These files replicate the results in the initial (April 2017) version of "Identification of and Correction for Publication Bias." Replication files (in Matlab) for more recent versions are available upon request.

This document includes instructions for estimating the selection model and calculating selection corrected estimates based on 1) replication studies and 2) meta studies using `ApplicationScript`, a wrapper script in both Matlab and R. It also provides documentation on Matlab and R functions called in the script. To replicate results in the paper, please refer to `AllApplication.m` for Matlab and `Main.R` for R.

For the selection model, we estimate mean $\bar{\theta}$ and variance τ^2 of μ, and the relative publication probability $\beta_{p,k}$. The distribution of the true effect Θ^*, denoted μ, is assumed to be normal. The publication probability $p(\cdot)$ is assumed to be a step function with cutoff specified by the users. Cutoffs can be conventional critical values or zero. For more details on the selection model, see Section 3 of the paper. For more details on parametric specifications, see Section C of the Supplement.

Selection corrected estimates are based on the estimated $p(\cdot)$ returned above. Section 4 of the paper derives median unbiased estimators and valid confidence sets for θ, assuming a normal distribution of the latent study estimates X^*. We apply this approach and calculate median unbiased estimators and confidence sets for each original study estimate.

In `ApplicationScript`, first set the path accordingly and create a folder `FiguresandTables` to save figures and tables following comments in the script. Based on your study type (replication studies or meta studies), refer to the corresponding section of this readme and follow instructions therein.

1 Replication studies

Follow the instructions below to estimate the selection model based on replication studies. In `ApplicationScript`, set the variable `identificationapproach` to one. Instructions below are specific to the corresponding block of code. See Section 5.1 and 5.2 of the paper for examples of the replication-studies approach.

1.1 Data preparation

To apply the replication-studies approach, we need z-statistics and standard errors from both the original and replication studies. To obtain effect size estimates, apply the Fisher transformation to standardized effect sizes reported by the studies. Dividing these estimates by the z-statistics recovers the standard error.

For examples of preparing original and replication estimates dataset, see
`./Data Cleaning/clean_econ_replication_data` and `./Data Cleaning/clean_psych_replication_data.m`.

`ApplicationScript` reads in a dataset with rows (X1, sigma1, X2, sigma2). X1 and sigma1 are standardized effect and standard error of the original study. X2 and sigma2 are standardized effect and standard error of its corresponding replication study. It can also read in a list of names of original studies for outputting purpose. This list should be stored in variable `Studynames`. This list needs to

be in the same order as the dataset of original and replication estimates dataset.

Assign name for outputs in variable `name`. All tables and figures from estimation are saved in `./FiguresandTables/`. with file names containing `name`.

Variable used for estimating the selection model are:

- `X`=X1
- `sigma`=sigma1
- `Z`=[X1/sigma1 X2/sigma1][1]
- `sigmaZ2`=sigma2/sigma1

1.2 Parametric specification

In `ApplicationScript`, we include an indicator variable `symmetric`. If `symmetric` is set to one, then we impose that $p(\cdot)$ is symmetric and all cutoffs should be positive. Furthermore, $\bar{\theta}$ is fixed at zero and excluded from estimation.[2] If `symmetric` is set to zero, then we do not impose that $p(\cdot)$ is symmetric and cutoffs should be specified in ascending order. Cutoffs are stored in vector `cutoffs`. We also need to specify starting values for $\bar{\theta}$,τ^2, and $\beta_{p,k}$ in the vector `Psihat0`. Note that when `symmetric` is set to one and $\bar{\theta}$ is fixed at zero, we do not specify a starting value for $\bar{\theta}$.

Assign these variables based on your dataset and parametric specification, but do not change the variable names `cutoffs`, `name`, `Psihat0`, `Studynames`, `symmetric`, `X`, `sigma`, `Z`, and `sigmaZ2`. The rest of the variables in this block of code, `n` and `cluster_ID`, can be left as they are. Then execute the section "data preparation and parametric specification" of `ApplicationScript`.

Based on your goal (to produce figures of study estimates, to estimate the selection model, or to calculate selection corrected estimates), run the corresponding section of code in `ApplicationScript` as described in Section 3 of this readme.

2 Meta studies

Follow the instructions below to estimate the selection model based on meta studies. In `ApplicationScript`, set the variable `identificationapproach` to two. Instructions below are specific to the corresponding block of code. See Section 5.3 and 5.4 of the paper for examples of the meta-studies approach.

2.1 Data preparation

To apply the meta-studies approach, we need estimates and standard errors from published studies. There can be multiple estimates per study, and for inference on estimates of the selection model, we cluster standard errors at the study-level.

[1] Both original and replication estimates are normalized by the variance of the original estimate.

[2] When the sign of the original estimates is normalized to be positive, we impose that $p(\cdot)$ is symmetric, and that the mean of Θ^*, $\bar{\theta}$ is zero. Details and further motivation for these specifications, as well as a specification for the model of Section 3.1.3, are discussed in Section C of the supplement.

For examples of preparing meta-studies dataset, see
`./Data Cleaning/clean_min_wage_data` and `./Data Cleaning/clean_deworming_data.m`.

`ApplicationScript` reads in a dataset with rows (X1, sigma1, cluster_id). X1 and sigma1 are estimate and standard error from some published study. cluster_id is a factor variable, assigning the estimate to its corresponding study. It can also read in a list of names of original studies for outputting purpose. This list should be stored in variable `Studynames`. This list needs to be in the same order as the dataset of original and replication estimates dataset.

Assign name for outputs in variable `name`. All tables and figures from estimation are saved in `./FiguresandTables/`. with file names containing `name`.

Variable used for estimating the selection model are:

- `X`=X1
- `sigma`=sigma1
- `cluster_ID`=cluster_id
- `includeinestimation`=an indicator variable for whether the estimate is used in estimating the selection model. Set it equal to a one vector if all estimates are used in the estimation.

2.2 Parametric specification

In `ApplicationScript`, we include an indicator variable `symmetric`. If `symmetric` is set to one, then we impose that $p(\cdot)$ is symmetric and all cutoffs should be positive. Furthermore, $\bar{\theta}$ is fixed at zero and excluded from estimation.[3] If `symmetric` is set to zero, then we need to specify a starting value for $\bar{\theta}$ in the vector `Psihat0`. We include another indicator variable `symmetric_p`. If `symmetric_p` is set to one, then we impose that $p(\cdot)$ is symmetric and all cutoffs should be positive; if `symmetric_p` is set to zero, then we do not impose that $p(\cdot)$ is symmetric and cutoffs should be specified in ascending order. We also need to specify starting values for $\bar{\theta}$,τ^2, and $\beta_{p,k}$ in the vector `Psihat0`. Note that when `symmetric` is set to one and $\bar{\theta}$ is fixed at zero, we do not specify a starting value for $\bar{\theta}$.

Assign these variables based on your dataset and parametric specification, but do not change the variable names `cutoffs`, `name`, `Psihat0`, `Studynames`, `symmetric`, `X`, `sigma`, `cluster_ID`, and `includeinestimation`. The rest of the variables `n` can be left as it is. Then execute the section "data preparation and parametric specification" of `ApplicationScript`.

Based on your goal (to produce figures of study estimates, to estimate the selection model, or to calculate selection corrected estimates), run the corresponding section of code in `ApplicationScript` as described in Section 3 of this readme.

3 Matlab and R functions

After "data preparation and parametric specification" section, there are three sections of code `ApplicationScript` that estimates the selection model and calculates selection corrected estimates.

[3] When the sign of the original estimates is normalized to be positive, we impose that $p(\cdot)$ is symmetric, and that the mean of Θ^*, $\bar{\theta}$ is zero. Details and further motivation for these specifications, as well as a specification for the model of Section 3.1.3, are discussed in Section C of the supplement.

1. Producing figures. This section produces figures of study estimates, which can be skipped if the users are only interested in estimates of selection model and/or selection corrected estimates. The figures produced are

   (a) A binned density plot for the normalized z-statistics of the original study estimates;

   (b) A scatter plot of the original study z-statistic against its corresponding replication study estimate, normalized by the variance of the original study estimate (replication study only);

   (c) A scatter plot of the original study estimate against its standard error.

2. Estimating the selection model. `EstimatingSelection` is a wrapper script for functions that estimate the selection model.

   - To estimate the selection model based on replication studies, this wrapper calls function `ReplicationAnalyticLogLikelihood`. For more details, see Section C.1 of the Supplement.
   - To estimate the selection model based on meta studies, this wrapper calls function `VariationVarianceLogLikelihood`. For more details, see Section C.2 of the Supplement.
   - To estimate variance of estimates of the selection model, this wrapper calls function `RobustVariance`, which then calls `Clustered_covariance_estimate`. Variance is robust to clustering if users specify cluster design in parametric specification.
   - To output estimates of the selection model along with their standard errors, this wrapper calls function `SelectionTable`, which produces a TeX table. For replication studies, estimates based on both replication studies and meta studies (using initial studies) are provided.

3. Calculating selection corrected estimates. `HorizontalBars` calls `Step_function_normal_cdf` to calculate the density of published study, which is then used to estimate median unbiased θ and provide inference for a given initial study Z score. Note that this section takes the estimated $p(\cdot)$ as an input, which is returned by `EstimatingSelection` above. In order to run `HorizontalBars` , please first execute `EstimatingSelection`. It then produces the following figures

   - A plot of original, corrected z-statistics. When there are replication studies, replication results are also plotted. See the upper panel of Figure 6 of the paper for an example.
   - A comparison between median unbiased estimators and corrected confidence sets and original estimates and confidence sets. Since estimates here are z-statistics, original confidence sets are simply original z-statistics plus and minus 1.96. See the lower panel of Figure 6 of the paper for an example.
   - A plot of median unbiased estimators and corrected confidence sets analogous to Figure 4 of the paper.

- Lastly, we output median unbiased estimators and corrected confidence sets corresponding to each original estimates to a csv file.