behav.* 5, 980–989 (2021).

39. Almaatouq, A. *et al*. Beyond playing 20 questions with nature: Integrative experiment design in the social and behavioral sciences. *Behav. Brain Sci.* 1–55 (2022) doi:10/grzgz4.

40. Delios, A. *et al*. Examining the generalizability of research findings from archival data. *Proc. Natl. Acad. Sci.* 119, e2120377119 (2022).

41. Borenstein, M., Hedges, L. V., Higgins, J. P. T. & Rothstein, H. R. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Res. Synth. Methods* 1, 97–111 (2010).

42. Higgins, J. P. T. & Thompson, S. G. Quantifying heterogeneity in a meta-analysis. *Stat. Med.* 21, 1539–1558 (2002).

43. Mittlböck, M. & Heinzl, H. A simulation study comparing properties of

heterogeneity measures in meta-analyses. *Stat. Med.* 25, 4321–4333 (2006).

44. Higgins, J. P. T., Thompson, S. G., Deeks, J. J. & Altman, D. G. Measuring inconsistency in meta-analyses. *BMJ* 327, 557–560 (2003).
45. Pigott, T. D. *Advances in Meta-Analysis*. (Springer US, 2012).
46. Klein, R. A. *et al.* Investigating variation in replicability: A "many labs" replication project. *Soc. Psychol.* 45, 142–152 (2014).
47. Ebersole, C. R. *et al.* Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *J. Exp. Soc. Psychol.* 67, 68–82 (2016).
48. Klein, R. A. *et al.* Many Labs 2: Investigating variation in replicability across samples and settings. *Adv. Methods Pract. Psychol. Sci.* 1, 443–490 (2018).
49. Klein, R. A. *et al.* Many Labs 4: Failure to replicate mortality salience effect with and without original author involvement. *Collabra Psychol.* 8, 35271 (2022).
50. Alogna, V. K. *et al.* Registered Replication Report: Schooler and Engstler-Schooler (1990). *Perspect. Psychol. Sci.* 9, 556–578 (2014).
51. Cheung, I. *et al.* Registered Replication Report: Study 1 From Finkel, Rusbult, Kumashiro, & Hannon (2002). *Perspect. Psychol. Sci.* 11, 750–764 (2016).
52. Eerland, A. *et al.* Registered Replication Report: Hart & Albarracín (2011). *Perspect. Psychol. Sci.* 11, 158–171 (2016).
53. Hagger, M. S. *et al.* A multilab preregistered replication of the ego-depletion effect. *Perspect. Psychol. Sci.* 11, 546–573 (2016).
54. Wagenmakers, E.-J. *et al.* Registered Replication Report: Strack, Martin, & Stepper (1988). *Perspect. Psychol. Sci.* 11, 917–928 (2016).
55. Bouwmeester, S. *et al.* Registered Replication Report: Rand, Greene, and Nowak (2012). *Perspect. Psychol. Sci.* 12, 527–542 (2017).
56. McCarthy, R. J. *et al.* Registered Replication Report on Srull and Wyer (1979). *Adv. Methods Pract. Psychol. Sci.* 1, 321–336 (2018).
57. O'Donnell, M. *et al.* Registered Replication Report: Dijksterhuis and van Knippenberg (1998). *Perspect. Psychol. Sci.* 13, 268–294 (2018).
58. Verschuere, B. *et al.* Registered Replication Report on Mazar, Amir, and Ariely (2008). *Adv. Methods Pract. Psychol. Sci.* 1, 299–317 (2018).
59. Landy, J. F. *et al.* Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychol. Bull.* 146, 451–479 (2020).
60. Huber, C. *et al.* Competition and moral behavior: A meta-analysis of forty-five crowd-sourced experimental designs. *Proc. Natl. Acad. Sci.* 120, e2215572120 (2023).
61. Silberzahn, R. *et al.* Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Adv. Methods Pract. Psychol. Sci.* 1, 337–356 (2018).
62. Huntington-Klein, N. *et al.* The influence of hidden researcher decisions in applied microeconomics. *Econ. Inq.* 59, 944–960 (2021).

63. Hoogeveen, S. *et al*. A many-analysts approach to the relation between religiosity and well-being. *Relig. Brain Behav.* 0, 1–47 (2022).

64. Gould, E. *et al*. Same data, different analysts: Variation in effect sizes due to analytical decisions in ecology and evolutionary biology. Preprint at https://doi.org/10.32942/X2GG62 (2023).

65. Auspurg, K. & Brüderl, J. Has the credibility of the social sciences been credibly destroyed? Reanalyzing the "Many Analysts, One Data Set" project. *Socius Sociol. Res. Dyn. World* 7, 237802312110244 (2021).

66. Auspurg, K. & Brüderl, J. Is social research really not better than alchemy? How many-analysts studies produce "a hidden universe of uncertainty" by not following meta-analytical standards. Preprint at https://doi.org/10.31222/osf.io/uc84k (2023).

67. Benjamin, D. J. *et al*. Redefine statistical significance. *Nat. Hum. Behav.* 2, 6–10 (2018).

68. Kenny, D. A. & Judd, C. M. The unappreciated heterogeneity of effect sizes: Implications for power, precision, planning of research, and replication. *Psychol. Methods* 24, 578–589 (2019).

69. Smaldino, P. E. & McElreath, R. The natural selection of bad science. *R. Soc. Open Sci.* 3, 160384 (2016).

70. Grainger, M. J., Bolam, F. C., Stewart, G. B. & Nilsen, E. B. Evidence synthesis for tackling research waste. *Nat. Ecol. Evol.* 4, 495–497 (2020).

71. Forscher, P. S. *et al*. The Benefits, Barriers, and Risks of Big-Team Science. *Perspect. Psychol. Sci.* 18, 607–623 (2023).

72. van Erp, S., Verhagen, J., Grasman, R. P. P. P. & Wagenmakers, E.-J. Estimates of between-study heterogeneity for 705 meta-analyses reported in Psychological Bulletin from 1990–2013. 5, 4 (2017).

73. Stanley, T. D., Carter, E. C. & Doucouliagos, H. What meta-analyses reveal about the replicability of psychological research. *Psychol. Bull.* 144, 1325–1346 (2018).

74. Kvarven, A., Strømland, E. & Johannesson, M. Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nat. Hum. Behav.* 4, 423–434 (2020).

75. Olsson-Collentine, A., Wicherts, J. M. & van Assen, M. A. L. M. Heterogeneity in direct replications in psychology and its association with effect size. *Psychol. Bull.* 146, 922–940 (2020).

76. Linden, A. H. & Hönekopp, J. Heterogeneity of research results: A new perspective from which to assess and promote progress in psychological science. *Perspect. Psychol. Sci.* 16, 358–376 (2021).

77. Jarke, H. *et al*. A Roadmap to Large-Scale Multi-Country Replications in Psychology. *Collabra Psychol.* 8, 57538 (2022).

78. McShane, B. B., Tackett, J. L., Böckenholt, U. & Gelman, A. Large-scale replication

projects in contemporary psychological research. *Am. Stat.* 73, 99–105 (2019).

79. Milkman, K. L. *et al.* A megastudy of text-based nudges encouraging patients to get vaccinated at an upcoming doctor's appointment. *Proc. Natl. Acad. Sci.* 118, e2101165118 (2021).

80. Milkman, K. L. *et al.* Megastudies improve the impact of applied behavioural science. *Nature* 600, 478–483 (2021).

81. Milkman, K. L. *et al.* A 680,000-person megastudy of nudges to encourage vaccination in pharmacies. *Proc. Natl. Acad. Sci.* 119, e2115126119 (2022).

82. Uhlmann, E. L. *et al.* Scientific utopia III: Crowdsourcing science. *Perspect. Psychol. Sci.* 14, 711–733 (2019).

83. Patel, C. J., Burford, B. & Ioannidis, J. P. A. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *J. Clin. Epidemiol.* 68, 1046–1058 (2015).

84. Silberzahn, R. & Uhlmann, E. L. Crowdsourced research: Many hands make tight work. *Nature* 526, 189–191 (2015).

85. Steegen, S., Tuerlinckx, F., Gelman, A. & Vanpaemel, W. Increasing Transparency Through a Multiverse Analysis. *Perspect. Psychol. Sci.* 11, 702–712 (2016).

86. Simonsohn, U., Simmons, J. P. & Nelson, L. D. Specification curve analysis. *Nat. Hum. Behav.* 4, 1208–1214 (2020).

87. Wagenmakers, E.-J., Sarafoglou, A. & Aczel, B. One statistical analysis must not rule them all. *Nature* 605, 423–425 (2022).

88. Ebersole, C. R. *et al.* Many Labs 5: Testing pre-data-collection peer review as an intervention to increase replicability. *Adv. Methods Pract. Psychol. Sci.* 3, 309–331 (2020).

89. Bastiaansen, J. A. *et al.* Time to get personal? The impact of researchers choices on the selection of treatment targets using the experience sampling methodology. *J. Psychosom. Res.* 137, 110211 (2020).

90. Botvinik-Nezer, R. *et al.* Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* 582, 84–88 (2020).

91. Schweinsberg, M. *et al.* Same data, different conclusions: Radical dispersion in empirical results when independent analysts operationalize and test the same hypothesis. *Organ. Behav. Hum. Decis. Process.* 165, 228–249 (2021).

92. Menkveld, A. J. *et al.* Non-standard errors. *J. Finance* forthcoming, (2023).

93. Breznau, N. *et al.* Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proc. Natl. Acad. Sci.* 119, e2203150119 (2022).

94. Stark, P. B. Before reproducibility must come preproducibility. *Nature* 557, 613–613 (2018).

95. Stodden, V., Seiler, J. & Ma, Z. An empirical analysis of journal policy effectiveness for computational reproducibility. *Proc. Natl. Acad. Sci.* 115, 2584–2589 (2018).

96. Trisovic, A., Lau, M. K., Pasquier, T. & Crosas, M. A large-scale study on research code quality and execution. *Sci. Data* 9, 60 (2022).

97. Dreber, A. & Johannesson, M. A framework for evaluating reproducibility and replicability in economics. SSRN Preprint at https://doi.org/10.2139/ssrn.4458153 (2023).

98. Pérignon, C. *et al*. Computational Reproducibility in Finance: Evidence from 1,000 Tests. SSRN Preprint at https://doi.org/10.2139/ssrn.4064172 (2023).

99. Viechtbauer, W. Conducting meta-analyses in R with the metafor package. *J. Stat. Softw.* 36, 1–48 (2010).

100. R Core Team. *R: A language and environment for statistical computing*. https://www.R-project.org/ (2022).

101. Viechtbauer, W. Confidence intervals for the amount of heterogeneity in meta-analysis. *Stat. Med.* 26, 37–52 (2007).

102. Viechtbauer, W. Bias and efficiency of meta-analytic variance estimators in the random-effects model. *J. Educ. Behav. Stat.* 30, 261–293 (2005).

103. DerSimonian, R. & Laird, N. Meta-analysis in clinical trials. *Control. Clin. Trials* 7, 177–188 (1986).

104. Hartung, J. & Knapp, G. On tests of the overall treatment effect in meta-analysis with normally distributed responses. *Stat. Med.* 20, 1771–1782 (2001).

Figures

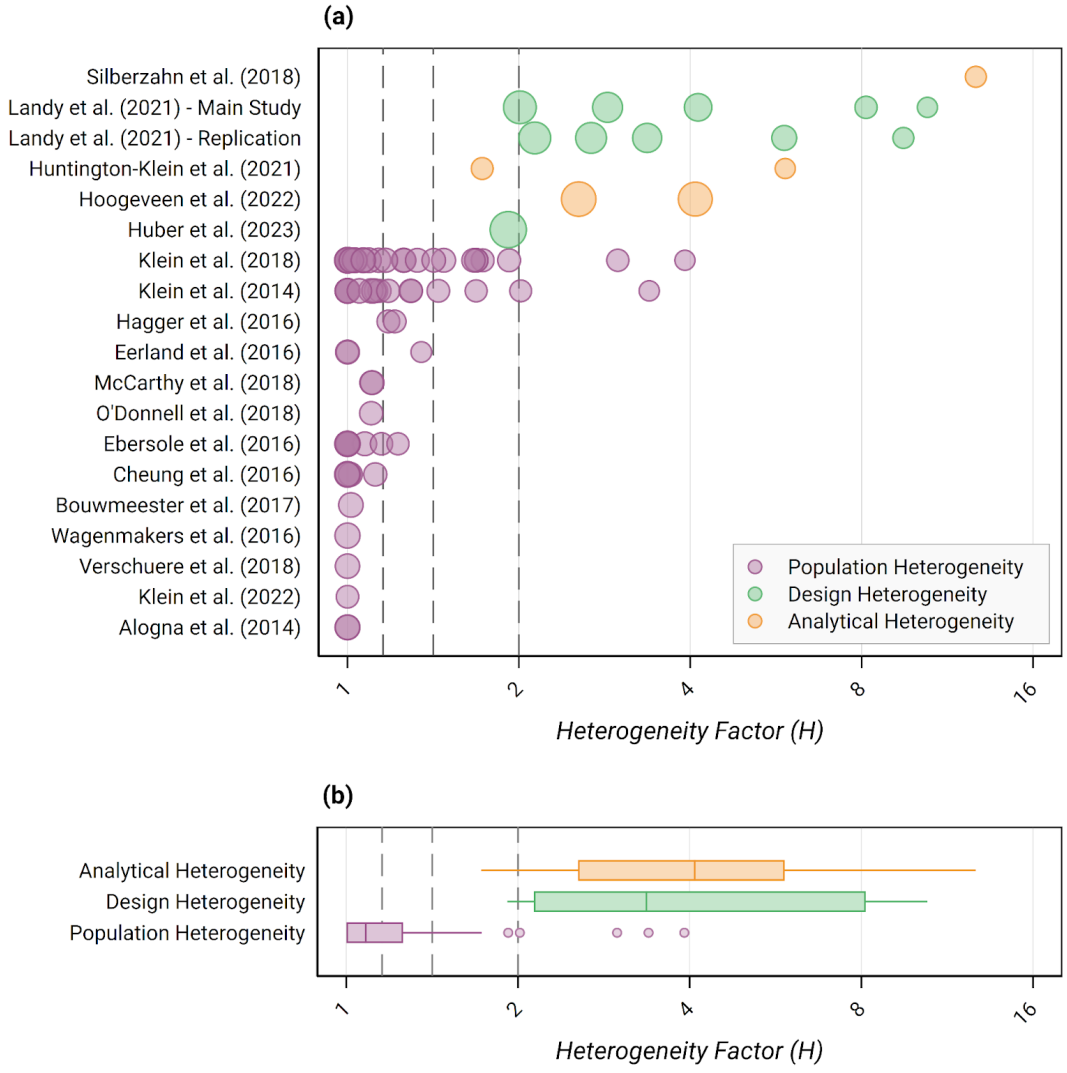

Fig. 1. Empirical estimates of population, design, and analytical heterogeneity. **(a)** The figure shows estimates of the heterogeneity factor H for 70 estimates from 13 papers isolating population heterogeneity[46–58], 11 estimates from 2 papers isolating design heterogeneity[59,60], and 5 estimates from 3 papers isolating analytical heterogeneity[61–63]. The vertical reference lines indicate benchmark levels for small, medium, and large heterogeneity based on I^2 values of 25% (H = 1.15), 50% (H = 1.41), and 75% (H = 2), respectively. **(b)** The figure shows box plots of the distribution of heterogeneity factors H, separated by the source of heterogeneity, illustrated in panel (a).

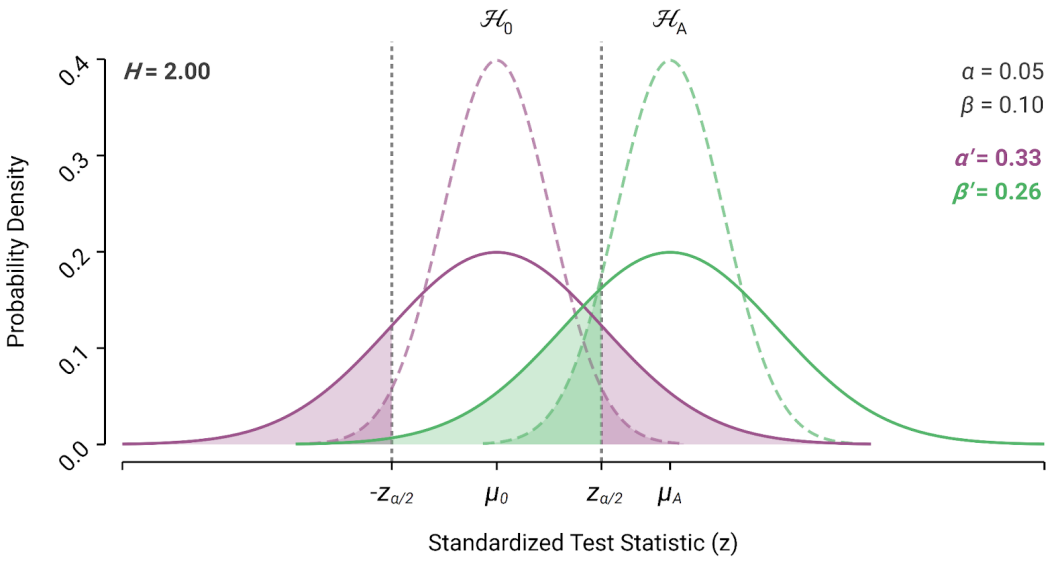

Fig. 2. The figure shows the probability density function of an effect under the null hypothesis ($\mathcal{H}_0$; purple density functions) and the alternative hypothesis ($\mathcal{H}_A$; green density functions) for a two-tailed z-test with 90% nominal power (π) at a 5% nominal significance level (α) assuming homogeneity (i.e., H = 1; dashed lines) and the implications of disregarded heterogeneity of H = 2.0 (solid lines) on the effective type-I error rate α' and statistical power π'. Areas shaded in purple indicate the test's nominal and effective false positive rates (α and α'); areas shaded in green correspond to the test's nominal and effective false negative rates (β and β').

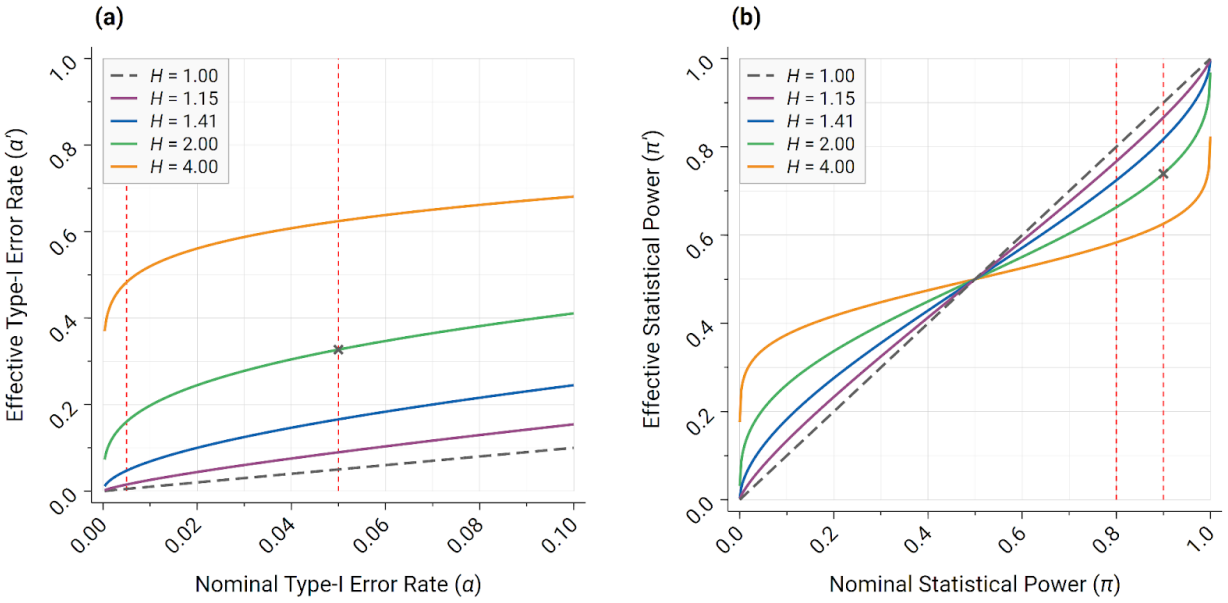

Fig. 3. (a) The figure illustrates the effective type-I error rates α' as a function of the nominal type-I error rate α for various levels of heterogeneity. $H = 1$ implies the absence of heterogeneity; $H = 1.15$, $H = 1.41$, and $H = 2.00$ correspond to the commonly used I^2 thresholds of 25%, 50%, and 75% (i.e., small, medium, and large heterogeneity); $H = 4.00$ corresponds to "extreme" heterogeneity (equivalent to $I^2 = 93.75\%$). The dashed vertical lines indicate the 5% and 0.5% nominal significance thresholds. **(b)** The figure illustrates the effective statistical power (π') as a function of nominal statistical power (π) for the same values of the heterogeneity factor H as shown in (a). The dashed vertical lines indicate the 80% and 90% nominal statistical power levels. The x-markers in both panels map the values in the generic example illustrated in Fig. 1.

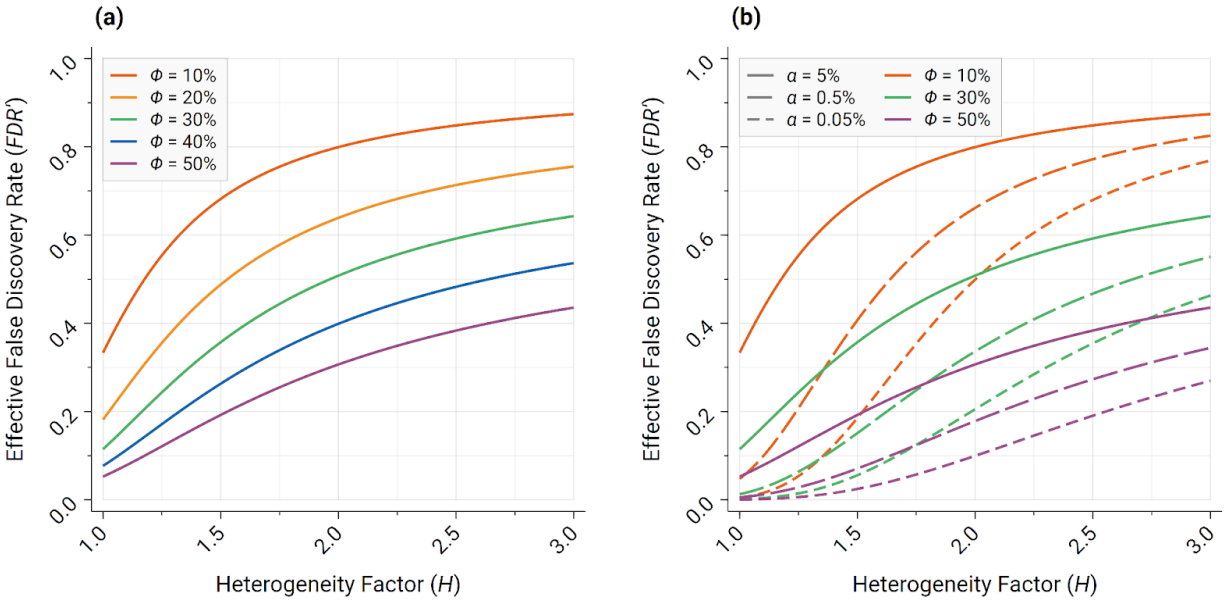

Fig. 4. (a) The panel illustrates the effective false discovery rate (*FDR'*), i.e., the ratio of false positive results to the total number of positive classifications in the presence of heterogeneity, for different prior probabilities for the alternative hypothesis being genuinely true (ϕ), as a function of the heterogeneity factor H for a two-tailed z-test with nominal statistical power of π = 90%. **(b)** The panel illustrates the *FDR'*, for different prior probabilities ϕ and various significance thresholds α, as a function of the heterogeneity factor H for a two-tailed z-test with nominal statistical power of π = 90%.