ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2020)
183, *Part* 3, *pp.* 1145–1166

New statistical metrics for multisite replication projects

Maya B. Mathur

Stanford University, USA

and Tyler J. VanderWeele

Harvard University, Boston, USA

[Received April 2018. Final revision March 2020]

Summary. Increasingly, researchers are attempting to replicate published original studies by using large, multisite replication projects, at least 134 of which have been completed or are on going. These designs are promising to assess whether the original study is statistically consistent with the replications and to reassess the strength of evidence for the scientific effect of interest. However, existing analyses generally focus on single replications; when applied to multisite designs, they provide an incomplete view of aggregate evidence and can lead to misleading conclusions about replication success. We propose new statistical metrics representing firstly the probability that the original study's point estimate would be at least as extreme as it actually was, if in fact the original study were statistically consistent with the replications, and secondly the estimated proportion of population effects agreeing in direction with the original study. Generalized versions of the second metric enable consideration of only meaningfully strong population effects that agree in direction, or alternatively that disagree in direction, with the original study. These metrics apply when there are at least 10 replications (unless the heterogeneity estimate $\hat{\tau} = 0$, in which case the metrics apply regardless of the number of replications). The first metric assumes normal population effects but appears robust to violations in simulations; the second is distribution free. We provide R packages (Replicate and MetaUtility).

Keywords: Effect sizes; Heterogeneity; Replication; Reproducibility

1. Introduction

Several social science disciplines have recently moved to assess replicability of the published literature empirically through systematic, third-party replications. Investigators conducting replications often seek to assess, firstly, how similar the results of the replication studies are to those of the original studies, i.e. the extent to which the original studies are statistically consistent or inconsistent with their replications (Anderson and Maxwell, 2016). Second, investigators often aim to use replications to reassess the strength of evidence for the scientific effect under investigation (Anderson and Maxwell, 2016), ideally while minimizing bias (e.g. through protocol and analysis preregistration and *a priori* editorial approval (Simons *et al.*, 2014) and while ensuring high statistical power.

Novel designs for replication research now exist to address these objectives with more sophistication than simple designs involving a single replication of an original study. Some high impact experimental psychology journals now encourage projects in which multiple independent sites

Address for correspondence: Maya B. Mathur, Quantitative Sciences Unit, Stanford University, 1701 Page Mill Road, Palo Alto, CA 94304, USA.
E-mail: mmathur@stanford.edu

attempt to replicate a single, published original study by using a standardized experimental protocol closely approximating the original and developed with input from the original authors (Simons *et al.*, 2014). Extensions (sometimes called 'Many Labs' projects) select multiple original studies and subject each to a multisite replication (Ebersole *et al.*, 2016; Klein *et al.*, 2014), and others have applied a similar approach to replicate new original research before its publication (Schweinsberg *et al.*, 2016). We use the term 'many-to-one design' to refer generically to any design in which an original study is replicated in multiple sites. Many-to-one replication research is a nascent, but rapidly expanding, field: we are aware of at least 79 completed and 55 on-going many-to-one replication studies to date, all completed or initiated since 2014 and in experimental psychology and experimental philosophy alone (completed, Alogna *et al.* (2014), Bouwmeester *et al.* (2017), Cheung *et al.* (2016), Cova *et al.* (2018), Ebersole *et al.* (2016), Eerland *et al.* (2016), Hagger *et al.* (2016), Klein *et al.* (2014), Schweinsberg *et al.* (2016) and Wagenmakers *et al.* (2016); on going, Association for Psychological Science (2018), Ebersole *et al.* (2018), Klein *et al.* (2018) and Schweinsberg and Uhlmann (2018)).

However, the adoption of many-to-one designs in the social sciences has outpaced the development of corresponding statistical analyses. Existing work (Andrews and Kasy, 2019; Etz and Vandekerckhove, 2016; Patil *et al.*, 2016; Simonsohn, 2015; Verhagen and Wagenmakers, 2014) has proposed analytic approaches for a single replication of a single study or designs in which numerous original studies across a discipline or domain are each replicated once (here termed 'one-to-one designs'), as in Open Science Collaboration (2015) and in Camerer *et al.* (2016). Other potentially relevant work has not been adapted to the replication context (Gadbury and Iyer, 2000; Gadbury *et al.*, 2001; Heckman *et al.*, 1997; Longford, 1999). However, many-to-one designs pose unique statistical challenges and opportunities. Results of many-to-one replications often suggest effect heterogeneity across sites despite use of standardized protocols (for example, eight of 16 replications in Klein *et al.* (2014) suggested 'statistically significant' evidence of heterogeneity), yet current analysis approaches do not adequately account for heterogeneity. As we shall discuss, this can lead to misleading assessments of consistency between the original study and the replications and of the strength of evidence for the effect under investigation. Additionally, results of many-to-one designs often lead to unresolved debates regarding the extent to which the original study 'replicated' or 'did not replicate', but these debates remain highly speculative, perhaps partly because few directly relevant quantitative metrics are currently available.

We therefore propose new statistical metrics that are specifically designed for many-to-one designs. To assess statistical consistency, we provide a metric (P_{orig}) representing the probability that the effect estimate from the original study would be as extreme or more extreme than it actually was if, in fact, the original study and the replications were statistically consistent in the sense of being drawn from the same distribution. To assess the strength of evidence, we provide a metric ($\hat{P}_{>0}$) estimating the proportion of population effects agreeing in direction with the original effect estimate. Because replication effects that agree in direction with the original, but are very weak, may in fact be considered insufficient evidence to support the original effect, we also demonstrate how to generalize this metric to consider the proportion of population effects that not only agree in direction with the original but are also stronger than a user-chosen threshold of meaningfully strong size ($\hat{P}_{>q}$). Lastly, we also provide a counterpart metric estimating the proportion of population effects in the *opposite* direction of the original ($\hat{P}_{<q^*}$). In contrast with existing metrics, the metrics proposed account for all relevant sources of statistical uncertainty in many-to-one replication designs, including heterogeneity (Kenny and Judd, 2019), and they harness the specific strengths of many-to-one designs. These metrics are mathematically straightforward but, to the best of our knowledge, have not yet been reported in

any published many-to-one replication. We provide R functions in the packages `Replicate` and `MetaUtility` to conduct all the analyses proposed.

2. Applied example

As a running example, we shall consider one of several many-to-one replication attempts that were conducted by Ebersole *et al.* (2016). Specifically, each of 21 independent laboratories used a common protocol to replicate a classical psychology experiment (Monin and Miller's (2001) experiment 1) on 'moral credentialling' theory, which proposes that people who are given an initial opportunity to demonstrate that they are not prejudiced (and thus establish 'moral credentials') are more likely to display apparently prejudiced attitudes in subsequent tasks (having licensed themselves to do so because of their previously established credentials). In the replicated experiment, the initial task required subjects to agree or disagree with potentially sexist statements. In the initial task, subjects were randomized to a credentialling condition in which the statements described 'most women' (e.g. 'Most women need a man to protect them') or to a control condition, in which the same statements described only 'some women'. Thus, credentialling statements were designed to induce higher disagreement than control statements, allowing subjects in the former condition to establish themselves more clearly as non-sexists. The dependent variable was subjects' degree of preference for male candidates in an imagined hiring scenario. As predicted, subjects in the credentialling condition more strongly preferred to hire male candidates than did control subjects (corresponding to an effect size of $r = 0.21$ on Pearson's correlation scale (95% confidence interval (CI) 0.09, 0.32; $p = 7 \times 10^{-4}$). Monin and Miller (2001) also reported an unexpected interaction of credentialling condition with the subject's sex, and Ebersole *et al.* (2016) attempted to replicate both the main effect and the interaction. For brevity, we focus on the main effect only.

3. Existing metrics

We first review metrics that are commonly reported in many-to-one designs as well as those developed for other designs but that are frequently reported in many-to-one designs. First, nearly all many-to-one designs report a pooled estimate of the effect size in the replications. The pooled estimate is usually estimated by meta-analysing effect sizes from the replications or by fitting a mixed model to individual subject data. For example, fitting a random-effects meta-analysis model to replication studies of Ebersole *et al.* (2016) on moral credentialling estimates an average effect size of 0.07 (95% CI 0.02, 0.12) on Pearson's correlation scale; both the replicators and the lead author of the original study (Monin, 2016) interpreted this finding as a successful replication supporting moral credentialling. Regardless of modelling approach, this metric estimates the average population effect size across the replications. This is adequate if replications exhibit little heterogeneity but provides an incomplete picture in the presence of heterogeneity across replication studies. Such heterogeneity may occur, for example, if replication studies differ with respect to subjects' demographic characteristics (e.g. age, sex, race or geographic region) or the setting in which the study is conducted (e.g. time of day or physical setting). As the proposed metrics will formalize, the forest plot in Fig. 1 suggests heuristically that, although a group of replication point estimates were clustered around the pooled point estimate, several point estimates were in fact in the direction opposite the original, and several were even larger than the original.

As discussed elsewhere in the context of meta-analyses rather than replications (Mathur and VanderWeele, 2019), under moderate or substantial heterogeneity, a pooled estimate near

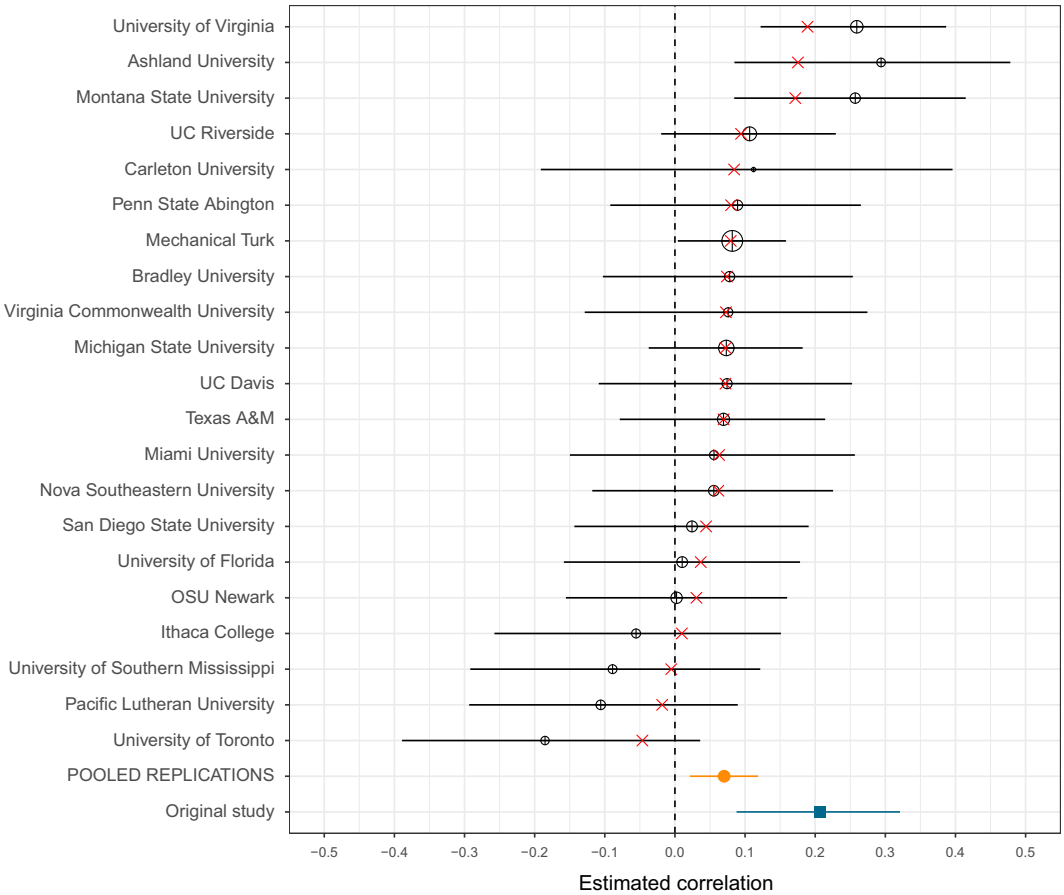

Fig. 1. Estimated correlation in each of the replications of Ebersole *et al.* (2016) (○), ordered by the calibrated estimate of the replication's population effect size (×) discussed in Section 4.2 (circle areas are proportional to the replication's relative inverse variance weight in the meta-analysis; error bars represent 95% CIs): ●, meta-analytic pooled estimate across replications; ■, estimate in the original study of Monin and Miller (2001); •, 5% weight; ●, 10% weight; ●, 15% weight

the null can belie the existence of strong effects in some replication settings. Thus because of heterogeneity, a many-to-one replication design whose pooled estimate appears not to support the hypothesized effect may nevertheless provide evidence of meaningfully strong effects in favour of the original hypothesis in some contexts (e.g. locations, subject demographics and variations in protocol administration). Conversely, if the pooled estimate is in the same direction as the original estimate but is smaller, we cannot directly discern whether the population effects are never as large as originally reported (and perhaps are too small to warrant scientific interest) or whether they may, in fact, be as large as or larger than the original estimate in some settings. For these reasons, we shall recommend supplementing the pooled point estimate with new metrics that additionally characterize heterogeneity.

A widespread metric of statistical consistency assesses whether the replication study obtains a 'statistically significant' p-value and an effect estimate in the same direction as in the original study (assuming that the original study itself obtained a 'significant' p-value). This 'significance agreement' metric is widely reported in single replications (Anderson and Maxwell, 2016), in

one-to-one designs (Camerer *et al.*, 2016; Open Science Collaboration, 2015) and in many-to-one designs. However, as others have noted (Patil *et al.*, 2016; Simonsohn, 2015), significance agreement is challenging to interpret because it is a function not only of the nominal α-level (e.g. 0.05), but also of power in both the original and the replication study. Thus, the expected probability of significance agreement may be quite low (Andrews and Kasy, 2019; Patil *et al.*, 2016), though it can be simulated (Patil *et al.*, 2016) or derived (on-line supplement) for a given original and replication study and then compared with the observed probability. In our running example regarding moral credentialling, 24% of replications (five of 21) obtained results agreeing in statistical significance and direction of effect with the original, which appears much lower than the 62% than we would expect theoretically (based on the original estimate and its standard error, as well as the standard error of each replication). Assessing significance agreement across replication studies is a direct analogue to the outdated practice of synthesizing evidence across studies in a systematic review by counting the number of studies obtaining 'statistically significant' results. The latter 'vote counting' method has extremely poor power and, surprisingly, in some settings the power can decrease as the number of synthesized studies increases (Hedges and Olkin, 1980).

A variety of more interpretable metrics have been developed for one-to-one replications, and some have also been reported in many-to-one designs. Patil *et al.* (2016) proposed to use the original study to construct a prediction interval representing a plausible range for the effect estimate in the replication study, assuming that the replication and the original study are generated from the same distribution (i.e. they are statistically 'consistent'). If indeed the two studies are generated from the same distribution, then regardless of power in either study there is, by construction, a 95% probability that the replication effect estimate will fall inside the prediction interval. Simonsohn (2015) proposed a hypothesis test of the replication estimate *versus* a non-zero null value chosen as the smallest effect size that the original study would have had an estimated 33% power to detect; this approach can help to assess whether the original study was adequately powered to detect the effect studied but does not directly assess consistency between the original study and the replications nor strength of evidence in the replications themselves. Andrews and Kasy (2019) developed a sophisticated, general statistical model for median-unbiased effect size estimation in one-to-one replication designs such as Open Science Collaboration (2015). Several researchers (e.g. Etz and Vandekerckhove (2016) and Verhagen and Wagenmakers (2014)) have recommended using Bayes factors to quantify evidence for and against the null hypothesis.

In a many-to-one design, some of these metrics can be applied individually to each replication study or to the pooled estimate. The former analysis can be informative but does not aggregate evidence and statistical power across all replications. The latter analysis is subject to the same limitations as the pooled estimate itself, namely that it summarizes a potentially heterogeneous distribution of replication effects by only its mean. In fact, analyses that fail to account for heterogeneity can underestimate (or potentially overestimate) consistency when there is in fact heterogeneity, leading to conclusions that can be unduly unfavourable (or potentially unduly favourable) to the original study, as proven in the on-line supplement.

4. Proposed new analyses

As discussed above, few statistical methods have been developed specifically for many-to-one designs, and those that were developed for other replication designs have limitations when applied to many-to-one designs, particularly in the presence of heterogeneous effects. We therefore propose new metrics to address central objectives of replication research while accounting for

all relevant sources of statistical uncertainty, namely statistical error in the original, statistical error in the replication, and heterogeneity. All proposed analyses can be conducted by using the R packages `Replicate` and `MetaUtility`, which are described in the on-line supplement.

4.1. Consistency of original with replications (P_{orig})

Our first proposed metric assesses statistical consistency. Rather than assuming that the replications and the original study measure exactly the same underlying effect size—an assumption that is implicit in most metrics for single replications—we instead assume that they measure potentially heterogeneous, normally distributed effects. We shall then say that the original study is 'consistent' with the replications if it is generated from the same underlying distribution as the replications, i.e. its population effect size comes from the same distribution as those of the replications. This measure directly accounts for heterogeneity because it compares the original study with the heterogeneous *distribution* of population effects underlying the replications, rather than to each replication point estimate individually. Then, we define the first proposed metric, called P_{orig}, as the probability that, if indeed the original is consistent with the replications in this sense, its estimate would be as extreme or more extreme than it actually was. A small value of P_{orig} would indicate strong evidence that the original study is inconsistent with the replications, whereas a large value would suggest relatively good consistency. In practice, if the original study is highly inconsistent with the replications, even accounting for heterogeneity, then we might consider it an anomaly. Future meta-analyses of the published literature might then present analyses both including and excluding such potentially anomalous studies. Additionally, others describe meta-analytically pooling results of an original study with those of a replication (Anderson and Maxwell, 2016; Open Science Collaboration, 2015); high inconsistency would suggest interpreting such analyses with greater caution.

To estimate P_{orig}, we first define $\hat{\theta}_{\text{orig}}$ and $\widehat{\text{SE}}_{\text{orig}}$ as the point estimate and standard error estimate in the original study, $\hat{\mu}$ and $\widehat{\text{SE}}_{\hat{\mu}}$ as estimates of the average population effect size in the replications and its standard error respectively and $\hat{\tau}^2$ as an estimate of the variance of the population effect sizes across replications. The effect sizes should be estimated on a scale for which the normality assumption is plausible (e.g. Fisher's z-scale), though simulation results suggest robustness to violations of this assumption (Section 6). In practice, $\hat{\mu}$ and $\hat{\tau}^2$ are most commonly estimated by using the pooled estimate and heterogeneity estimate from a random-effects meta-analysis of the replication sites' estimates. Alternatively, they could be estimated by fitting a mixed model to the individual observations themselves (which is also known as an 'individual patient data meta-analysis' (Stewart *et al.*, 2012)); both approaches are discussed further in the on-line supplement. In the main text, for simplicity, we illustrate the common meta-analytic approach, but all analyses can be conducted by using any unbiased estimates $\hat{\mu}$ and $\hat{\tau}^2$ arising from a model with the given distributional assumptions (supplement).

Then, if the original study is in fact consistent with the replications, the probability that its estimate would be as extreme as we observe it to be is approximately

$$P_{\text{orig}} = 2\left[1 - \Phi\left\{\frac{|\hat{\theta}_{\text{orig}} - \hat{\mu}|}{\sqrt{(\hat{\tau}^2 + \widehat{\text{SE}}^2_{\text{orig}} + \widehat{\text{SE}}^2_{\hat{\mu}})}}\right\}\right]. \tag{4.1}$$

For example, we used the R package `metafor` (Viechtbauer, 2010) to fit a random-effects meta-analysis to the site level point estimates of Ebersole *et al.* (2016) on Fisher's z-scale via restricted maximum likelihood with inverse variance weights. We thus estimated on the z-scale $\hat{\mu} = 0.07$, $\widehat{\text{SE}}_{\hat{\mu}} = 0.02$ and $\hat{\tau}^2 = 0.003$. We computed $\hat{\theta}_{\text{orig}} = 0.21$ and $\widehat{\text{SE}}_{\text{orig}} = 0.06$ for the original

study by converting the reported η^2-scale to Fisher's z (Lakens, 2013). Then, using equation (4.1), we estimated that, if the population effect in the original study indeed arose from the same distribution from which the replications were drawn, there would be a 10% chance that the original effect estimate would be as extreme as or more extreme than the observed 0.21. We can interpret this fairly low, but non-negligible, probability as being only weakly suggestive of inconsistency. On the basis of a visual assessment and the Shapiro–Wilk test (Shapiro and Wilk, 1965), the normality assumption did not appear to be violated, as described below in Section 5.

In contrast, previously discussed metrics indicating a low proportion of replications agreeing in statistical significance (24% *versus* 62% expected) and falling within the original prediction interval (76% *versus* 95% expected) might appear to suggest inconsistency more strongly. These relatively more pessimistic conclusions (compared with the conclusions that we might draw from P_{orig}) reflect these previous metrics' failure to account for heterogeneity in the effects across replications. To illustrate quantitatively, we can recompute P_{orig}, but this time setting $\hat{\tau}^2 = 0$ to assume no heterogeneity in the effects across replications. We then obtain a probability of only 4%. This is considerably lower than the 10% figure that is obtained by properly accounting for heterogeneity: a heterogeneous distribution of effects in the replications typically allows a higher chance that any given study would measure a very large or very small point estimate (as shown mathematically in the on-line supplement).

4.2. Proportion of population effects agreeing in direction with the original ($\hat{P}_{>0}$)

To address a second central objective of replication—reassessing the strength of evidence for the scientific effect of interest—we propose a metric ($\hat{P}_{>0}$) to supplement the usual pooled effect estimate and its CI. Unlike these existing metrics, which characterize only the mean of the distribution of population effects in the replications, $\hat{P}_{>0}$ characterizes both the mean and the heterogeneity of this distribution, and it addresses effect size rather than statistical significance. Specifically, $\hat{P}_{>0}$ represents the proportion of population effects, among the potentially heterogeneous population from which the replications are a sample, that agree in direction with the original. That is, any non-zero population effect agreeing in direction can be interpreted as a 'real' effect supporting the original study's theory (albeit potentially of a smaller effect size). This metric provides additional information beyond that provided by P_{orig}. That is, P_{orig} helps to assess whether the replications were 'successful' in the sense that their results are similar to those of the original study. In contrast, $\hat{P}_{>0}$ helps to assess whether the replications were successful in the sense of providing evidence for the effect under investigation, regardless of the results of the original study. We return to this difference in interpretation in analysing the applied example below as well as in the conclusion in Section 8.

To estimate $\hat{P}_{>0}$, it is not sufficient simply to compute the observed proportion of replication estimates agreeing in direction with the original; such an approach would fail to account for statistical error in the replication estimates. That is, the challenge is to use the distribution of the replication estimates (which has variability due to both heterogeneity and statistical error) to estimate the distribution of population effects (which has variability due only to heterogeneity). As in the context of meta-analysis rather than replication (Mathur and VanderWeele, 2020), the sample proportion can instead be estimated by using 'calibrated' estimates that have been appropriately shrunk to correct the point estimates' overdispersion due to statistical error (Wang and Lee, 2019). Let $\hat{\theta}_{\text{rep},i}$ and $\widehat{\text{SE}}_{\text{rep},i}$ respectively denote the point estimate and standard error estimate in the ith replication study, and, as before, let $\hat{\mu}$ and $\hat{\tau}^2$ represent estimates of the mean and variance of the population effects in the replications. Then the calibrated estimate of the population effect in the ith replication study is (Wang and Lee (2019), equation (4.1))

$$\tilde{\theta}_{\text{rep},i} = \hat{\mu} + (\hat{\theta}_{\text{rep},i} - \hat{\mu}) \sqrt{\left(\frac{\hat{\tau}^2}{\hat{\tau}^2 + \widehat{\text{SE}}^2_{\text{rep},i}} \right)}.$$

Intuitively, the calibrated estimate $\widetilde{\theta}_{\text{rep},i}$ shrinks the point estimate $\hat{\theta}_{\text{rep},i}$ towards the estimated mean $\hat{\mu}$ with a degree of shrinkage that is inversely proportional to the study's precision: relatively imprecise estimates $\hat{\theta}_{\text{rep},i}$ (i.e. those with large $\widehat{\text{SE}}_{\text{rep},i}$) receive strong shrinkage towards $\hat{\mu}$, whereas relatively precise estimates receive less shrinkage and remain closer to their original values. (Readers who are familiar with empirical Bayes estimation will note that the coefficient $\sqrt{\{\hat{\tau}^2/(\hat{\tau}^2 + \widehat{\text{SE}}^2_{\text{rep},i})\}}$ imposes less shrinkage than the coefficient $\hat{\tau}^2/(\hat{\tau}^2 + \widehat{\text{SE}}^2_{\text{rep},i})$ that would be used in the empirical Bayes estimate. By construction, the latter estimates minimize componentwise losses but produce calibrated estimates that are underdispersed compared with the population effects (Louis, 1984). The present coefficient $\sqrt{\{\hat{\tau}^2/(\hat{\tau}^2 + \widehat{\text{SE}}^2_{\text{rep},i})\}}$ minimizes the distance between the empirical cumulative distribution functions of the calibrated estimates and of the population effects (Louis, 1984), which is the relevant loss function for estimation of $\hat{P}_{>q}$ and $\hat{P}_{<q^*}$.)

One can then estimate $\hat{P}_{>0}$ as the sample proportion of calibrated estimates above 0, i.e. letting k denote the number of replication studies,

$$\hat{P}_{>0} = \frac{1}{k} \sum_{i=1}^{k} \mathbb{1}(\tilde{\theta}_{\text{rep},i} > 0)$$

(Mathur and VanderWeele, 2020).

For inference, one can bootstrap pairs of $(\hat{\theta}_{\text{rep},i}, \widehat{\text{SE}}_{\text{rep},i})$ by drawing with replacement from the original sample and estimating in turn $\hat{\mu}$ and $\hat{\tau}^2$, $\tilde{\theta}_{\text{rep},i}$ for each replication, and finally $\hat{P}_{>0}$ (Mathur and VanderWeele, 2020). A bias-corrected and accelerated CI (Carpenter and Bithell, 2000; Efron, 1987) can then be constructed from the bootstrapped values of $\hat{P}_{>q}$.

4.3. Proportion of meaningfully strong population effects ($\hat{P}_{>q}$ and $\hat{P}_{<q^*}$)

The aforementioned $\hat{P}_{>0}$ treats all effects that agree in direction with the original estimate, even those that are very close to the null, as evidence in favour of the scientific effect under investigation. This is generous towards the original study and therefore might serve as a useful default analysis. Alternatively, as a more stringent measure of evidence strength, it can also be useful to consider a generalized metric ($\hat{P}_{>q}$) representing the proportion of effects that are stronger than a non-null threshold, q. This approach is similar to equivalence testing and minimal effects testing, which compare a point estimate with null values other than 0 (Lakens *et al.*, 2018). An extensive interdisciplinary literature has provided recommendations on how to choose thresholds for meaningfully strong effect sizes, which we summarize briefly in the online supplement. For example, suppose that, through comparison with well-established effects on similar dependent variables (supplement), we select a threshold at an effect size of Cohen's $d = 0.20$ or, equivalently, an approximate correlation of $r = 0.10$ (Cohen, 1977). If $\hat{P}_{>q}$ is large (e.g. 85%), this suggests that, when drawing from the population distribution of effect sizes underlying the replications, a high proportion of population effects are sufficiently large to warrant scientific interest (e.g. larger than Cohen's $d = 0.20$). We might therefore conclude that the replications provide strong evidence that the scientific effect of interest is meaningfully strong in many settings. In contrast, if $\hat{P}_{>q}$ is small, we might instead conclude that the replications fail to support meaningfully strong effects in most contexts.

Conversely, it can also be useful to consider effects in the direction opposite to the original estimate by using a second threshold-based metric, $\hat{P}_{<q^*}$. That is, one could select a second

threshold representing a meaningfully strong effect size in the opposite direction (e.g. Cohen's $d = -0.20$) and estimating the proportion of population effects *below* this threshold. If the pooled estimate is fairly close to the null or if heterogeneity is substantial, this probability may be non-neglible, suggesting that the experimental manipulation may (perhaps unexpectedly) induce meaningful effects in the opposite direction in some replication settings. Such a finding may help to stimulate hypotheses regarding important moderators or boundary conditions on the effect of interest. Additionally, effects in the opposite direction from theoretical predictions may actively support competing theories. Indeed, when evaluating competing theories, researchers sometimes deliberately design experimental manipulations that are expected to induce opposing effects under each candidate theory. Returning to moral credentialling, the theory under investigation predicts that credentialling opportunities would *increase* subsequent attitudes that are consistent with prejudice; however, other theories suggest that credentialling opportunities might sometimes *decrease* such attitudes by prompting self-consistency or by priming personal values that discourage prejudice (Monin and Miller, 2001). Using $\hat{P}_{<q^*}$ to characterize effects in the opposite direction explicitly (rather than simply allowing them to dilute the pooled estimate without additional consideration) may help to identify situations, possibly supported by alternative theories, in which such competing effects occur.

These threshold-based metrics are particularly informative when the pooled estimate in the replications is smaller than that of the original study, as is often the case (e.g. Ebersole *et al.* (2016)). The proportion of population effects above a threshold ($\hat{P}_{>q}$) may then help to identify whether

(a) the population effects are closely clustered around a small average effect size, providing little evidence for meaningfully strong effects *versus*
(b) the population effects are quite variable around a small average effect size, such that there is in fact compelling evidence that meaningfully strong effects occur in some settings (and thus suggesting the importance of examining possible moderators).

For example, suppose that the original study estimates an effect size of $d = 0.85$, but the replications estimate a much smaller pooled effect size. Exclusive focus on the existing metrics may then mislead us into considering the replication effort to have succeeded completely (if the pooled point estimate is also 'statistically significant') or to have failed completely (if the pooled point estimate is not 'statistically significant'). However, if we additionally choose a threshold of scientific importance at, for example, $d = 0.20$ and estimate a reasonably high percentage (e.g. 75%) of population effects exceeding this threshold, then we might instead consider the replications to provide moderately strong evidence for meaningful effect sizes in some replication settings, warranting an assessment of possible moderators. In contrast, if we instead find that only, for example, 8% of population effects exceed $d = 0.20$, then we might instead conclude that the replications provide little evidence to support meaningfully strong effect sizes (even if the pooled point estimate is 'statistically significant'). Similarly to $\hat{P}_{>0}$, these metrics can be estimated by using the appropriate sample proportions of calibrated estimates, i.e. letting q be a chosen threshold defining a meaningfully strong effect size,

$$\hat{P}_{>q} = \frac{1}{k} \sum_{i=1}^{k} \mathbb{1}(\tilde{\theta}_{\text{rep},i} > q)$$

and

$$\hat{P}_{<q^*} = \frac{1}{k} \sum_{i=1}^{k} \mathbb{1}(\tilde{\theta}_{\text{rep},i} < q^*)$$

(Mathur and VanderWeele, 2020). If the point estimates have been transformed for analysis (e.g. from Pearson's r- to Fisher's z-scale), the calibrated estimates will also be on the transformed scale, and the threshold q should therefore be chosen on the transformed scale as well. CIs can again be obtained via bias-corrected and accelerated bootstrapping.

4.4. Proportion of replication effects supporting moral credentialling

In the moral credentialling example, the original study had a point estimate of 0.21 (95% CI 0.09, 0.32) on the Pearson correlation scale, whereas our meta-analysis of the replications of Ebersole *et al.* (2016) had a pooled estimate of 0.07 (95% CI 0.02, 0.12) with estimated heterogeneity $\hat{\tau}^2 = 0.003$. As discussed previously, the 'statistically significant' result in the replications might appear to suggest that the replication effort was successful. But does the small pooled estimate in the replications, despite its 'statistical significance', correspond to a high proportion of replication effects supporting credentialling theory? First, we can use $\hat{P}_{>0}$ to estimate the proportion of population effects above 0 (91% with 95% CI 57%, 100%). Alternatively, to choose a threshold representing a minimum effect size that is meaningfully strong, we might consider the effect sizes of well-established interventions on the same outcome measure, namely prejudice (see the on-line supplement for more details on choosing a threshold). For example, a meta-analysis of the enormous literature on intergroup contact and prejudice obtained a point estimate of $r = -0.21$ among all study designs and $r = -0.33$ among experimental studies (Pettigrew and Tropp, 2008). We might treat experimental intergroup contact interventions as a 'gold standard' representing the effect sizes on prejudice that are achievable through purposefully designed interventions. In contrast, the proposed moral credentialling effect is not a designed intervention on prejudice but rather a specific, potentially more subtle, cognitive mechanism of prejudice. Thus, to select an effect size threshold for moral credentialling, we might somewhat reduce the magnitude of the gold standard interventions to, for example, $|r| = 0.10$, which corresponds to Fisher's $z \approx 0.10$ on the analysis scale.

Then we can estimate the sample proportion of calibrated estimates above 0.10 as

$$\hat{P}_{>q} = \frac{1}{k} \sum_{i=1}^{k} \mathbb{1}(\tilde{\theta}_{\text{rep},i} > 0.10).$$

This analysis suggests that a minority of effects (14% with 95% CI 0%, 43%) surpass $r = 0.10$. We can also estimate $\hat{P}_{<q^*}$, i.e. the proportion of meaningfully strong population effects in the direction opposite to Monin and Miller's (2001) original findings. We might choose a conservative second threshold at, for example, $r = 0$ ($z = 0$ on the analysis scale) and use the sample proportion of calibrated estimates below 0,

$$\frac{1}{k} \sum_{i=1}^{k} \mathbb{1}(\tilde{\theta}_{\text{rep},i} < 0),$$

to estimate that few effects (14% with 95% CI 0%, 43%) are negative. (For this particular example, the point estimates and CIs are the same for $\hat{P}_{>0.10}$ and $\hat{P}_{<0}$ because, in each case, exactly three (14%) of the calibrated estimates were more extreme than the chosen threshold, as can be seen in Fig. 1.)

Ultimately, although these replications produce a 'statistically significant' point estimate in the same direction as the original study's estimate, we might nevertheless caution that they provide little evidence for effect sizes of strength comparable with that of the original estimate across replication settings. In the distribution of population effects, there is a high proportion of non-zero effects in the direction of the original estimate, but most of these effects are considerably

smaller than the original estimate. Considering these results along with the previously discussed consistency metric ($P_{\text{orig}} = 10\%$), we might say, overall, that the moral credentialling main effect 'replicated' in the sense that there is not compelling evidence for inconsistency between the original study and replications (once we account for heterogeneity), yet strength of evidence for meaningfully strong effect sizes of moral credentialling is considerably weaker than suggested by the original study. These complementary findings further illustrate the conceptual distinction between statistical consistency and strength of evidence for meaningfully strong effects of interest.

5. Reporting recommendations and statistical diagnostics

5.1. Calibrated forest plot

To supplement the statistical metrics proposed, we suggest creating a forest plot displaying the point estimates and 95% CIs for each replication study, for all replication studies combined and for the original study (e.g. Fig. 1). It may be informative additionally to indicate the calibrated estimate $\tilde{\theta}_{\text{rep},i}$ in each replication study (e.g. the crosses in Fig. 1) and to order the replication studies by their calibrated estimates. The calibrated estimate in each replication can serve as an estimate of that replication study's population effect size on accounting for the study's precision along with information that is provided by the other replications. Taken together, the calibrated estimates can provide a heuristic sense of the distribution and variability of the population effect sizes themselves, rather than of the potentially noisy point estimates. Furthermore, if there are sites with apparently extreme point estimates or calibrated estimates, one might conduct sensitivity analyses in which the metrics proposed, as well as the diagnostics that are described below, are recalculated after excluding these potential outliers.

5.2. Diagnostics for distributional assumptions

Our proposed metrics assume that the point estimates are normally distributed around their corresponding population effects, which is reasonable when the replication studies have reasonably large sample sizes and when their estimates are transformed to an appropriate scale for analysis (Sutton *et al.*, 2000). Further, the metric P_{orig} assumes that the population effect sizes themselves in the replication studies are normally distributed, though the metrics $\hat{P}_{>0}$, $\hat{P}_{>q}$ and $\hat{P}_{<q^*}$ do not use this assumption. In most many-to-one designs, which use mixed modelling or parametric random-effects meta-analysis to estimate the pooled effect, this assumption is already implicit. Nevertheless, we recommend assessing its plausibility as follows. If the normality assumption holds, then the standardized estimates in the replication studies, $(\hat{\theta}_{\text{rep},i} - \hat{\mu})/\sqrt{(\hat{\tau}^2 + \widehat{\text{SE}}^2_{\text{rep},i})}$, will be approximately normal (Hardy and Thompson, 1998). Thus, one could test for normality of the standardized estimates via the Shapiro–Francia or Shapiro–Wilk test, though power may be limited when the number of replication studies is small (Shapiro and Francia, 1972; Shapiro and Wilk, 1965).

Additionally, sensitivity analyses could be conducted by meta-analysing the replication study estimates under distribution-free approaches (Fisher and Tipton, 2015; Hedges *et al.*, 2010) or under more flexible distributional assumptions (see Higgins *et al.* (2009) for a review). If the robustly estimated $\hat{\mu}$ and $\hat{\tau}^2$ differ substantially from those obtained through parametric meta-analysis, this suggests that the normality assumption may be violated. Because assessment of normality and accurate heterogeneity estimation are challenging with small sample sizes (Hardy and Thompson, 1998; Veroniki *et al.*, 2016), these proposed replication metrics should generally be applied only when there are at least 10 replication studies (unless there is no heterogeneity). To the best of our knowledge, this was true in each of the 79 completed many-to-one designs that were discussed in Section 1. An exception to this rule of thumb is when there is no heterogeneity,

as discussed in the following section. Additionally, simulation results suggested that P_{orig} is in fact quite robust to violations of the normality assumption (Section 6).

5.3. Diagnostics for uncertainty in P_{orig}

Additionally, estimation uncertainty in $\hat{\tau}^2$ may be considerable when the number of replications is small (Veroniki *et al.*, 2016); it is then possible that many values of τ^2 are plausible given the data, resulting in considerable uncertainty about P_{orig} itself. Three types of diagnostic plots may help to identify such situations (see Rubin (1981)). First, one could plot the marginal likelihood of τ^2 for a range of hypothetical values τ^2_*, including values both smaller than and larger than the actual point estimate $\hat{\tau}^2$. In practice, to display the likelihood on an interpretable scale, we suggest plotting the marginal likelihood ratio of the hypothetical values τ^2_* against the actual estimate $\hat{\tau}^2$ (Figs 2(a) and 2(b)), e.g. using the R function `Replicate::t2_lkl`. Ideally, the peak of the likelihood ratio curve would span a narrow range of values around the actual estimate $\hat{\tau}^2$, declining sharply on either side of $\hat{\tau}^2$; this would indicate relatively good certainty about the true value of τ^2. If instead the likelihood ratio spans a wide range of values, remaining close to 1 even for values τ^2_* that are several fold smaller or larger than the actual estimate $\hat{\tau}^2$, this suggests that there may be considerable uncertainty in P_{orig} itself.

For example, Figs 2(a) and 2(b) display this likelihood ratio for all 21 moral credentialling replications and for only the first three replications respectively. Because the 21 replications provide more information about the amount of heterogeneity than do only the first three, the likelihood curve is much more peaked in Fig. 2(a) than in Fig. 2(b) and, in Fig. 2(b), a wide range of heterogeneity values appear plausible given the very limited data. A plot similar to Fig. 2(b) would suggest that the replication data might not be sufficiently informative to assess consistency via P_{orig}. As second and third diagnostic plots, one could similarly plot a range of values τ^2_* against the ratio $\hat{\mu}/\tau^2_*$, which plays a central role in the calculation of P_{orig}, and also against P_{orig} itself (Figs 2(c) and 2(d)). These plots could help to indicate how much P_{orig} would change if a different heterogeneity estimate had been obtained. If, for example, there are hypothetical values τ^2_* for which the likelihood ratio is high (i.e. close to 1 based on plots such as Figs 2(a) and 2(b)) but for which $\hat{\mu}/\tau^2_*$ and P_{orig} differ substantially from their actual estimates, this again suggests that there may be much uncertainty associated with P_{orig} given the replication data.

5.4. Diagnostics for bias in $\hat{P}_{>q}$

As discussed in Section 6 below, the metric $\hat{P}_{>q}$ and the related proportion metrics can potentially be biased if, for example, there is low heterogeneity, the true proportion (i.e. $E[\hat{P}_{>q}]$) is close to 0 or 1, or if there are limited replication data. A useful diagnostic for potential bias is the mean of the bootstrapped estimates that are used to construct the CI. The difference between the mean of the bootstrap estimates and the sample estimate $\hat{P}_{>q}$ is an estimate of the bias of $\hat{P}_{>q}$ itself (Davison and Hinkley (1997), section 2.2.1), so caution is warranted if the mean of the bootstrap estimates differs considerably from $\hat{P}_{>q}$. The R function `MetaUtility::prop_stronger` automatically returns the mean of the bootstrap estimates for this purpose.

6. Simulation study

6.1. Methods

We conducted a simulation study with two objectives. First, we assessed the type I error and power of P_{orig} when it is treated as a level $\alpha = 0.05$ test of the hypothesis that the population effect size in the original study is consistent with the replication studies, i.e. letting μ and τ^2 denote

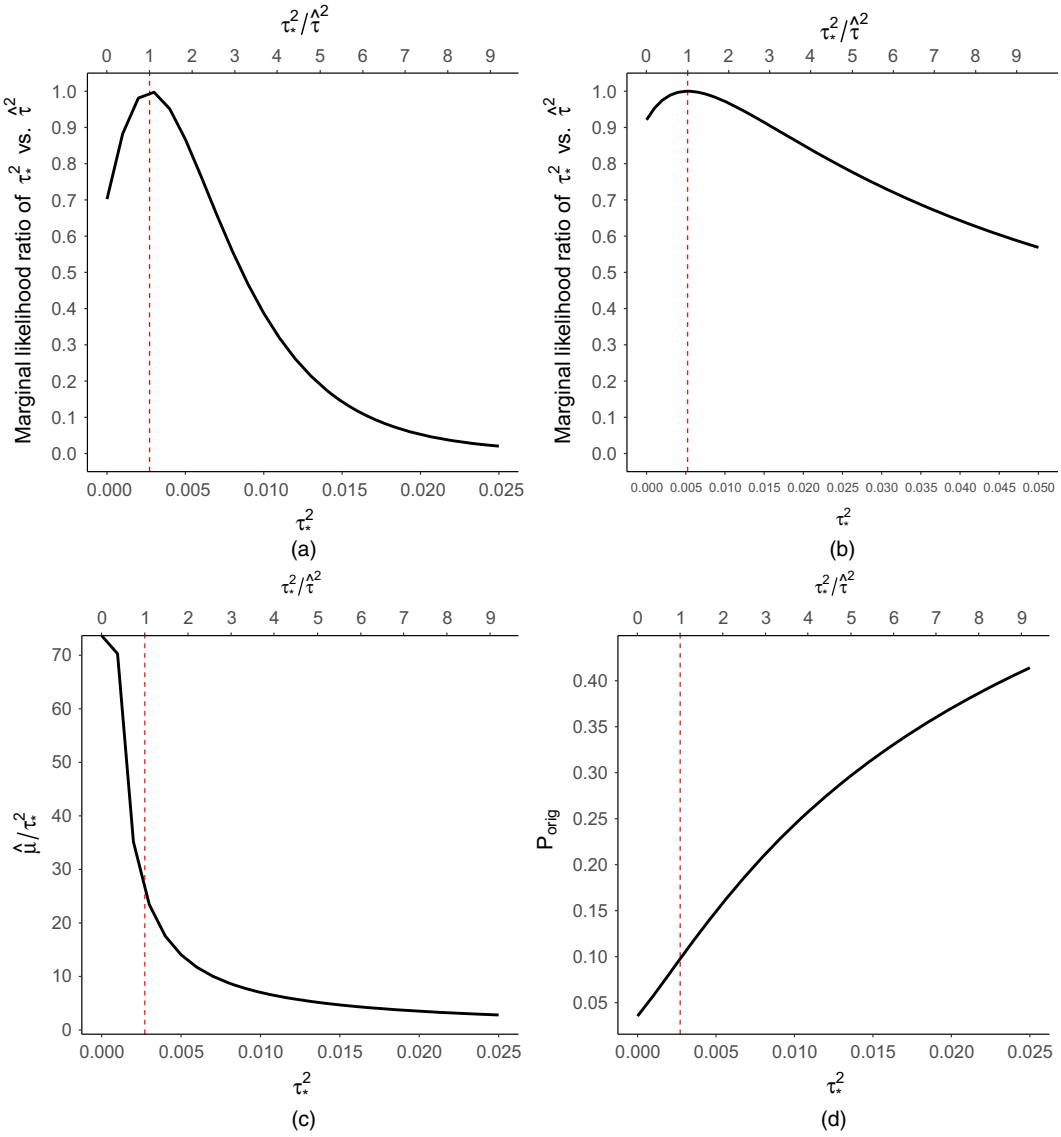

Fig. 2. Diagnostic plots (see Rubin (1981)) relating hypothetical heterogeneity values τ_*^2 to the marginal likelihood ratio of the hypothetical τ_*^2 *versus* (a), (b) the actual restricted maximum likelihood estimate $\hat{\tau}^2$, (c) to the ratio $\hat{\mu}/\tau_*^2$ and (d) to the resulting P_{orig}: (a) (c) and (d) were created by using the applied example data; (b) was created by using only the first three of the moral credentialling replications and serves as a comparison for (a); the upper *x*-axis is the ratio of the hypothetical value τ_*^2 to the actual estimate $\hat{\tau}^2$ (!, actual estimate $\hat{\tau}^2$)

the unknown true mean and variance of the distribution of effects underlying the replications, we tested $H_0 : \theta_{\text{orig}} \sim N(\mu, \tau^2)$ *versus* $H_A : \theta_{\text{orig}} \not\sim N(\mu, \tau^2)$, rejecting H_0 when $P_{\text{orig}} < 0.05$. (As described in Section 8, in practice, we recommend reporting P_{orig} as a continuous measure rather than relative to a dichotomous α-threshold. We use the hypothesis testing framework for the simulation study merely as a means of benchmarking the metric's performance.) Second, we assessed the bias, root-mean-square error RMSE and CI coverage of $\hat{P}_{>q}$.

To simulate data sets, we fixed the mean of the population effects to $\mu = 0.50$ on the mean difference scale while varying the number of replications, k, between 5 and 25, the heterogeneity $\tau^2 \in \{0.002, 0.01, 0.25\}$ and the sample size in each individual replication $N_{rep} \in \{50, 300\}$. Relevant to the estimation of P_{orig}, we varied the sample size in the original study $N_{orig} \in \{30, 50, 200\}$ and the difference Δ between the population effect size in the original study and the true mean in the replications ($\Delta \in \{0, 0.20, 0.50, 1.0\}$). Relevant to the estimation of $\hat{P}_{>q}$, we chose different thresholds q so that the expectation of $\hat{P}_{>q}$ varied in $\{0.05, 0.10, 0.50\}$. In each scenario, we calculated $\hat{P}_{>q}$ by using two recommended heterogeneity estimators (Veroniki *et al.*, 2016): the restricted maximum likelihood estimator (Raudenbush and Bryk, 1985) and the estimator of Paule and Mandel (1982).

For each of k replications, we generated a population effect, $\theta_{rep,i}$, on the raw mean difference scale from a normal distribution, a shifted exponential distribution, a scaled and shifted t-distribution with 3 degrees of freedom or a bimodal uniform mixture distribution. For all distributions, we chose the parameters to provide the desired mean of $\mu = 0.50$ and heterogeneity τ^2. Fig. S1 in the on-line supplement shows population effects simulated from each of the four distributions for each value of τ^2. We also generated a population effect for the original study, θ_{orig}, from a comparable distribution but with mean $\mu + \Delta$. Thus, the null hypothesis regarding P_{orig} held when $\Delta = 0$. For each replication and for the original study, we then simulated subject level data for a control group with mean 0 and for an intervention group with mean $\theta_{rep,i}$ or θ_{orig} respectively; each group was of size $N_{rep}/2$ or $N_{orig}/2$ respectively and had a standard deviation of 1. Thus, the true within-study standard errors of the estimated mean differences in each replication and in the original study were $\widehat{SE}_{rep,i} = \sqrt{(4/N_{rep})}$ and $\widehat{SE}_{orig} = \sqrt{(4/N_{orig})}$. We ran scenarios representing all 4320 possible combinations of the varying parameters, using 5000 bootstrap iterates to estimate the CI for $\hat{P}_{>q}$. We ran at least 1000 simulation iterates per scenario.

6.2. Results

Comprehensive simulation results for all 4320 scenarios are presented in Figs S2–S8 in the on-line supplement and can also be explored interactively via the website `https://mmathur.shinyapps.io/RRR_sims/` or downloaded as a data set from `https://osf.io/ufjg4/`. The two heterogeneity estimators performed almost identically, so we report here only the results by using the restricted maximum likelihood estimator. Across all scenarios with $k \geqslant 10$ (supplement), including both normal and non-normal distributions, the average type I error was 5%. The maximum type I error rate was 11%, which occurred in scenarios with high heterogeneity as well as few or low powered replications (e.g. Fig. S2). When heterogeneity was low to moderate ($\tau^2 = 0.001$ or $\tau^2 = 0.01$), the maximum type I error rate was 7% across all sample sizes and numbers of replications and was 6% when considering only normally distributed population effects. Alternatively, when there was a fairly large number ($k = 25$) of large replication studies ($N_{rep} = 300$), the maximum type I error rate was 7% across all values of heterogeneity and was 6% when considering only normally distributed population effects. These results suggest that, when $k \geqslant 10$, P_{orig} is quite robust to non-normal population effects, but that results should be interpreted cautiously if both

(a) the amount of replication data is limited (because of small sample sizes in the replication studies or a small number of replication studies) and
(b) heterogeneity appears to be high.

Table 1 summarizes the results for P_{orig} for scenarios with $k \geqslant 10$, since we do not recommend using these methods for $k < 10$ (unless there is no heterogeneity). For legibility, Table 1 displays results only for scenarios with $N_{orig} = 50$ and $N_{rep} = 300$. (Note that the means and maxima that

Table 1. Type I error ($\Delta = 0$ rows) and power ($\Delta > 0$ rows) of P_{orig} for scenarios with $k \geqslant 10$†

Distribution	τ^2	Δ	*Mean rejection rate*	*Maximum rejection rate*
Normal	0.002	0.00	0.04	0.05
Normal	0.002	0.20	0.09	0.09
Normal	0.002	0.50	0.34	0.34
Normal	0.002	1.00	0.89	0.90
Normal	0.010	0.00	0.05	0.05
Normal	0.010	0.20	0.09	0.09
Normal	0.010	0.50	0.33	0.33
Normal	0.010	1.00	0.86	0.87
Normal	0.250	0.00	0.06	0.07
Normal	0.250	0.20	0.07	0.08
Normal	0.250	0.50	0.13	0.15
Normal	0.250	1.00	0.37	0.39
Exponential	0.002	0.00	0.05	0.05
Exponential	0.002	0.20	0.09	0.10
Exponential	0.002	0.50	0.34	0.35
Exponential	0.002	1.00	0.89	0.90
Exponential	0.010	0.00	0.05	0.05
Exponential	0.010	0.20	0.09	0.09
Exponential	0.010	0.50	0.31	0.32
Exponential	0.010	1.00	0.85	0.86
Exponential	0.250	0.00	0.06	0.07
Exponential	0.250	0.20	0.08	0.09
Exponential	0.250	0.50	0.16	0.18
Exponential	0.250	1.00	0.37	0.38
t	0.002	0.00	0.05	0.05
t	0.002	0.20	0.09	0.10
t	0.002	0.50	0.33	0.34
t	0.002	1.00	0.88	0.89
t	0.010	0.00	0.05	0.06
t	0.010	0.20	0.09	0.09
t	0.010	0.50	0.28	0.30
t	0.010	1.00	0.80	0.81
t	0.250	0.00	0.07	0.08
t	0.250	0.20	0.08	0.09
t	0.250	0.50	0.11	0.12
t	0.250	1.00	0.23	0.24
Uniform mixture	0.002	0.00	0.04	0.04
Uniform mixture	0.002	0.20	0.09	0.09
Uniform mixture	0.002	0.50	0.34	0.35
Uniform mixture	0.002	1.00	0.89	0.89
Uniform mixture	0.010	0.00	0.05	0.05
Uniform mixture	0.010	0.20	0.09	0.09
Uniform mixture	0.010	0.50	0.32	0.34
Uniform mixture	0.010	1.00	0.86	0.87
Uniform mixture	0.250	0.00	0.03	0.04
Uniform mixture	0.250	0.20	0.05	0.06
Uniform mixture	0.250	0.50	0.14	0.15
Uniform mixture	0.250	1.00	0.40	0.42

†For legibility, results are shown only for scenarios with $N_{\text{orig}} = 50$ and $N_{\text{rep}} = 300$, and statistics aggregate over the other manipulated simulation parameters that are not listed in the column headings.

Table 2 Bias and RMSE of $\hat{P}_{>q}$ for scenarios with $k \geqslant 10$†

Distribution	τ^2	N_{rep}	*Mean bias*	*Mean RMSE*
Normal	0.002	50	0.15	0.29
Normal	0.002	300	0.07	0.19
Normal	0.010	50	0.06	0.20
Normal	0.010	300	0.01	0.12
Normal	0.250	50	−0.00	0.10
Normal	0.250	300	−0.00	0.09
Exponential	0.002	50	0.15	0.28
Exponential	0.002	300	0.08	0.18
Exponential	0.010	50	0.08	0.19
Exponential	0.010	300	0.04	0.13
Exponential	0.250	50	0.02	0.10
Exponential	0.250	300	0.00	0.09
t	0.002	50	0.12	0.26
t	0.002	300	0.04	0.17
t	0.010	50	0.04	0.17
t	0.010	300	0.01	0.11
t	0.250	50	−0.00	0.09
t	0.250	300	−0.00	0.09
Uniform mixture	0.002	50	0.16	0.30
Uniform mixture	0.002	300	0.08	0.21
Uniform mixture	0.010	50	0.08	0.22
Uniform mixture	0.010	300	0.03	0.14
Uniform mixture	0.250	50	0.01	0.11
Uniform mixture	0.250	300	0.01	0.09

†Statistics aggregate over the other manipulated simulation parameters that are not listed in the column headings.

are reported in the text consider each simulation scenario individually and include scenarios with all sample sizes, and so differ from the means and maxima that are listed in aggregated tables.)

For all scenarios with $k \geqslant 10$, the bias of $\hat{P}_{>q}$ was 0.05 on average and was at maximum 0.30. Bias occurred in scenarios with low heterogeneity, an extreme true proportion ($E[\hat{P}_{>q}]$) and either a small number of replication studies or small sample sizes in the replication studies. When there was a fairly large number ($k = 25$) of large replication studies ($N_{\text{rep}} = 300$), the average bias was 0.02, and the maximum bias was 0.12. Alternatively, when heterogeneity was high ($\tau^2 = 0.25$), the maximum bias was 0.07 across all sample sizes and true proportions. The bootstrapping-based inference appeared to compensate for occasional bias in $\hat{P}_{>q}$, as the minimum coverage across all scenarios with $k \geqslant 10$ was close to nominal (94%). Table 2 summarizes the results for $\hat{P}_{>q}$ for all scenarios with $k \geqslant 10$. These findings also point to the importance of designing multisite replications with a sufficient number of sites and sufficient sample sizes within each site, and we would encourage researchers designing multisite replication projects to use the interactive simulation results (`https://mmathur.shinyapps.io/RRR_sims/`) to provide some preliminary guidance.

7. Applications to other replication designs

We have primarily discussed our metrics in the context of many-to-one designs conducted under a shared replication protocol and in which population effects are heterogeneous. Here, we discuss other designs and settings to which the metrics proposed apply, potentially with modified interpretations.

7.1. Replications with no apparent heterogeneity

If a random-effects meta-analysis or similar individual subject level mixed model yields a negligible statistical estimate of heterogeneity ($\hat{\tau}^2 \approx 0$) with a CI including only small values, then $\hat{\mu}$ can be interpreted as an estimate of the single population effect size underlying all replication studies and will approximately coincide with the corresponding estimate from a fixed effects meta-analysis or linear regression (Rice *et al.*, 2018; Riley *et al.*, 2011). Because heterogeneity estimators can be highly variable when the number of studies is small (Veroniki *et al.*, 2016), it is important to report the CI for τ^2 when adopting this 'fixed effects' approach. Under this framework, P_{orig} can still be informative to assess consistency and is interpretable as the probability that the original study's point estimate would be at least as extreme as that observed if the original study had unbiasedly measured the same population effect as the replication studies. Without heterogeneity, P_{orig} does not require a normality assumption and can be reported with as few as one replication study, and it becomes a continuous counterpart to a prediction interval in which all replication data are analysed in aggregate, without regard to site (on-line supplement). The metrics $\hat{P}_{>0}$, $\hat{P}_{>q}$ and $\hat{P}_{<q^*}$ are no longer relevant because all population effects are estimated to be identical.

7.2. Single replications and one-to-one replications

Single replications or one-to-one replication projects preclude estimating heterogeneity for any given replication study, and existing analysis approaches implicitly assume no heterogeneity. If such an assumption is reasonable, then P_{orig} can be computed by setting $\hat{\tau}^2 = 0$, retaining the same interpretation as above; of course, such analyses would need to be interpreted with caution because the assumption of no heterogeneity is not testable in these designs.

7.3. 'Many Labs' designs

In designs in which multiple original studies are each replicated in many sites (Ebersole *et al.*, 2016; Klein *et al.*, 2014; Schweinsberg *et al.*, 2016), the metrics proposed permit direct comparison or aggregation of results across many-to-one replications of multiple original studies. For example, one could estimate the proposed metrics for each original study and report the average consistency, P_{orig}, as a global summary measure of replication success. The average $\hat{P}_{>0}$ could also be reported as a global summary of replication evidence strength across numerous scientific effects.

7.4. Conceptual replications

We have so far considered contexts in which all replications share a single protocol closely approximating that of the original study (sometimes called 'direct replications'). However, some researchers question using only direct replications in many-to-one designs, arguing that these designs assess replicability of a specific operationalization of a theory, rather than of the theory itself (Baumeister and Vohs, 2016). Others advocate supplementing direct replications with 'conceptual replications' that assess the same theory as the original study, but using a different operationalization (Crandall and Sherman (2016), Lynch *et al.* (2015) and Monin *et al.* (2014); see also dissent by Nosek *et al.* (2012) and Simons (2014)). For example, replication sites in a conceptual many-to-one design could implement different experimental protocols, each approved by the original researchers. Conceptual replications create heterogeneity by design, which exacerbates problems with the metrics that have been proposed before this paper (e.g. often leading to particularly unfavourable assessments of consistency and inadequately

characterizing the strength of evidence). In contrast, our proposed metrics could simply be applied without modification as they take into account heterogeneity across replications. They would retain their original interpretations, but $\hat{P}_{>0}$ could then additionally be interpreted as the probability that a new operationalization of the theory at stake would yield a population effect size in the same direction as the theoretical prediction. Such an interpretation holds only when the new operationalization under consideration can be treated as comparable with the range of protocols that are considered in the conceptual replications.

8. Discussion

We have proposed intuitively tractable metrics (implemented in the R packages `Replicate` and `MetaUtility`) for statistical consistency between the original study and replications and for the strength of evidence in many-to-one replication designs with potential heterogeneity. Such replication projects could report the new metric P_{orig} to convey consistency and could report the usual pooled estimate $\hat{\mu}$ and heterogeneity estimate $\hat{\tau}^2$, plus $\hat{P}_{>0}$ (and possibly also $\hat{P}_{>q}$ and $\hat{P}_{<q^*}$) to reassess the strength of evidence for the scientific effect of interest. The metrics proposed account for all relevant sources of statistical uncertainty and can therefore yield different conclusions from existing metrics when the replications are heterogeneous. These metrics can also help to identify situations in which there is good statistical consistency, but weak strength of evidence for meaningfully strong effects (and vice versa). For example, a set of replications estimating a small average effect size might be statistically consistent with a low powered original study that estimated a large effect size yet may provide little evidence that the effects of interest are of meaningfully strong size. In this case, P_{orig} would be fairly large, indicating good consistency, but $\hat{P}_{>q}$ would be small, indicating a low proportion of meaningfully strong effect sizes. Conversely, a set of replications estimating a moderate effect size may appear statistically inconsistent with an original study estimating a large effect size but may nevertheless provide strong evidence for meaningfully strong effect sizes.

The analyses proposed have limitations. We have assumed that the replications yield unbiased estimates, which is often reasonable when the replications are preregistered and conducted by third-party investigators. In contrast, other forms of replications, such as multiple experiments reported in a published, non-registered paper, may be subject to the same biases as seen in the published original research (Francis, 2012). As discussed, the metric P_{orig} assumes that the population effects are normally distributed; this assumption is already often used in pooled effect estimation and is often testable in practice, and simulation results indicated that P_{orig} is in fact quite robust to non-normal distributions. The metric P_{orig} also relies on accurate statistical estimation of both the pooled effect size and its variance. When estimating these parameters via random-effects meta-analysis, there are many possible choices of heterogeneity estimator, and it is important to choose one that is known to perform well for the effect measure of choice, particularly when the number of replication studies is relatively small (Veroniki *et al.*, 2016). P_{orig} should be interpreted cautiously if both

(a) the amount of replication data is limited and
(b) heterogeneity appears high,

as these situations can lead to poor estimation of $\hat{\tau}^2$. We have also suggested diagnostic plots to help to identify problems in estimating $\hat{\tau}^2$ that could propagate to P_{orig} (Section 5). Additionally, we do not recommend using P_{orig} to conduct a dichotomous 'hypothesis test' of consistency (by assessing whether $P_{\text{orig}} < 0.05$) between the original study and the replications; rather, P_{orig} is a continuous measure and is more informative when reported as such.

Simulation results indicated that CI coverage for $\hat{P}_{>q}$ and the related proportion metrics remained near the nominal 95% in all scenarios, though point estimates may be biased when there are few replications or small sample sizes in the replications. Therefore, we recommend reporting CIs along with the proportion metrics and using simple bootstrapping-based diagnostics (Section 5) to identify potential bias. When there are little replication data, estimating the amount of heterogeneity can be inherently imprecise (Veroniki *et al.*, 2016). This uncertainty propagates to the CIs for the proportion metrics and, when there are little replication data, may result in confidence intervals that span most or all of the possible range [0, 1]. Reporting CIs in these settings may nevertheless be informative: a very wide CI may instill appropriate circumspection about what can be learned from the replications, even if $\hat{\mu}$ itself may have a narrow CI. Additionally, the bootstrapped CI may sometimes fail to converge when there are few replication studies or if the threshold q is very far from $\hat{\mu}$; in these cases, one can try choosing a less extreme threshold.

In summary, the newly proposed metrics assess consistency of the original and replication studies and also assess evidence for meaningfully strong effects while accounting for heterogeneity across the population effects. Such heterogeneity is fairly common in practice and can arise due to differing subject demographics or protocol variations. If reported in many-to-one replication projects, the metrics proposed could help to address directly and intuitively the central objectives of replication research. These metrics are mathematically simple but are nevertheless, we believe, a useful supplement to current reporting practices to help to ground speculation about 'replication success' quantitatively.

9. Reproducibility

All data, materials and code required to reproduce the applied analyses and simulation study are publicly available, along with all code for the R packages `MetaUtility` and `Replicate` (`https://osf.io/ufjg4/`). The simulation data can also be browsed interactively (`https://mmathur.shinyapps.io/RRR_sims/`).

Acknowledgements

Maya Mathur conceptualized the research, conducted the data analysis and simulations, wrote the R packages and led writing. Tyler VanderWeele provided key input into the development of these methods and into writing. Both authors approved the final manuscript and declare that they have no conflicts of interest with respect to the authorship or publication of this paper.

Maya Mathur was supported by National Defense Science and Engineering graduate fellowship 32 CFR 168a and Franklin E. Fetzer Memorial Trust grant R2020-16. Tyler VanderWeele was supported by National Institutes of Health grants ES017876 and CA222147. The funders had no role in the design, conduct or reporting of this research.

References

Alogna, V., Attaya, M. K., Aucoin, P., Bahník, Š., Birch, S., Birt, A. R., Bornstein, B. H., Bouwmeester, S., Brandimonte, M. A., Brown, C. and Buswell, K. (2014) Registered replication report: Schooler and Engstler-Schooler (1990). *Perspect. Psychol. Sci.,* **9**, 556–578.

Anderson, S. F. and Maxwell, S. E. (2016) There's more than one way to conduct a replication study: beyond statistical significance. *Psychol. Meth.*, **21**, no. 1, 1.

Andrews, I. and Kasy, M. (2019) Identification of and correction for publication bias. *Am. Econ. Rev.*, **109**, 2766–2794.

Association for Psychological Science (2018) Ongoing replication projects. Association for Psychological Science, Washington DC. (Available from `https://www.psychologicalscience.org/publications/replication/ongoing-projects`.)

Baumeister, R. F. and Vohs, K. D. (2016) Misguided effort with elusive implications. *Perspect. Psychol. Sci.*, **11**, 574–575.

Bouwmeester, S., Verkoeijen, P. P., Aczel, B., Barbosa, F., Bégue, L., Brañas-Garza, P., Chmura, T. G., Cornelissen, G., Dessing, F. S., Espan, A. M. and Evans, A. M. (2017) Registered replication report: Rand, Greene, and Nowak (2012). *Perspect. Psychol. Sci.*, **12**, 527–542.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T. and Helkensten, E. (2016) Evaluating replicability of laboratory experiments in economics. *Science*, **351**, 1433–1436.

Carpenter, J. and Bithell, J. (2000) Bootstrap confidence intervals: when, which, what?: A practical guide for medical statisticians. *Statist. Med.*, **19**, 1141–1164.

Cheung, I., Campbell, L., LeBel, E. P., Ackerman, R., Aykutoğlu, B., Bahník, Š., Bowen, J. D., Bredow, C. A., Bromberg, C., Caprariello, P. A. and Carcedo, R. J. (2016) Registered replication report: Study 1 from Finkel, Rusbult, Kumashiro, & Hannon (2002). *Perspect. Psychol. Sci.*, **11**, 750–764.

Cohen, J. (1977) *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press.

Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., Beebe, J., Berniñas, R., Boudesseul, J., Colombo, M. and Cushma, F. (2018) Estimating the reproducibility of experimental philosophy. *Rev. Philos. Psychol.*, **14**, 1–36.

Crandall, C. S. and Sherman, J. W. (2016) On the scientific superiority of conceptual replications for scientific progress. *J. Exptl Socl Psychol.*, **66**, 93–99.

Davison, A. C. and Hinkley, D. V. (1997) *Bootstrap Methods and Their Application*, vol. 1. New York: Cambridge University Press.

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B., Boucher, L. and Brown, E. R. (2016) Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *J. Exptl Socl Psychol.*, **67**, 68–82.

Ebersole, C. R., Nosek, B. A., Kidwell, M. C., Buttrick, N., Baranski, E., Chartier, C. R., Mathur, M., Campbell, L., Izerman, H. and Lazarevic, L. (2018) Many Labs 5: Can conducting formal peer review in advance improve reproducibility? (Available from `https://osf.io/7a6rd/`.)

Eerland, A., Sherrill, A. M., Magliano, J. P., Zwaan, R. A., Arnal, J., Aucoin, P., Berger, S. A., Birt, A. R., Capezza, N., Carlucci, M. and Crocker, C. (2016) Registered replication report: Hart & Albarracín (2011). *Perspect. Psychol. Sci.*, **11**, 158–171.

Efron, B. (1987) Better bootstrap confidence intervals. *J. Am. Statist. Ass.*, **82**, 171–185.

Etz, A. and Vandekerckhove, J. (2016) A Bayesian perspective on the Reproducibility Project: Psychology. *PLOS One*, **11**, no. 2, article e0149794.

Fisher, Z. and Tipton, E. (2015) Robumeta: an R-package for robust variance estimation in meta-analysis. *Preprint arXiv:1503.02220*.

Francis, E. (2012) The psychology of replication and replication in psychology. *Perspect. Psychol. Sci.*, **7**, 585–594.

Gadbury, G. L. and Iyer, H. K. (2000) Unit-treatment interaction and its practical consequences. *Biometrics*, **56**, 882–885.

Gadbury, G. L., Iyer, H. K. and Allison, D. B. (2001) Evaluating subject-treatment interaction when comparing two treatments. *J. Biopharm. Statist.*, **11**, 313–333.

Hagger, M. S., Chatzisarantis, N. L., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., Brand, R., Brandt, M. J., Brewer, G., Bruyneel, S. and Calvillo, D. P. (2016) A multilab preregistered replication of the ego-depletion effect. *Perspect. Psychol. Sci.*, **11**, 546–573.

Hardy, R. J. and Thompson, S. G. (1998) Detecting and describing heterogeneity in meta-analysis. *Statist. Med.*, **17**, 841–856.

Heckman, J. J., Smith, J. and Clements, N. (1997) Making the most out of programme evaluations and social experiments: accounting for heterogeneity in programme impacts. *Rev. Econ. Stud.*, **64**, 487–535.

Hedges, L. V. and Olkin, I. (1980) Vote-counting methods in research synthesis. *Psychol. Bull.*, **88**, no. 2, 359.

Hedges, L. V., Tipton, E. and Johnson, M. C. (2010) Robust variance estimation in meta-regression with dependent effect size estimates. *Res. Synth. Meth.*, **1**, 39–65.

Higgins, J. P. T., Thompson, S. G. and Spiegelhalter, D. J. (2009) A re-evaluation of random-effects meta-analysis. *J. R. Statist. Soc.* A, **172**, 137–159.

Kenny, D. A. and Judd, C. M. (2019) The unappreciated heterogeneity of effect sizes: implications for power, planning of research, and replication. *Psychol. Meth.*, Feb.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, Jr, R. B., Bahnik, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C. and Cemalcilar, Z. (2014) Investigating variation in replicability. *Socl Psychol.*

Klein, R. A., *et al.* (2018) Many Labs 2: Investigating variation in replicability across sample and setting. (Available from `https://osf.io/8cd4r/`.)

Lakens, D. (2013) Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Front. Psychol.*, **4**.
Lakens, D., Scheel, A. M. and Isager, P. M. (2018) Equivalence testing for psychological research: A tutorial. *Adv. Meth. Pract. Psychol. Sci.*, **1**, no. 2, 259–269.
Longford, N. T. (1999) Selection bias and treatment heterogeneity in clinical trials. *Statist. Med.*, **18**, 1467–1474.
Louis, T. A. (1984) Estimating a population of parameter values using Bayes and empirical Bayes methods. *J. Am. Statist. Ass.*, **79**, 393–398.
Lynch, J. G., Bradlow, E. T., Huber, J. C. and Lehmann, D. R. (2015) Reflections on the Replication Corner: In praise of conceptual replications. *Int. J. Res. Marktng*, **32**, 333–342.
Mathur, M. B. and VanderWeele, T. J. (2019) New metrics for meta-analyses of heterogeneous effects. *Statist. Med.*, **38**, 1336–1342.
Mathur, M. B. and VanderWeele, T. J. (2020) Robust metrics and sensitivity analyses for meta-analyses of heterogeneous effects. *Epidemiology*, **31**, 356–358.
Monin, B. (2016) Be careful what you wish for: Commentary on Ebersole et al. (2016). *J. Exptl Socl Psychol.*, **67**, 95–96.
Monin, B. and Miller, D. T. (2001) Moral credentials and the expression of prejudice. *J. Personlty Socl Psychol.*, **81**, no. 1, 33–43.
Monin, B., Oppenheimer, D. M., Ferguson, M. J., Carter, T. J., Hassin, R. R., Crisp, R. J., Miles, E., Husnu, S., Schwarz, N., Strack, F. and Klein, R. A. (2014) Commentaries and rejoinder on Klein et al. (2014). *Socl Psychol.*, **45**, 299–300.
Nosek, B. A., Spies, J. R. and Motyl, M. (2012) Scientific utopia: II, Restructuring incentives and practices to promote truth over publishability. *Perspect. Psychol. Sci.*, **7**, 615–631.
Open Science Collaboration (2015) Estimating the reproducibility of psychological science. *Science*, **349**, article aac4716.
Patil, P., Peng, R. D. and Leek, J. T. (2016) What should researchers expect when they replicate studies?: A statistical view of replicability in psychological science. *Perspect. Psychol. Sci.*, **11**, 539–544.
Paule, R. C. and Mandel, J. (1982) Consensus values and weighting factors. *J. Res. Natn. Bur. Stand.*, **87**, 377–385.
Pettigrew, T. F. and Tropp, L. R. (2008) How does intergroup contact reduce prejudice?: Meta-analytic tests of three mediators. *Eur. J. Socl Psychol.*, **38**, 922–934.
Raudenbush, S. W. and Bryk, A. S. (1985) Empirical Bayes meta-analysis. *J. Educ. Statist.*, **10**, no. 2, 75–98.
Rice, K., Higgins, J. and Lumley, T. (2018) A re-evaluation of fixed effect(s) meta-analysis. *J. R. Statist. Soc.* A, **181**, 205–227.
Riley, R. D., Higgins, J. P. and Deeks, J. J. (2011) Interpretation of random effects meta-analyses. *Br. Med. J.*, **342**, article d549.
Rubin, D. B. (1981) Estimation in parallel randomized experiments. *J. Educ. Statist.*, **6**, 377–401.
Schweinsberg, M., Madan, N., Vianello, M., Sommer, S. A., Jordan, J., Tierney, W., Awtrey, E., Zhu, L. L., Diermeier, D., Heinze, J. E. and Srinivasan, M. (2016) The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *J. Exptl Socl Psychol.*, **66**, 55–67.
Schweinsberg, M. and Uhlmann, E. (2018) The Pipeline Project 2. European School of Management and Technology. (Available from `https://osf.io/skq2b/`.)
Shapiro, S. and Francia, R. (1972) An approximate analysis of variance test for normality. *J. Am. Statist. Ass.*, **67**, 215–216.
Shapiro, S. and Wilk, M. B. (1965) An analysis of variance test for normality (complete samples). *Biometrika*, **52**, 591–611.
Simons, D. J. (2014) The value of direct replication. *Perspect. Psychol. Sci.*, **9**, 76–80.
Simons, D. J., Holcombe, A. O. and Spellman, B. A. (2014) An introduction to Registered Replication Reports at *Perspectives on Psychological Science*. *Perspect. Psychol. Sci.*, **9**, 552–555.
Simonsohn, U. (2015) Small telescopes: detectability and the evaluation of replication results. *Psychol. Sci.*, **26**, 559–569.
Stewart, G. B., Altman, D. G., Askie, L. M., Duley, L., Simmonds, M. C. and Stewart, L. A. (2012) Statistical analysis of individual participant data meta-analyses: a comparison of methods and recommendations for practice. *PLOS One*, **7**, no. 10, article e46042.
Sutton, A. J., Abrams, K. R., Jones, D. R., Sheldon, T. A. and Song, F. (2000) *Methods for Meta-analysis in Medical Research*. Chichester: Wiley.
Verhagen, J. and Wagenmakers, E.-J. (2014) Bayesian tests to quantify the result of a replication attempt. *J. Exptl Psychol.*, **143**, 1457.
Veroniki, A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J. P., Langan, D. and Salanti, G. (2016) Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Res. Synth. Meth.*, **7**, 55–79.
Viechtbauer, W. (2010) Conducting meta-analyses in r with the metafor package. *J. Statist. Softwr.*, **36**, no. 3.

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, Jr, R., Albohn, D. N., Allard, E. S., Benning, S. D., Blouin-Hudon, E. M. and Bulnes, L. C. (2016) Registered Replication Report: Strack, Martin, & Stepper (1988). *Perspect. Psychol. Sci.*, **11**, 917–928.
Wang, C.-C. and Lee, W.-C. (2019) A simple method to estimate prediction intervals and predictive distributions: summarizing meta-analyses beyond means and confidence intervals. *Res. Synth. Meth.*, **10**, 255–266.

Supporting information
Additional 'supporting information' may be found in the on-line version of this article:

'Supplement: New statistical metrics for multisite replication projects'.