Supplement: Estimating Publication Bias in Meta-Analyses

Contents

1. Supplementary methods

1.1. Data extraction methods

We eliminated studies from the grey literature as follows. For many meta-analyses, the methods section suggested no attempts to include studies from the grey literature, in which case we simply included all point estimates in our analysis. When the methods section instead indicated some attempt to include studies from the grey literature, we contacted the meta-analysts or examined the reference lists of included studies in order to identify and exclude the studies from the grey literature. When possible, we treated as "published" any article published in a peer-reviewed journal or conference proceeding; however, we sometimes had to rely instead on the meta-analysts' own definitions of "published".

1.2. Dual coding and data entry quality checks

A team of six research assistants (Acknowledgments) and MBM extracted data, with two coders independently extracting data for every analyzed meta-analysis (except Metalab, which was singly coded because the publicly available datasets were already curated in an analysis-friendly format). The research assistants completed a customized 54-minute course of training videos (`https://osf.io/6dbhp/`) detailing how to extract data from the meta-analyses. To code journal tiers, we downloaded the full SciMago database of 2018 ratings (*Scimago Journal and Country Rank*, n.d.). We used an R script to standardize and merge journal titles from the SciMago database with those in our meta-analysis corpus. Some journals in our corpus did not have an exact match in the SciMago database because, for example, the title included a subtitle or section within the journal, the title we entered was abbreviated whereas the SciMago title was unabbreviated, the title had special characters or accents, the citation in paper was incomplete or misspelled, etc. For all such unmatched journals, we manually coded their Scimago ratings by splitting the work across four coders; finally, MBM manually checked and, if necessary, corrected every entry. We then used an R script to merge the resulting SJR dataset with our meta-analysis corpus, conducting sanity checks and data cleaning. For example, some journals had multiple, discrepant rankings because they had non-unique titles (e.g., *Surgery*); we removed these ambiguous journals from our SJR database so that they would result in missing data. Ultimately, 1.4% of point estimates had missing data on journal. Of the 1107 unique journals in our meta-analysis

corpus, rankings were hand-coded for 233 (23%) journals. Rankings were hand-coded for 11% of all point estimates in our corpus.

Upon the completion of data entry, we used an R script to check for extreme or incompatible values for all numerical entries, manually confirming or correcting each of these entries. We additionally compared the dual-coded datasets; where there were discrepancies, coders attempted to resolve these through discussion. When discrepancies of $\geq 5\%$ remained on analysis variables (e.g., the estimated selection ratio), MBM manually reviewed both coders' datasets, choosing one dataset for analysis by preferring datasets that: (1) proved to be correct on manual review of each point estimate and inference entry; (2) were prepped automatically in R by MBM rather than entered manually; and/or (3) exactly reproduced the paper's reported estimates when this was expected because the meta-analysis contained no studies from the grey literature. For two meta-analyses, limitations of analytic reproducibility precluded resolution of discrepancies (e.g., because there were inherent ambiguities in how to link study citations with study abbreviations in forest plots or because documentation regarding which studies were unpublished was unclear). Specifically, for PMID 26724178, the forest plot listed trial acronyms rather than unique publications, with each trial potentially yielding many separate publications. For PMID 28159391, most point estimates were from large public databases rather than publications, and point estimates from publications had no indication of which publication they were from. We excluded these meta-analyses from analysis as depicted in the PRISMA flowchart.

As a post hoc addition introduced during peer review, the first author (MBM) coded each meta-analysis by study design (all observational, all randomized, or both). We coded as "all randomized" meta-analyses whose inclusion criteria required randomization or the use of "within-subjects manipulations" or "between-subjects manipulations". The latter two terms are often used in experimental psychology, in which randomization is often assumed. We did not attempt to distinguish whether studies were randomized with respect to the actual exposure of interest or were randomized with respect to some other exposure. We treated all other study designs, including quasi-experiments, as "observational". We also coded as "all observational" some meta-analyses that made no reference to study designs, but in which the research question strongly suggested that only observational designs would be possible (e.g., the exposure was gender) *and* in which no included study's title contained the string "random*". We coded as "unclear" those meta-analyses that, for example: (1) made no reference to study designs and in which the research question did not clearly preclude randomized studies; or (2) whose inclusion criteria specified that both study designs were eligible, but that did not provide information on what designs were ultimately represented in

the analyzed studies.

2. Supplementary results

2.1. Sensitivity analyses for violations of model assumptions

To assess for violations of the assumption that publication bias operates in favor of affirmative results (i.e., those with $p < 0.05$ and point estimates in the desired direction), we calculated and plotted one-tailed p-values from all studies in our dataset, treating the direction of the meta-analytic point estimate as the desired direction (Figure S1). The much larger mass of one-tailed p-values below 0.025 (50% of all p-values) versus those above 0.975 (4% of p-values) suggested that selection indeed was primarily one-directional, though a small mass above 0.975 suggests some weak two-tailed selection (i.e., selection favoring "significant" results regardless of sign). As a simple measure of apparent two-tailed selection in each meta-analyses, we calculated the ratio of the observed proportion of nonaffirmative studies with one-tailed $p > 0.975$ to its expectation under the assumption that the p-values of all non-affirmative studies are uniformly distributed. Since nonaffirmative studies are those with a one-tailed $p > 0.025$, the expectation is therefore $0.025/0.975 \approx 0.0256$. In the below sensitivity analyses, we excluded meta-analyses for which this ratio exceeded 3.

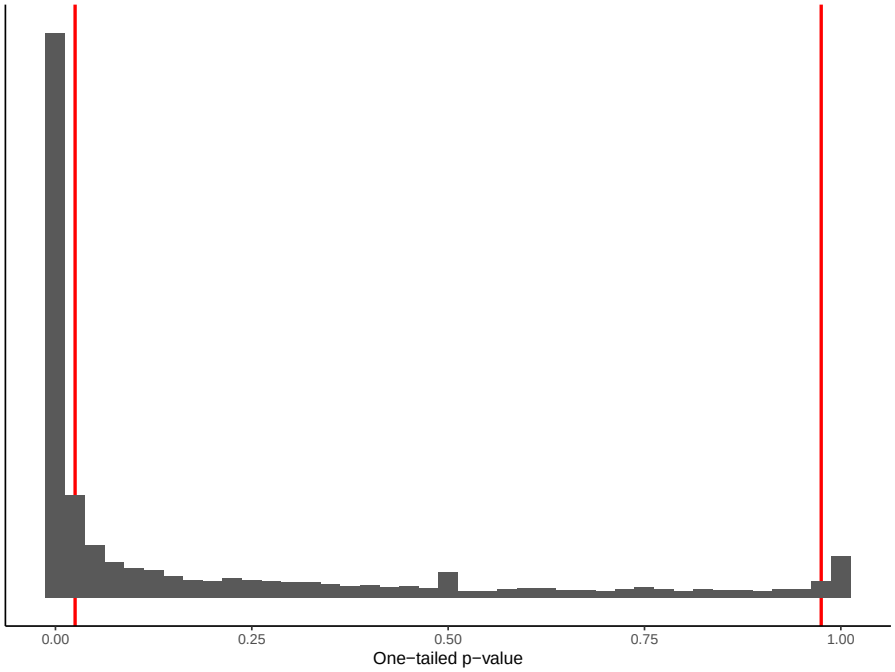

Figure S1: *One-tailed p-values from all meta-analyses, treating the direction of the meta-analytic point estimate as the desired direction. Red lines indicate the 0.025 and 0.975 thresholds, i.e., the thresholds at which the corresponding two-tailed p-value would be* < 0.05 *and in the desired direction and at which the two-tailed p-value would be* < 0.05 *but in the unanticipated direction.*

A second plausible threat to model assumptions is non-normal true effects, which we assessed by excluding meta-analyses for which a Shapiro test of the normalized point estimates yielded $p < 0.05$ (Hardy & Thompson, 1998; Shapiro & Wilk, 1965). This criterion is conservative in that the selection model assumes that the latent true effects are normal *prior to* selection due to publication bias, so meta-analyses with non-negligible publication bias may have normal true effects in the latent population despite having non-normal point estimates. A third potential threat is our inclusion of at least two meta-analyses (PMID 27416099 and 27835651) in which the authors coded as "0" the point estimate for any study that reported only a "nonsignificant" effect, creating point masses of estimates at exactly 0. These point masses would violate the normality assumption as well as produce a downward-biased estimate of the selection ratio. Our sensitivity analyses, below, excluded meta-analyses in which $> 5\%$ of estimates were coded as exactly 0. Finally, in one meta-analysis (PMID 28700728), the original dataset coded effects were coded in an internally inconsistent manner, rendering the direction of the point estimates meaningless. We additionally excluded this meta-analysis in sensitivity analyses. Table S1 summarizes the effects of applying each exclusion criterion, or all criteria simultaneously, on an overall estimate of the selection ratio. These results suggests

that while many meta-analyses failed the stringent sensitivity analysis criteria, the resulting pooled point estimates were not substantially affected.

Possible threat	k	$\widehat{SR}$ [95% CI]	Max $\widehat{SR}$	q_{95}
Two-tailed selection	36	1.05 [0.80, 1.38]	7.80	2.26
Non-normality	32	1.32 [1.03, 1.70]	7.80	2.52
Point mass at zero	51	1.21 [0.94, 1.56]	54.77	3.96
Other	58	1.17 [0.93, 1.47]	54.77	3.51
All of above	17	1.29 [0.87, 1.92]	7.80	2.57

Table S1: *Effect of sensitivity analyses on overall estimate of selection ratio. k: number of meta-analyses included in the sensitivity analysis.*

2.2. Sensitivity analyses excluding *Journal of Educational Psychology*

Because a single journal (*Journal of Educational Psychology*) contributed a particularly large percentage of higher-tier point estimates (47%), we conducted a sensitivity analysis in which we recoded this journal as "lower-tier". After doing so, we estimated that higher-tier results were 0.82 (95% CI: [0.72, 0.94]; $p = 0.00$) times as likely to be affirmative as lower-tier results.

3. Supplementary figures

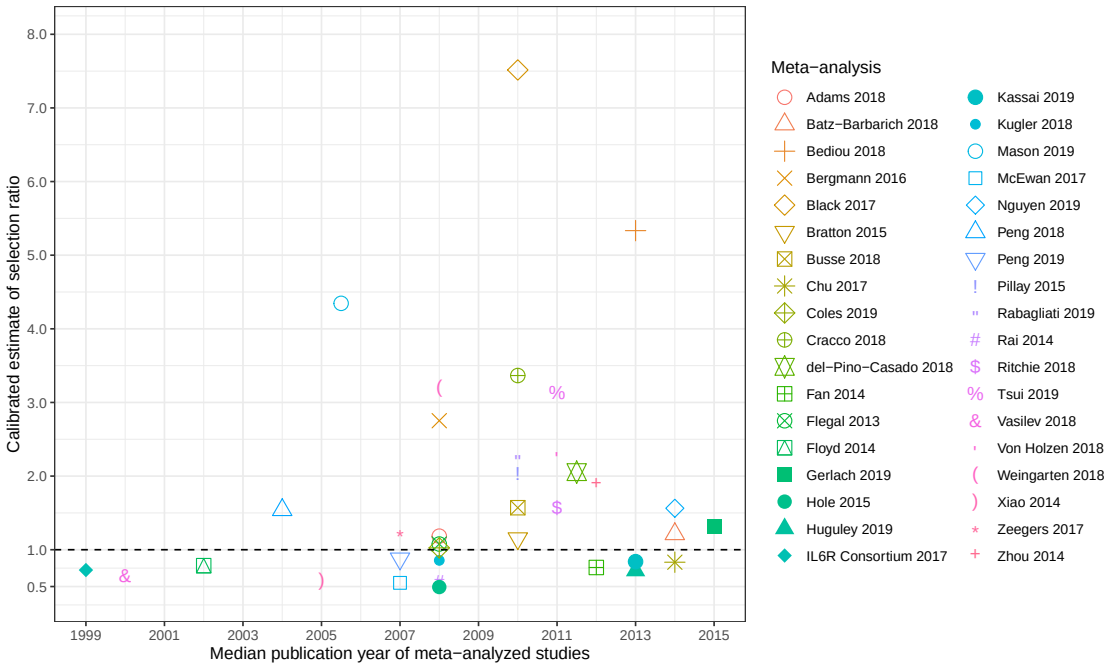

Figure S2: *Median publication year of meta-analyzed studies versus calibrated estimate of selection ratio. Horizontal dashed line: null.*

4. Changes and additions to preregistered protocol

During article review, we decided to exclude network meta-analyses because these typically do not have study-level point estimates, though we made exceptions for network meta-analyses that also presented standard pairwise meta-analyses. We had originally planned to classify as "early" those studies that were "among the chronologically first three point estimates"; however, due to the large number of overlapping study years, this criterion appeared too lenient, so we adopted the criterion described in the main text. Regarding thresholds for "higher-tier" journals, we had initially planned to set the threshold for psychology to 3.25 and the threshold for medicine to 7.4 so that the lowest-ranked higher-tier journals in each category would be *Journal of Experimental Psychology: General* and *Annals of Internal Medicine*; we revised this threshold when new rankings became available after the preregistration was published. The preregistration indicated that we would consider the percentage of "statistically significant" results without specifying whether this would include results with point estimates in either direction or only affirmative results. For consistency with the selection models in main

analyses, we chose to use affirmative status as the primary outcome and secondarily present analyses for "significant" results in either direction. The preregistration did not describe how we would conduct inference for the study-level measures, leading us to introduce the robust GEE models post hoc. All analyses described as sensitivity analyses or exploratory analyses were introduced post hoc before or during peer review.

References

Hardy, R. J., & Thompson, S. G. (1998). Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine*, *17*(8), 841–856.

Scimago journal and country rank. (n.d.). https://www.scimagojr.com/. (Accessed: 2019-07-08.)

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3/4), 591–611.