

EEF Evidence Database Phase 2:

Protocol and Analysis Plan

Steve Higgins, Berenice Villanueva Aguilera, Emma Dobson, Louise Gascoine, Afroditi Kalambouka, Maria Katsipataki, Jonathan Reardon, Margherita Rovida and Germaine Uwimpuhwe

January 2024



EEF Evidence Database Phase 2 Protocol and Analysis Plan

Protocol and Analysis Plan

Principal investigator: Professor Steve Higgins

Table of contents

Table of contents 2

Background and review rationale 3

Objectives 3

Research questions 4

Methodology 4

Recording and reporting 11

Personnel 11

Conflicts of interest 11

Ethical issues 11

Acknowledgements 11

References 12

EEF Evidence Database Phase 2 Protocol and Analysis Plan
Protocol and Analysis Plan
Principal investigator: Professor Steve Higgins

Background and review rationale

- The Education Endowment Foundation (EEF) - Sutton Trust Teaching and Learning Toolkit ('Toolkit') is an accessible summary of 30 areas of educational practice that is used by practitioners in England and by EEF's international partners around the globe. The evidence summaries for each strand draw on a series of meta-analyses, completed in Autumn 2021 (EEF Evidence Database Phase 1), of the studies that were included in the reviews on which the previous version of the Toolkit was based (see Higgins et al. 2022 for a review of the methodologies). A protocol and analysis plan for Phase 1 of the Database project was completed in June 2019 (Higgins et al. 2019). A version for early years practitioners was updated in February 2023 following the same methodology.
- A limitation of the current version of the database is that the coverage of studies in each strand varies according to the inclusion criteria of the underpinning reviews which were used to identify the screened studies. To ensure that the coverage of studies is equivalent across all of the Toolkit strands a systematic search needs to be conducted across each of the strands.
- The aim of this phase of the database project, therefore, is to update each of the strands by systematically searching for missing studies using OpenAlex[1], an open source bibliographic database, and some of the machine learning tools for search and screening available within EPPI-Reviewer[2].
- The ultimate aim (Phase 3) is to create a series of 'living reviews' for each of the Toolkit strands which can be updated as new evidence is identified.

Objectives

The specific objectives for this phase of the project are:

- To update the web-based database of the single studies included in the current Toolkit, which meet a consistent set of inclusion criteria, with a series of systematic reviews for missing studies for each Toolkit strand;
- To re-estimate the pooled effect for each Toolkit strand based on this database using meta-analytic approaches;
- To explore variation in impact systematically, both within each strand and across strands:
    - for pedagogical variation (such as by subject, by age of learner or by intervention duration and intensity);
    - for methodological variation (such as design features, sample size or outcome and measurement artefacts)
    - for features associated with rigour and robustness of evaluation methods (such as randomisation, attrition, independence of evaluation) to identify potential bias.
- To create a series of reviews and machine learning classifiers with which to update the Toolkit as a set of linked living reviews.

[1] https://openalex.org
[2] https://eppi.ioe.ac.uk/CMS/Portals/35/machine_learning_in_eppi-reviewer_v_7_web_version.pdf

EEF Evidence Database Phase 2 Protocol and Analysis Plan
Protocol and Analysis Plan
Principal investigator: Professor Steve Higgins

Research questions

The overarching primary research question is framed as follows:

- What is the average impact in each of the areas of the Teaching and Learning Toolkit on pupils' attainment in schools, compared with pupils who do not experience these interventions or approaches?
  What is the average impact in each of the areas of the Early Years Toolkit on pupils' attainment in schools, compared with pupils who do not experience these interventions or approaches?

In addition, there are two supplementary questions:

- How consistent is this impact by age of pupils and by subject outcome?
- What factors help to explain any variation in impact?

Methodology

Inclusion and exclusion criteria for the review

- The inclusion criteria were defined for Phase 1 of the database project (see the protocol and analysis plan: Higgins et al. 2019) and are set out below. These have been revised and updated to address inconsistencies and ambiguities identified during their use in Phase 1.

Inclusion criteria	**Excluded**
Population The majority of the sample (>50%) on which the analysis of impact in the study is based are learners or pupils aged between 3-18 (further education or junior college students are included where their study is for school level qualifications).	The majority of the sample (>50%) are: in post-secondary education; in higher education; adults; infants under 3; other students over 18. The sample consists only or mainly of learners of a second or additional language where these are a minority in the setting and where the language is not the usual language(s) of instruction for that setting.
Intervention Evaluates the impact of an educational intervention or approach, including named or clearly defined programmes and recognisable approaches classifiable according to the Toolkit strand definitions (see Appendix 1..	The intervention or approach is not applicable to the current Toolkit strand definitions (see Appendix 1). Interventions designed only for second or additional language learners of the language(s) of instruction.
Setting	

The intervention or approach is undertaken in a normal educational setting or environment for the learners involved, such as a nursery or school or a typical educational setting for school age pupils (e.g. an outdoor field centre or museum).	Laboratory studies. Specially created environments (either physical or virtual) designed for theoretical research questions, rather than for learners' educational benefit.
Study design A valid counterfactual comparison between those receiving the educational intervention or approach and those not receiving it.	Single group and single subject designs where there is no control for maturation or growth.
Outcomes Assessment of educational or cognitive achievement which reports quantitative results from testing of attainment or learning outcomes such as by standardised tests or other appropriate curriculum assessments or school examinations or appropriate cognitive measures relating to the curriculum.	Attitudinal, affective or motivational outcomes. Cognitive or metacognitive measures of learning processes (e.g. working memory, or metacognitive monitoring).
A quantitative estimate of the impact of the intervention or approach on the educational attainment of the sample involved in the intervention or approach can be calculated or estimated in the form of an effect size (standardised mean difference) with its standard error based on a counterfactual comparison.	Purely qualitative outcomes. Studies where an effect size (standardised mean difference) and standard error cannot be identified, calculated or estimated with reasonable precision.[3]
Language Studies written in English.	Studies in languages other than English.

Search strategy for identification of studies

Studies for each strand will be identified using keywords and terms developed in Phase 1 (e.g. using the Toolkit strand specific keywords and some named programmes and interventions), together with the studies screened previously for each of the Toolkit strands. These studies will then be used to identify potentially similar studies via network graph searches in OpenAlex[4], a database of millions of interconnected research references or 'entities', using the machine learning tools in EPPI-Reviewer 6. This approach needs data from which to 'learn'; the previous screening decisions from Phase 1 of

[3] Such as by using the conversions available in David B. Wilson's online conversion tool: https://campbellcollaboration.org/escalc/html/EffectSizeCalculator-Home.php .
[4] https://openalex.org

the project should provide sufficient information to begin this process (Stansfield et al. 2019). A separate search and retrieval review for each Toolkit strand will be set up in EPPI-Reviewer 6 to manage the searching and screening process. Studies from other sources, screened for strand relevance, may also be used to teach the machine learning tools at this stage.

Additional comparison searches will be undertaken in Proquest Dissertations and Theses Global[5] and a database of UK dissertation abstracts (EThOS[6]), as preliminary exploration indicates these sources may not be comprehensively included in OpenAlex. These searches will be discontinued if the studies identified are found to be included in the OpenAlex searches.

Selection of studies

Customised machine classifiers will be developed for each of the Toolkit strands to reduce the total number of studies that need to be screened manually[7]. The aim is to create a search and screening process for each Toolkit strand which can then be used to update the strands as a series of linked 'living reviews' (Elliot et al. 2017). The studies will be screened using the 'priority screening' tools in EPPI-Reviewer 6[8] (Thomas et al. 2017).

Screening will be undertaken in two stages. The first uses the title and abstract, and the focus is on identifying studies which can clearly be *excluded*. The second uses the full text, where the focus is on ensuring that the study meets *all* of the inclusion criteria. A study will be excluded once a single inclusion criterion is identified that it does not meet. A minimum of 10% of studies will be double screened at each stage.

Studies which are included at title and abstract stage will then be retrieved for full text screening. Studies which are identified as borderline will be reviewed by two coders and referred to a third if agreement cannot be reached, or where there is still some uncertainty.

At this stage, the focus is on identifying studies for data extraction (Lefebvre et al. 2022). Sometimes, a report includes more than one study, in which case the report will be duplicated for each included study (for example a journal article or a dissertation describing a series of experiments with different samples of pupils). In some cases, there are multiple reports on the same study. This can be challenging to identify, especially where there are different authoring teams. These will usually be identified and linked at full text screening, but further checks will be made at data extraction, data cleaning and analysis. Linked reports will usually have copies of all of the relevant reports made available for data extraction in EPPI-Reviewer 6.

Results will be tracked in EPPI-Reviewer 6 and recorded using Excel to keep track of the process, equivalent to the information in a PRISMA flow chart. Each strand will be updated once following this approach, so that future updates can adopt a living review methodology. The order for updating strands will take account of EEF priorities and will be agreed with the Synthesis Team.

[5] https://about.proquest.com/en/products-services/pqdtglobal/
[6] https://ethos.bl.uk/
[7] https://eppi.ioe.ac.uk/CMS/Portals/35/machine_learning_in_eppi-reviewer_v_7_web_version.pdf
[8] https://eppi.ioe.ac.uk/CMS/Portals/35/Manuals/ER%20Web%20-%20Priority%20Screening.pdf

Data extraction and management

Data extraction will follow the same procedures as in Phase 1 using the same data extraction tools: EEF main data extraction (v 2.0 March 2022), EEF Toolkit effect size data extraction (v 2.0 March 2022) and EEF Toolkit strand specific data extraction (see Appendices 1-3). Guides for coders are available which have been updated with additional examples to clarify codes where issues had been identified in Phase 1 (Main Data Extraction and Effect Size Data Extraction Coding Guides: see Appendices 4 and 5).

A minimum 10% sample of studies will be double coded (by strand and by coder) to ensure reliability.

A data review and cleaning process will be undertaken for each strand prior to analysis. This includes the facility to review the text highlighted in each study report and to compare it with the associated code.

Appraisal of included studies

Two approaches to the appraisal of included studies will be undertaken. The first is to identify a 'padlock' rating for each strand. This is based on an aggregation of a number of the data extraction questions and includes the total number of studies in a strand, how recent the studies are, and the proportion of studies which are randomised, are independently evaluated, and have high levels of attrition.

Second, an additional risk of bias summary is being developed for each of the single studies based on a synthesis of a number of existing risk of bias and quality appraisal tools[9]. This will similarly be based on an aggregation of a number of the data extraction questions including randomisation, sample size, attrition, type of analysis and ecological validity.

Effect size calculation

The database includes and summarises quantifiable school attainment outcomes from experimental studies which meet the inclusion criteria and match the Toolkit themes. The key metric used is the Standardised Mean Difference or effect size (Cohen's d or Hedges' g) as a range of tests and outcome measures are used for educational attainment.

Where possible, a post-test comparison of the impact of the intervention or approach, adjusted for baseline and clustering of pupils in classes, will be selected as the primary outcome. If this is calculable[10] from other reported data then this effect size will be used. Where this is not available the following order of preference will be used:

- Post-test comparison adjusted for baseline (such as from ANCOVA or MLM/HLM);

[9] Common risk of bias tools include the Risk of Bias 2 tool for RCTs and ROBINS-I for quasi-experimental studies. Tools for assessing the quality of qualitative research include the Critical Appraisal Skills Programme (CASP) checklist for qualitative research or adapted versions of this checklist.

[10] For example, by using the effect size calculator created by David Wilson (Lipsey & Wilson, 2001) and available on the Campbell website: https://www.campbellcollaboration.org/research-resources/effect-size-calculator.html or Psychometrika: https://www.psychometrica.de/effect_size.html

- Post-test only comparison (for randomised studies or matched studies where there is no or little baseline imbalance[11]);
- Gain score comparison (for non-randomised studies or where there is notable imbalance at baseline);

For studies reporting only descriptive statistics for continuous measures of pupil attainment outcomes, the post-intervention mean of the control group will be subtracted from the post-intervention mean of the intervention group and the resulting difference will be divided by the pooled standard deviation, adjusted for sample size if necessary (Hedges' g). An accompanying standard error (representing the 95% confidence interval) will also be recorded[12]. Wherever possible the descriptive outcome statistics (N, means and standard deviations for control and intervention groups) will be collected, even where the study report reports an effect size and accompanying standard error, or where an effect size can be calculated from other inferential statistics.

All effect sizes will be coded according to the type of outcome calculation used for both risk of bias and sensitivity analysis (see Appendix 2). These effect sizes can then be meta-analysed separately as they may represent different metrics (such as when the intervention affects the relative spread of the intervention group, see Xiao et al. 2016). For studies where there is substantial baseline imbalance, a gain score effect size will be selected (such as in quasi-experimental designs or natural experiments where groups are not well matched).

The Toolkit focuses on academic attainment outcomes. It is likely that in some studies there will be several measures of the same or similar outcomes from the same sample of learners. When this happens, we will select the most representative measure of overall attainment, such as a standardised test of reading comprehension. When no single outcome is judged to be appropriate in relation to the design of the study and the Toolkit strand it is included in, we will average effects deriving from similar or complementary measures of school attainment. If the same group of participants is used more than once (such as the same control group compared with two different treatment groups, each applicable to the Toolkit) the sample size and associated standard error will be adjusted to ensure the study contributes fairly to the overall average.

Dealing with missing data

Outcome data are likely to be reported in a variety of formats in educational studies and there are likely to be studies with incomplete or missing descriptive data. For studies that report inferential statistics such as t, F, or p-values only, the appropriate conversion formula will be applied to calculate the *d*-index as the effect size estimate (Lipsey & Wilson, 2001; Hedges, Shymansky, & Woodworth, 1989; Hedges & Olkin, 1985). To ensure appropriate corrections for small sample size bias, all *d*-indices will be converted to the unbiased Hedges' *g* statistic where necessary. Adjustment for clustering will also be undertaken where there is sufficient information to do this.

Due to the large number of studies involved in the Toolkit and the possible age of the older studies we will not normally contact the authors about incomplete, missing or apparently incorrect data. This may mean some studies are excluded at a late stage in the review process.

[11] Chance imbalance is likely to occur in some randomised studies (the smaller the study the greater the risk) but can usually be taken into account through a baseline adjusted analysis.
[12] The standard error can be calculated from the confidence intervals or estimated from p-values: https://handbook-5-1.cochrane.org/chapter_7/7_7_7_2_obtaining_standard_errors_from_confidence_intervals_and.htm

Data synthesis

For each strand, a primary standardised mean difference (SMD) and its confidence interval will be identified and extracted from each of the selected studies for each Toolkit strand. Estimates from appropriate trials contributing to a specific strand will be synthesised into a single pooled effect using random effects meta-analysis using restricted maximum likelihood (REML) estimation method following Langan et al. (2019) and Viechtbauer's (2005) recommendations, using the R package Metafor (Viechtbauer, 2010). Wherever relevant, sub-group or 'mini-strand' analyses will be undertaken using a similar approach. A separate pooled effect will be calculated for each Toolkit strand for reading, mathematics and science outcomes and for early years, primary and secondary age pupils where data is available. (A sample R script for Metafor can be found in Appendix 6.)

As noted above, a single overall main effect will be selected for each study. Some studies may include multiple comparisons for different groups where the interventions or approaches for the different groups can be distinguished in the data extraction coding. In this case standard errors will be adjusted to take account of dependency or, where appropriate, a multi-level meta-analysis will be undertaken. Where there are fewer than ten studies in a strand (or mini-strand) a meta-analysis will not be undertaken as the estimate may be unreliable.

Investigation of heterogeneity

The I^2 statisticp will be used to assess the relative heterogeneity (Borenstein et al. 2010; Schmid, Stijnen & White 2020). It is possible that the estimated variability may not only be due to random factors but may be partially explained by some moderator(s). To explore the influence of moderators on the extent of heterogeneity, further analysis accounting for moderators will be undertaken using mixed-effects meta-regression models (Viechtbauer 2007)[13]. Exploration will not normally include variables where there are fewer than 30 studies in a category.

Sensitivity analysis

To assess potential bias associated with individual out-of-range calculated effect sizes which may potentially distort the overall interpretation of the findings, a sensitivity analysis will be undertaken (Hedges & Olkin, 1985). This is intended to determine whether the removal of a particular effect size

[13] The models for all possible combinations of the available moderators (n) will be fitted (the number of fitted models equals 2^n). Then the set of moderators that explain most of the heterogeneity will be selected based on the heterogeneity statistics to identify the Adjusted I^2. Since the more moderators there are in the model, the smaller the I^2 is likely to be, using I^2 alone tends to choose a model with too many moderators, which may not necessarily be the optimal model. Therefore, the Adjusted I^2 will be used in the model selection process to avoid the possibility of overfitting. This Adjusted I^2 will be obtained from the model I^2, penalised for the number of parameters in the model. The same penalty term $\frac{2P(P+1)}{(N-P-1)}$ from Wagenmakers and Farrell (2004) will be used to derive the corrected Akaike information criterion (AICc) from the Akaike information criterion (AIC) applied to the I^2 statistics to obtain Adjusted I^2. Where P denotes the number of parameters in the model (1 + number of moderators). All studies with missing values on any of the selected moderators will be ignored while fitting the related mixed-effect model (optimal model). The same studies will be excluded in the random effects model (the model without any moderators), to compare its I^2 and that of the optimal model. These analyses will be conducted using the Metafor R package.

increases the fit of the remaining effect sizes in a homogeneous distribution while not substantially affecting the interpretation of the recalculated mean effect size. Various approaches to identifying potential outliers will be used, including visual examination of data organized into forest plots and also performing "one study removed" (Borenstein et al. 2009 - for a more exploratory approach see Baker & Jackson, 2008). Identified outliers will be examined with the potential to remove them from the final dataset. Potential sources of bias, such as study design, type of treatment, publication type, missing data, sample size, or attrition, will be carefully examined through the corresponding moderator variable analyses.

Risk of bias

Risk of bias will be assessed at both study and Toolkit strand levels using a number of categories in the data extraction tools (see above: appraisal of included studies). These study level variables will be included in the moderator analysis. A further analysis will then be undertaken to see to what extent the risk of bias variables explain systematic variation in the pooled effect.

Relying on available and published studies may bias or inflate the overall intervention effect, particularly in education, which has a relatively large proportion of smaller studies. To evaluate potential publication bias across the database, we will review the associations between publication type and the pooled effect (i.e. journal article, dissertation or thesis, technical report, book or book chapter, conference paper, and other) in each Toolkit strand. Thesis completion is not usually influenced by the size of the effect, unlike journal articles, so this can be a useful category to use for comparison.

A further exploration using selection models will be undertaken. This method was initially proposed by Hedges (1984) to adjust for publication bias. It assumes that homogeneous effect sizes and only significant studies are published. Since these assumptions are questionable, especially in education (Higgins, 2018), a further three-parameter variant of Hedges (1984) recommended by McShane, Böckenholt, and Hansen (2016), will be adopted which relaxes these assumptions.

Other methods for assessing small study bias will be explored, such as a visual inspection of a funnel plot or Duval & Tweedie's (2000) trim and fill. Becker (2005) and Banks et al. (2012) recommend the discontinuation of the use of the failsafe N to assess publication bias, as the results are often inconsistent with the results from other publication bias methods (see also Page et al. 2020). This is likely to be because the small-study effect is only one of the reasons among many others for such variation such as heterogeneity in the effects or study-level factors (Marks-Anglin & Chen 2020). In education all of the methods to detect publication bias are problematic due to the negative association between sample size and effect size (e.g. Slavin and Smith, 2009) and the likelihood of 'super-realization bias' (Cronbach et al., 1980) in smaller studies.

Recording and reporting

The Metafor outputs from the analysis of each of the Toolkit strands will inform any revision needed to the Toolkit website[14] ; these revisions will be developed in collaboration with the EEF's Synthesis Team. Any academic publications arising from this work will be shared with EEF before publication.

Personnel

Durham core team: Steve Higgins, Berenice Villanueva Aguilera, Emma Dobson, Louise Gascoine, Afroditi Kalambouka, Maria Katsipataki, Jonathan Reardon, Margherita Rovida, Germaine Uwimpuhwe.

Coding and data extraction team: John Brown, Diwakar Basnet, Ryan Gray, Nicole Wilson, Vicky Byrne and Maria Athanasiou.

Conflicts of interest

There is no conflict of interest in this project, so far as the review team are aware.

Ethical issues

The use of information and data already in the public domain is usually regarded as of low ethical risk. The project team aim to adhere to the ethical guidelines of the British Educational Research Association[15] and ethical approval was obtained from the Ethics Committee of the School of Education at Durham University.

Acknowledgements

The review team would like to acknowledge the help and support from the EPPI-Centre at the Institute of Education, UCL, especially James Thomas, Jeff Brunton, Sergio Graciosi and Ian Shemilt.

The ongoing support and enthusiasm of the Synthesis Team at the EEF is a vital aspect of the continuing evolution of the Toolkit and the development of the EEF Education Database: Jon Kay, Mohammad Zaman, Isabel Kempner and Hannah Shackleton.

In addition, the tireless support of the coding team needs to be acknowledged: John Brown, Diwakar Basnet, Ryan Gray, Nicole Wilson, Vicky Byrne and Maria Athanasiou (as well as members of the team who have contributed to earlier stages of the project).

[14] https://educationendowmentfoundation.org.uk/education-evidence/teaching-learning-toolkit
[15] Ethical Guidelines for Educational Research, fourth edition (2018) available at: https://www.bera.ac.uk/publication/ethical-guidelines-for-educational-research-2018

References

Baker, R., & Jackson, D. (2008). A new approach to outliers in meta-analysis. *Health Care Management Science*, *11*(2), 121-131. https://doi.org/10.1007/s10729-007-9041-8

Banks, G. C., Kepes, S., & Banks, K. P. (2012). Publication bias: The antagonist of meta-analytic reviews and effective policymaking. *Educational Evaluation and Policy Analysis*, *34*(3), 259-277. https://doi.org/10.3102/0162373712446144

Becker, B. J. (2005). Failsafe N or file-drawer number. In H. R. Rothstein, A. J. Sutton, M. Borenstein (Eds) *Publication bias in meta-analysis: Prevention, assessment and adjustments*, 111-125. London: John Wiley and Sons

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*(2), 97-111. https://doi.org/10.1002/jrsm.12

Borenstein M., Hedges L.V., Higgins, J.P.T., Rothstein, H.R. (2009) Subgroup analyses. In: *Introduction to Meta-Analysis*.London: John Wiley & Sons, Ltd, pages 59-86;

Duval, S., & Tweedie, R. (2000). Trim and fill: a simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455-463. https://doi.org/10.1111/j.0006-341X.2000.00455.x

Cronbach, L. J., Ambron, S. R., Dornbusch, S. M., Hess, R. O., Hornik, R. C., Phillips, D. C., Walker, D. F. & Weiner, S. S. (1980) *Toward reform of program evaluation: Aims, methods, and institutional arrangements* (San Francisco, CA, Jossey-Bass).

Elliott, J. H., Synnot, A., Turner, T., Simmonds, M., Akl, E. A., McDonald, S., ... & Pearson, L. (2017). Living systematic review: 1. Introduction—the why, what, when, and how. *Journal of Clinical Epidemiology*, 91, 23-30. https://doi.org/10.1002/9781119536604.ch4

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics*, *9*(1), 61-85. https://doi.org/10.3102/10769986009001061

Hedges, L. V., Shymansky, J. A., & Woodworth, G. (1989). Modern methods of meta-analysis: A Practical Guide. National Science Teachers Association, Washington, DC (ERIC ED309952: https://eric.ed.gov/?id=ED309952

Hedges, L., & Olkin, I. (1985). *Statistical models for meta-analysis*. New York: Academic Press.

Higgins, S. (2018) *Improving Learning: Meta-analysis of Intervention Research in Education* Cambridge: Cambridge University Press.

Higgins, S., Aguilera, A. B. V., Dobson, E., Gascoine, L., Katsipataki, M., & Rajab, T. (2019). *Education Endowment Foundation evidence database: Protocol and analysis plan.* London: Education Endowment Foundation.

Higgins, S., Katsipataki, M. Villanueva Aguilera, A.B., Dobson, E., Gascoine, L., Rajab, T., Reardon, J., Stafford, J. & Uwimpuhwe, G. (2022) The Teaching and Learning Toolkit: Communicating Research

Evidence to Inform Decision-making for Policy and Practice in Education *Review of Education* 10.1 http://dx.doi.org/10.1002/rev3.3327

Higgins, J.P.T., Thompson, S.G., Deeks, J.J. & Altman, D.G. (2003) Measuring inconsistency in meta-analyses. *British Medical Journal*;327:557-560. https://doi.org/10.1136/bmj.327.7414.557

Langan, D., Higgins, J. P., Jackson, D., Bowden, J., Veroniki, A. A., Kontopantelis, E., ... & Simmonds, M. (2019). A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Research Synthesis Methods*, 10(1), 83-98. https://doi.org/10.1002/jrsm.1316

Lefebvre C, Glanville J, Briscoe S, Featherstone R, Littlewood A, Marshall C, Metzendorf M-I, Noel-Storr A, Paynter R, Rader T, Thomas J, Wieland LS. Chapter 4: Searching for and selecting studies. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). (2022) *Cochrane Handbook for Systematic Reviews of Interventions* version 6.3 (updated February 2022). Available from www.training.cochrane.org/handbook.

Lipsey, M. W. & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks CA, US: Sage

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science*, *11*(5), 730-749. https://doi.org/10.1177/1745691616662243

Marks-Anglin, A., & Chen, Y. (2020). A historical review of publication bias. *Research Synthesis Methods*, 11(6), 725-742. https://doi.org/10.1002/jrsm.1452

Page, M. J., Sterne, J. A., Higgins, J. P., & Egger, M. (2021). Investigating and dealing with publication bias and other reporting biases in meta-analyses of health research: A review. *Research Synthesis Methods*, *12*(2), 248-259. https://doi.org/10.1002/jrsm.1468

Schmid, C.H., Stijnen, T. & White, I. (2020) *Handbook of Meta-Analysis*. 1st ed. Chapman; Hall/CRC. https://doi.org/10.1201/9781315119403.

Slavin, R., & Smith, D. (2009). The relationship between sample sizes and effect sizes in systematic reviews in education. *Educational Evaluation and Policy Analysis*, *31*(4), 500-506. https://doi.org/10.3102/0162373709352369

Stansfield, C., Stokes, G., & Thomas, J. (2022). Applying machine classifiers to update searches: Analysis from two case studies. *Research Synthesis Methods*, 13(1), 121-133. https://doi.org/10.1002/jrsm.1537

Thomas, J., Noel-Storr, A., Marshall, I., Wallace, B., McDonald, S., Mavergames, C., ... & Pearson, L. (2017) Living systematic reviews: 2. Combining human and machine effort. *Journal of Clinical epidemiology*, 91, 31-37. https://doi.org/10.1016/j.jclinepi.2017.08.011

Viechtbauer, W. (2005) Bias and Efficiency of Meta-Analytic Variance Estimators in the Random-Effects Model. *Journal of Educational and Behavioral Statistics* 30 (3): 261–93. https://doi.org/10.3102/10769986030003261

Viechtbauer, W. (2007) Accounting for heterogeneity via random-effects models and moderator analyses in meta-analysis. *Zeitschrift für Psychologie/Journal of Psychology*, 215(2), 104-121. https://doi.org/10.1027/0044-3409.215.2.104

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software,* 36(3), 1-48. https://cran.r-project.org/web/packages/metafor/vignettes/metafor.pdf

Wagenmakers, E-J., & Farrell, S. 2004) AIC Model Selection Using Akaike Weights. *Psychonomic Bulletin & Review* 11 (1): 192–96. https://doi.org/10.3758/BF03206482

Xiao Z., Kasim, A., Higgins, S.E. (2016) Same Difference? Understanding Variation in the Estimation of Effect Sizes from Educational Trials *International Journal of Educational Research* 77: 1-14 http://dx.doi.org/10.1016/j.ijer.2016.02.001

The appendices are available as a separate document from EEF or the research team.