Final submitted version of the manuscript, please check with the published version if citing or quoting.
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),. https://doi.org/10.1002/rev3.3327

BERA, Review of Education, Special Issue: Evidence use in Policy and Practice

The Teaching and Learning Toolkit: Communicating Research Evidence to Inform Decision-making for Policy and Practice in Education

Authors: Steve Higgins* [1,2], Maria Katsipataki[1], Alaidde Berenice Villanueva Aguilera[1], Emma Dobson[1], Louise Gascoine[1], Taha Rajab[1], Jonathan Reardon[1], Jade Stafford[1], and Germaine Uwimpuhwe[2]

***Corresponding author: School of Education, Durham University, Durham, DH1 1TA**
s.e.higgins@durham.ac.uk

[1]School of Education, Durham University
[2]Durham Research Methods Centre

Data availability statement
The Education Endowment Foundation's evidence database is continually updated. For access to the current version please contact the EEF. For a copy of the data used in this article please contact the corresponding author.

Funding statement
The original research for the 'Pupil Premium Toolkit' was funded by the Sutton Trust in 2010-11. The Education Endowment Foundation has funded the development of the Teaching and Learning Toolkit over a series of three research contracts with Durham University, running from 2011-2024.

Conflict of interest disclosure
There is no conflict of interest in undertaking this research, so far as the authors are aware.

Ethics approval statement
Ethical approval for each of the phases of the development of the Toolkit was granted by the Ethics Committee in the School of Education at Durham University. In addition the team seek to follow BERA's Ethical Guidelines for Education Research. The Toolkit uses

secondary sources of information and data in the public domain and available through academic libraries (journal articles, chapters in books, reports and dissertations).

Permission to reproduce material from other sources

Not applicable

Final submitted version of the manuscript, please check with the published version if citing or quoting.
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),. https://doi.org/10.1002/rev3.3327

Abstract

This article compares and contrasts two versions of the Education Endowment Foundation's (EEF) Teaching and Learning Toolkit ('Toolkit'), a web-based summary of international evidence on teaching 3-18 year-olds. The Toolkit has localised versions in six different languages in Australia, Cameroon, Chile, Jordan, and Spain. The initial Toolkit, created in 2011 with funding from the Sutton Trust and updated since then with funding from EEF, drew upon over 250 meta-analyses across 30 areas of education research. An updated version, drawing on a database of over 2,500 single studies from these meta-analyses was launched in Autumn 2021. This change was motivated by increased interest in evidence-use in education, and a desire to engage in more rigorous synthesis of primary studies. The article presents the rationale for these changes, outlines the methods adopted to populate and analyse the Toolkit database and presents results from this analysis. Findings indicate that although the broad picture of the relative benefits of the different approaches is similar, a more fine-grained analysis is possible. This deeper synthesis can provide more specific guidance about what has been successful in the different areas of the Toolkit in research studies and offers opportunities for further refinement and improvement. This increased specificity, however, comes at the cost of greater complexity in the findings and the implications for policy and practice, and it increases the challenge of ensuring findings are both accurate and accessible. A final section reflects on the challenges of summarising evidence from research to inform decision-making in education.

Keywords: evidence-based education, Tookit, meta-analysis, attainment, what works

Final submitted version of the manuscript, please check with the published version if citing or quoting.
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),. https://doi.org/10.1002/rev3.3327

The Teaching and Learning Toolkit: Communicating Research Evidence to Inform Decision-making for Policy and Practice in Education

Context & Implications

Rationale for the study

The aim of the research reported in this paper is to improve and accuracy and accessibility of the Education Endowment Foundation's Teaching and Learning Toolkit, which is a synthesis and summary of the impact of various educational approaches on learners' attainment in schools designed to support decisions in schools. An additional aim of this paper is to contrast two different approaches to research synthesis by comparing findings from an earlier reviews of reviews (meta-meta-analysis) with single study meta-analysis.

Why the new findings matter

The findings allow for more comparable analysis across the Toolkit strands and more informative exploration of what drives variation within each strand. This paper also provides a comparison of two approaches to research synthesis. Although there are some differences in the findings and implications, the broad picture is similar.

Implications for policy & practice

This study is relevant for researchers, teachers, practitioners and policy makers in education. The revised Toolkit is more transparent about the path from evidence in research studies to implications for policy and practice. The findings from 30 meta-analyses allow the exploration of patterns of effects occurring within each strand and provide more detailed information for practice, such as relating to effects by school phase or curriculum subject. This more detailed approach can therefore provide more specific indications for practice and and idea of the range of settings in which the approaches have been evaluated. Teachers can make evidence-informed decisions about what might work in their own context. Common inclusion and coding criteria also allow exploration of patterns across strands to identify variation related to methodological and pedagogical characteristics of the included studies.

Final submitted version of the manuscript, please check with the published version if citing or quoting.
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),. https://doi.org/10.1002/rev3.3327

Introduction

Evidence-based or evidence-informed policy and practice has become prevalent over recent decades as educators are encouraged to use and deliver interventions and programmes that research suggests can work to improve educational outcomes (Slavin, 2020). However, there are always challenges associated with using evidence, especially in the field of education. One of the most prominent challenges is to successfully balance the technicalities of research synthesis with dissemination which supports the translation, uptake and embedding of evidence into both policy and practice (Higgins, 2018).

The Teaching and Learning Toolkit exemplifies this tension between academic accuracy and effective uptake and application of research-based approaches. The overarching aim of the Toolkit is to provide accessible evidence-based information for policymakers and practitioners to inform educational decision-making. The first online version of the Toolkit summarised 34 different areas of educational practice by analysing information from over 250 meta-analyses in a meta-meta-synthesis (Higgins, 2018). Its findings were limited, however, by the level of aggregation at the level of the synthesis in each meta-analysis and, whilst the comparative inferences between the different areas were indicative of the relative benefit of different approaches, they were limited by the parameters of the underlying reviews. These issues have been addressed in a revised version of the Toolkit where the individual studies contained within these meta-analyses have been 'unzipped', reviewed and synthesised at study-level. This approach has allowed the application of consistent inclusion criteria and a common method to categorise and classify studies. As a result the comparative inferences between the different areas of the Toolkit are more rigorous, and the analysis of what drives variation within each area is more informative and can include features such as the age of the pupils, the phases of schooling or the subject areas being taught. These changes do not overcome the inherent tensions in evidence-based or evidence informed practice as the additional complexity increases the challenge of making the findings accessible in a way which supports their application in policy or practice.

Conceptual frameworks and rationale

The article draws on a number of frameworks of research use (e.g. Hemsley-Brown & Sharp, 2003; van Schaik, Volman, Admiraal & Schenke, 2018; see also Higgins, 2020) to explain conceptual aspects of the design of the original and updated Toolkit. The overarching rationale is the exploration of a number of tensions in research communication and impact such as the accessibility of evidence, balanced against the accuracy of summaries, and the usefulness of this information in terms of how actionable it is for the user (see Figure 1).

About here: Figure 1: A model of research and practice responsibilities

Some of the responsibilities in the model are from the perspective of the researcher. These involve the research being accessible, accurate and actionable. This immediately sets up a series of tensions for the researcher, represented by the connecting lines in the diagram to summarise findings accurately but succinctly in a way which educational practitioners can understand and put into practice. Accuracy refers mainly to how findings are summarised in relation to what was found (answering the question 'did it work there?' or addressing the internal validity of the study). This tension is also motivated by the aim to support practitioners in England in using the information to make informed decisions as to how and where to spend additional funding. There are numerous factors that can explain the observed difficulties that act as barriers for the dissemination and uptake of educational research (Cherney, Povey, Head, Boreham & Ferguson, 2012). These include the professional culture which influences teachers' attitudes and perceptions, time pressures and other commitments, access to relevant research and the perceived relevance and quality of that research (van Schaik et al., 2018). This sets up a series of responsibilities that can be understood as the responsibility of the user in terms of how applicable the research is to a different context, how appropriate to pupils' needs and how acceptable the research is in terms of values. Although it is beyond the scope of this paper to elaborate on all these challenges, the development of both versions of the Toolkit have taken into consideration many of these factors and have paid special attention to the presentation and communication of findings, with teachers and educational practitioners in mind. The creation of an accessible and "user-friendly" source of educational knowledge formed a large part of the rationale for this work, as did the introduction of the Pupil Premium in England which shaped the original form of the Toolkit. A description of the policy context in England and the Pupil Premium to support

the attainment of disadvantaged students, as well as the wider international context of an increasing focus on attainment outcomes (Slavin, 2020), are set out to indicate some of the drivers and tensions for evidence use in education in England. A summary of growth in Toolkit usage over the last decade will also be provided. This increased interest in evidence-based research among educators has further contributed to the need for rigorous research about school-based interventions. The following sections will provide a summary of some of the different aspects of the context of this research. First, some background on the Pupil Premium will be presented, then the introduction of the What Works Centres in the UK and then the wider international context.

Policy context in England

The Pupil Premium was first introduced in England in 2011 with the aim of providing additional school funding to raise the attainment of disadvantaged students (Foster & Long, 2017). Children eligible for free school meals (FSM) as well as looked after children were allocated additional funding to help raise their attainment. Schools were subsequently expected to publish their strategy for using this funding on their websites. This encouraged schools to justify their decisions and an increasing proportion drew on the evidence in the Toolkit. The amount of money allocated for each child has been increased since 2011, from about £450 to about £1,000 per pupil per year. A number of studies have investigated the impact of the Pupil Premium with the majority concluding that since the Pupil Premium was established attainment levels among the disadvantaged had significantly improved, even in the most challenging areas (Gorard, Siddiqui & Huat See, 2019 & Gorard, Siddiqui & Huat See, 2021). These findings suggest that the Pupil Premium initiative is working and should be retained. Thus the potential value of an evidence-based, accessible educational resource for schools such as the Toolkit is clear and further explains the current project's rationale. The following section presents a brief summary of the What Works Centres, another important element of the context in England.

What Works Centres

Evidence-based (or evidence-informed) education emerged from the need to establish a knowledge base where practitioners, policy makers, teachers and other related professionals could find guidance and advice on effective methods (often characterised as "what works") that could be used in the classroom to improve various educational outcomes (see Nelson &

Campbell, 2017 for a discussion of terminology). Attempts can be identified in a number of countries including the USA, the United Kingdom and Australia. For example, the USA introduced a legislation in 1998 stating that federal funding would only be available for programmes that have demonstrated evidence of their effectiveness (see Hempenstall, (2006), with further discussion of the global context below). The same year in the UK, a National Literacy Strategy was introduced indicating that school practice should be based on reliable and good quality research findings. Then in 2013 a UK government initiative created the "What Works Centres" (What Works Network, 2014). These comprised of a network of nine centres covering areas such as, education, crime, homelessness, well-being and health. The Education Endowment Foundation (EEF) was designated by the government as the What Works Centre for improving educational attainment in schools on the basis of the Toolkit (Higgins, 2020). This growing policy focus has also been reflected in an increased interest from practitioners and a growth in the commercialisation of this field (Menter, 2021). Organisations like ResearchEd, the appointment of research leads in schools, and the growth in professional publications with research and evidence in the title indicate corresponding a shift in interest from practitioners (Cain, 2018).

The wider international context of evidence in education

This development of a research and evidence-based resource for schools needs to be seen in the wider international context of evidence in education and in public policy more widely (Donaldson, 2009). This shift has been evident not only in the UK but also in Australia, New Zealand, the USA, Denmark , Norway and other countries (Hanne & Rieper, 2009; Petersen & Reimer, 2014; Slavin, 2002). In the USA, the "What Works ClearingHouse" (WWC) produced by the Institute for Education Sciences (IES) provides educators with research on different programmes to inform their decisions (https://ies.ed.gov/ncee/wwc/FWW). Another evidence-based resource drawing on research in the USA and the UK is the Best Evidence Encyclopaedia (BEE). The BEE has been created by Bob Slavin at John Hopkins University and provides summaries of programmes that are currently available to educators (https://bestevidence.org/). A full account of the development of clearinghouses and evidence repositories around the world is beyond the scope of this paper, but see Mayo Wilson, Grant and Supplee (2021) for a recent account of the development of clearinghouses in the USA and Hanne & Rieper (2009) for an account of some of the methodological developments associated with the growth of the evidence movement, especially in Europe. In

Australia the National Enquiry into the Teaching of Literacy in 2005 emphasised the role that evidence-based research can play in education (Hempenstall, 2006). The international context undoubtedly influences the underlying policy and practice culture for the Toolkit and the wider evidence-based or evidence-informed debates and developments support the expansion of systematic reviews and evidence production methodologies and approaches (Gamoran, 2018).

The Pupil Premium Toolkit and the EEF's Teaching and Learning Toolkit

Having previously reviewed the extent of evidence available in meta-analyses of intervention findings in education as part of an ESRC Researcher Development Initiative, we were initially approached by the Sutton Trust to develop a series of summaries which could help schools decide how to allocate any additional funding for the new Pupil Premium policy (Higgins, Kokotsaki & Coe, 2011). We developed these as a series of related "umbrella reviews" (Grant & Booth, 2009) which would provide a rigorous but accessible summary of the quantitative evidence with a common methodology across the different strands (Ioannidis, 2009). The Pupil Premium Toolkit was published in May 2011. The feedback from both policy and practice audiences convinced us that this was worth developing further and the Toolkit was adopted in 2011 by the newly formed EEF. The Foundation was established by Impetus and the Sutton Trust with an £125 million endowment from the Department for Education. The Toolkit became the focus for research synthesis, as the EEF commissioned large scale trials in schools to identify approaches to improving outcomes for disadvantaged pupils in schools. The first online version of the Toolkit was launched in 2013, with annual revisions and updates until 2019.

Growth in Toolkit usage 2012 – 2020

The policy context in England had a direct influence on awareness and use of the Toolkit. It also enabled us to track increasing interest and access to the information. To evaluate the Pupil Premium and better understand how schools were using the funding the Department of Education commissioned a number of studies. These usually included surveys asking teachers to select what tools they use to infrom their funding spending decisions. In 2013 Carpenter and colleagues found that from the 1,240 schools involved in the study a total of 124 were

using the Toolkit as their main resource when deciding how to spend the Pupil Premium (Carpenter et al., 2013). Additional data from the National Audit Office (NAO) in 2015 suggested that 64% of schools in England had consulted the Toolkit (National Audit Office, 2015). These figures showed an increase from earlier figures in 2012 where 36% of schools indicated using the Toolkit (NAO, 2015). Data regarding the use of the Toolkit was also tracked by a Sutton Trust survey conducted annually by the National Foundation for Education Research (NFER). This survey focused on a sample of about 1,500 teachers included questions on the pupil premium spending and decision-making. The survey indicates that the percentage of school leaders reporting that they consulted the Toolkit when making decisions about the pupil premium increased from 11% in 2012 to 69% in 2021 (see Figure 1).

About here: Figure 2: Percentage of school leaders in England reporting consulting the Teaching and Learning Toolkit as part of their Pupil Premium strategy.

As the chart indicates, there has been an increase in the Toolkit's usage over the years, suggesting both a shift in teachers' attitudes towards consulting evidence-based research to inform their decisions, and offering some indication of the value of accessible information. The uptake of the Toolkit was no doubt influenced by the requirement for schools to report on their websites how they were spending the pupil premium, but it seems likely that headteachers saw some value in this resource. We do not know, of course, how well the evidence is used and whether schools in England cite the Toolkit to justify decisions that have already been made or whether the information is used to identify evidence-based solutions to the educational challenges that they face.

EEF's International partners and localised versions of the Toolkit

Following the Toolkit's growing popularity and its widespread use in England, interest in the development of similar approaches has become evident not only in education but other areas of public life as well (Higgins, 2020). The Toolkit structure is based around a summary page of approaches, with links to successive layers of detail on each of these approaches. This has become characteristic of a number of What Works centre websites in the UK. This increase in popularity has also led to its replication and development in other countries such as Australia, Chile, Spain Jordan and Cameroon. The EEF worked with the education community in Australia and formed a partnership with Social Ventures Australia (SVA)in 2014. SVA

formed Evidence for Learning as an independent educational organisation with the support of the Commonwealth Bank taking the global evidence from the Toolkit and contextualising it with local research to make it relevant to the Australian context (https://evidenceforlearning.org.au/the-toolkits/). This became the pattern for the development of a number of other international partnerships. In 2017, the EEF collaborated with Fundación Chile and SUMMA to create the 'Effective Education Practices Platform'. This website synthesizes the global evidence from the Toolkit with quality evidence and academic research on school-level educational interventions for Latin America and the Caribbean in Spanish, Portuguese and English, (https://www.summaedu.org/plataforma-de-practicas-educativas-efectivas/?lang=en). In 2018 the EEF entered a partnership with EduCaixa to develop a translation for Spanish and Catalan versions (https://educaixa.org/es/home). In 2019 the EEF formed another partnership with Effective Basic Services Africa (eBase) which led to the adaptation of the Toolkit for teachers in Cameroon, Nigeria, Chad and Niger and will also include a French translation. Another international partner was added in 2019 with the Queen Rania Foundation supporting the translation of the Toolkit into Arabic for the Middle East and North African regions (https://www.qrf.org/en). See Figure 3 for the locations of EEF's international partners with local versions of the Toolkit. The darker colour represents the host country for the Toolkit with their regional network indicated in the ligher shade. In each instance the local partner, is independent of the government in each jurisdiction and is usually supported by charitable donations, has created a contextualised version of the Toolkit with relevant local research in each region and translated into other languages as necessary.

About here: Figure 3: EEF's global partnerships with local versions of the Toolkit

The interaction, contributions and constructive feedback provided by these international partners has provided valuable insights for our research and our understanding of research communication and impact. The independence from government is seen as an essential component for building trust with potential users of the evidence. The importance of local contextualisation is also acknowledged, as well as the emphasis on comparative information detailing the effectiveness of different approaches to support local decision-making. It is far from clear in the evidence how widely findings from education research might apply in different contexts or jurisidictions.

Development of an education evidence database: methods

This section of the paper describes the methods underpinning the two versions of the Toolkit, and sets out the rationale for the changes so as to provide a basis for the comparison of findings that follows**.** The methods for the first version of the Tookit are set out in more detail in earlier publications, so only a brief summary is included here (see Higgins, Kokotsaki & Coe, 2011; Higgins & Katsipataki, 2016; Higgins, 2016; Higgins, 2018; EEF, 2018 for further details). The methods for the updated version are set out in more detail below in terms of the systematic application of inclusion criteria, data extraction from individual studies (using the EPPI-Reviewer software) and the approach to synthesis using meta-analysis as these methods underpin the warrant or the claims for the findings.

The Sutton Trust- Education Endowment Foundation Teaching and Learning Toolkit

The Sutton Trust wanted to engage with the developing policy context in 2010 for the Pupil Premium in England which was mooted as support for the educational attainment of children from disadvantaged backgrounds. The aim of the review was to provide an overview, akin to a 'Which?' guide to summarise the cost/benefit of different educational approaches to provoke discussion and to support schools in how to allocate the new funding. The review was able to draw on a database of meta-analyses that had been developed to support the teaching of meta-analysis for social sciences, funded by the ESRC. The selection of the topics covered was made based on three criteria: first, approaches that were commonly mentioned in connection with education policy; second, schools' suggestions for how additional resource could be used; and third, approaches with a strong evidence of effectiveness not covered by either previous criterion (Higgins, Kokotsaki & Coe, 2011). The main aspects of the original Toolkit to communicate our findings were: potential gain presented in the form of an effect size and also translated into months progress, cost estimates, applicability and evidence strength (Higgins, & Katsipataki, 2016). When the Toolkit was first developed it covered 21 themes or strands. The translation of the effect size (standardised mean difference) to months progress was motivated by the need for an easy to understand measure. It is problematic for a number of reasons, not least of which is the variation by age (for further discussion of this see Higgins, 2018). The fledgling EEF adopted the Toolkit and funded its translation into a website in 2013. Since then, repeated systematic searches have been undertaken for

systematic reviews with quantitative data and meta-analyses of educational interventions in the relevant Toolkit areas. This first online version of the Teaching and Learning Toolkit (2013-2019) eventually comprised 34 topics, along with an Early Years version with 12 separate strands. The online presentation of the Toolkit provides both a brief summary overview of all of the topics on the main page as well as a more comprehensive presentation for each strand for those who are interested in additional information about each topic (EEF, 2018). A further successive level of detail provided technical information, including effect sizes and references to the meta-analyses and systematic reviews for transparency. Toolkit impact estimates were based on the quantitative syntheses in these meta-analyses, combined as a basic fixed effect average, making the Toolkit a "review of reviews" or 'super-synthesis'(Higgins, 2016). This has a number of important limitations: first, the lack of granularity that is available for the findings for each of the areas of the Toolkit; second, the synthesis is at the level of the review so patterns in features of individual studies cannot easily be identified; third, moderators in the individual meta-analyses differ according to the questions for the different reviews and the availability of data so are not comparable across reviews; fourth, the inclusion criteria for each meta-analysis differ so do not provide a consistent set of underlying studies; fifth, the meta-analyses provided uneven coverage on the underlying literature, with differing time periods with some overlapping (and therefore risking some studies being double counted) and some missing particular years or even decades. In sum, the quality of the reviews in each area of the Toolkit determines the quality of the synthesis, rather than the quality of the individual studies. These limitations in terms of the accuracy and applicability of the information led to the decision to create a database of single studies for a more consistent and comprehensive analysis.

'Unzipping' the meta-analyses

The meta-analyses used in the online version of the EEF-Sutton Trust Teaching and Learning Toolkit in 2018 served as the base for the EEF Evidence Database. These meta-analyses and systematic reviews have been identified through a systematic updating process (EEF, 2018) since the initial version of the Toolkit was published by the Sutton Trust in 2011 (Higgins, Kokotsaki & Coe, 2011). A new search for reviews was conducted in 2020 and studies from these syntheses were also included.

References to studies included in these meta-analyses were systematically 'unzipped' so that each of the included studies which contributed to the overall pooled effect were identified and screened (a two-stage process of title and abstract and then full text screening) for inclusion

in the database. A flow diagram describing this process can be found in the Supplementary materials S1: Study identification flow diagram.

Inclusion criteria for the EEF Evidence Database

The inclusion criteria aim to identify relevant educational evidence for schools and policy makers interested in school-based education, consistent with the mission of the EEF, which is dedicated to breaking the link between family income and educational achievement. Specifically, the EEF aims to:

- raise the attainment of 3-18 year-olds, particularly those facing disadvantage;
- develop their essential life skills; and
- prepare young people for the world of work and further study.

A PICOS and SPIDER analysis (Methley et al., 2014) were used to define the database scope (see Supplementary materials S2) which was then used to refine more specific inclusion and exclusion criteria (see Table 1).

About here: Table 1: Database inclusion and exclusion criteria

Search strategy for identification of relevant single studies

Where there were no existing meta-analyses or systematic reviews with quantitative data covering the existing Toolkit strands, a new systematic search was undertaken for primary studies to update the existing single studies identified for the Toolkit. This included the following Toolkit strands: Aspiration intervention, Teaching assistants and School Uniforms. These sources were used (gateways and databases): First search (Article First, ECO, Papers First, World Cat Dissertations); EBSCO (BEI, Education Abstracts, Education Administration Abstracts, ERIC, PsycArticles, PsycINFO); Taylor and Francis (Educational Research Abstracts Online); ProQuest (ProQuest Dissertations and theses (Global)); Elsevier (Science Direct); Thomson Reuters (Web of Science).

In addition, informal searching for 'grey' literature (reports and unpublished studies) was undertaken using Google, Google Scholar and Microsoft Academic. Our approach does not use citation searching, 'pearl growing' (Schlosser et al., 2006) or expert nomination, though we used these techniques at an exploratory stage to ensure the adequacy of search terms (Papaioannou, 2010). Our rationale for this is that the use of such approaches on their own, without subsequently adapting the search criteria in the light of what is found are likely to increase the risk of publication bias (Higgins, 2018). If we identified includable studies from non-systematic approaches we then refined our search criteria and ran additional searches to find other similar studies retrieved with the amended search strings which could then be

applied systematically. More details about this approach and the search strings used can be found in the database protocol (Higgins et al. 2019), Following the systematic search and study retrieval, the next stage was study screening. This took place in three stages described in more detail below.

Title and abstract screening

Screening was conducted using EPPI Reviewer, systematic review software developed by the EPPI-Centre at the Social Science Research Unit of the UCL Institute of Education (Thomas et al., 2020). Having this tool was central to manage our references, to store the document files of the reports of the studies and to facilitate analyses such as mapping the studies and conducting a meta-analysis. At the initial stage of screening, each title and abstract was reviewed based on the inclusion criteria (please see Table 1 & Supplementary Materials S2). At this stage, if there was uncertainty a study would be included for full text screening.

Study retrieval

Once the studies had been screened at the title and abstract level then those included were retrieved as documents and uploaded into EPPI Reviewer. If a study was not available through Durham's library access it would be marked for "manual search" to see if it was available online. If not it would be requested as an interlibrary loan. This was more common with older articles, dissertations and book chapters. If the item could not be retrieved from the library we tried to contact the author(s) as the final step. Ultimately, in cases where the study could not be retrieved it was marked as 'excluded/not retrieved'.

Full text screening

Once the full texts had been retrieved, each study was reviewed for the final decision to ensure that it met all of the inclusion criteria .Once all the studies had been screened a 10% sample was assigned to a second reviewer for blind double coding (agreement at this stage was typically above 95%). EPPI-Reviewer facilitates this process by recording the potential agreements and disagreements and through the reconciliation function for two reviewers to resolve any disagreements. Upon completion of the reconciliation stage the remaining studies were then allocated to a trained team of coders for data extraction.

The following adapted PRISMA diagram (Figure 4) provides an overview of the studies included in the database (Page et al., 2021: For more information, see: http://www.prisma-statement.org/).

About here: Figure 4: Adapted PRISMA diagram

Description of methods used in the included studies

The inclusion criteria aimed to identify studies with a valid counterfactual comparison between those receiving the educational intervention or approach and those not receiving it. True experimental (randomised) and quasi-experimental studies (both prospective and retrospective) designs were included when they featured two educational conditions addressing the central theme of each Toolkit strand (e.g. peer tutoring compared with no peer tutoring or studies contrasting reduced class size with normal or usual sized classes). Other designs such as interrupted time series or regression discontinuity were included where they similarly provided an estimate of the effect of the intervention or approach by comparing the attainment of different groups of pupils. Design features were coded to allow for exploratory analysis. The different counterfactual conditions included were: an active control (i.e. there is control for novelty such as with another introduced new intervention or 'treatment'); business as usual (i.e. comparison group having their usual learning experience); no equivalent teaching (i.e. additional learning time, where the control or comparison group have no typical educational experience, such as in a Summer School intervention or a Before or After school club).

Identifying the best single outcome from a study is not always straightforward as the study aims are not necessarily the same as the Toolkit aims. Three key principles were adopted to support outcome identification. First, a good test of the impact of the intervention for the Toolkit synthesis. The main issue to consider here was the alignment of the study with the Toolkit in terms of the research design and the research questions. We needed the best estimate of the difference between pupils experiencing the intervention or approach with the most appropriate counterfactual condition (those not experiencing the intervention or approach). Second, an appropriate measure of educational attainment. This created the challenge of identifying which specific curriculum or cognitive outcome was most appropriate. In general, the focus is on outcomes which are good indicators of overall educational attainment, such as standardised tests of reading comprehension or mathematics. These are also good predictors of subsequent educational attainment. Standardised tests or national tests and examinations tend to be better overall indicators of educational performance than researcher-designed measures or teacher-designed class tests (see Sammons et al., 1995; Tymms, 1999) as these can inflate effect size estimates (Martin &

Shapiro, 2011; Ainsworth et al., 2015). Third, we were looking for as direct and fair a measure as possible. Single outcomes rather than ones combined across subjects are usually preferable (so reading *or* mathematics rather an overall score that combines both). This is not always straightforward. In a pedagogical intervention where the focus is on general strategies and is taught across several curriculum subjects it can be difficult to decide which is the best outcome for the Toolkit. Peer-tutoring delivered in reading and mathematics may have one designated as the primary and another as the secondary outcome or they may be combined and the average reported. It may be appropriate to combine them when they are equally valid possible outcomes. In this case the separate scores for each subject were also recorded so that subject specific analyses could be conducted.

Another issue to consider was the distinction between treatment inherent and treatment independent measures (e.g. Slavin & Madden, 2011). In practice they can be hard to separate. Criterion-referenced measures can be particularly problematic here. In a spaced-learning intervention in history, for example, a school history knowledge test is only fair if the control group were also taught the same topic in history. Another example might be a phonics intervention where the intervention group are taught letter sounds and compared with a business-as-usual control. Here a letter recognition test may not provide a fair comparison as it is likely to over-estimate the impact on reading (as opposed to impact on letter recognition). In fact, such a measure might be a better measure of implementation fidelity. On the other hand, evaluating the impact of teaching number fact recall with a standardised test of mathematics may similarly under-estimate effects if number forms only a limited part of the standardised test.

These considerations resulted in the development of a flow diagram to aid coders in identifying the primary outcome (see Supplementary material S3) where the aim is to identify in each study the most comparable effect size for each Toolkit strand, but which also takes into account the nature of the particular intervention in the study. Additional secondary outcomes (such as alternative measures of attainment), or equivalent measures in different subjects (where applicable for the intervention) were also identified and extracted.

Statistical independence of findings from a single study

There are a number of threats to the validity of findings in a meta-analysis related to statistical dependence (Scammacca, Roberts & Stuebing, 2014). These are the use of data from the same participants for different outcomes; reporting multiple outcomes of the same type; and aggregating outcomes of different types for the same sample of participants. We

selected one primary outcome for the Toolkit strand from each study or independent comparison in a study report on the basis of a pre-specified protocol. Other equivalent academic and cognitive outcomes were recorded as secondary outcomes. Where it was not possible to identify a single preferred outcome (such as a reading intervention where a standardised test of reading comprehension is not reported), comparable outcomes may be combined to produce one overall effect for the study (such as word reading, reading fluency or decoding skills). There is a case for extracting all outcomes from a study and running a multivariate model (van den Noortgate et al., 2015) but we chose not to do this for three main reasons. First, we wanted to select as common an outcome as possible across the Toolkit strands (usually reading comprehension assessed using a standardised measure); second, it can be difficult to judge whether all outcomes are an equally valid measure of the effect of the intervention (some measures may be used to check for fidelity or used as placebo controls); and third, some studies may have used a large number of outcomes and we wanted to control the costs of the project for the sponsor.

Data extraction: details of study coding categories and quality assurance procedures

Study coding was undertaken with three data extraction tools: EEF main data extraction (v 1.0 June 2019), used for all studies (Supplementary materials S4); EEF Toolkit effect size data extraction (v 1.0 June 2019), used for all studies, which extracts the effect sizes for attainment from each study along with information to categorise this (see Supplementary materials S5); and Strand specific information: this is a set of additional codes for each Toolkit strand, such as information about tutors and tutees in Peer tutoring, or group size in Small group tuition – used for studies in each Toolkit strand (see Supplementary materials S6). These first two data extraction tools are now available in EPPI-Reviewer as public codesets.

The main data extraction tool was developed based on a comparison of available and relevant alternative coding frameworks (e.g. EPPI Centre Education guidelines (version 0.97/2003), Lipsey and Wilson (2001), the Institute for Education Sciences What Works Centre Study Review Guides (https://ies.ed.gov/ncee/wwc/StudyReviewGuide), and the coding developed by 3iE (http://www.3ieimpact.org/en/). We did not adopt a specific, separate quality appraisal or risk of bias tool as the evidence is limited about the validity of these tools (in medicine at least: Hartling et al., 2009; Katikireddi et al., 2009) and the choice of tool has a direct impact on the outcomes of a meta-analysis (Voss & Rehfuess, 2013). We included aspects of study quality (such as design, randomisation and attrition) so that relationship

between particular features of study quality and risk of bias could be explored at the analysis stage.

Demographic features include learners' age, socio-economic background and attainment level, as well as subject matter studied. Substantive features across studies were used to explore variation in terms of pedagogical differences such as treatment duration, provision of professional development for teachers and training for students, or involvement of parents or digital technology, depending on approach. These study features were analysed as moderators for their potential relationship with outcome effects. Methodological quality in each Toolkit strand was assessed using features such as design, the unit of assignment/analysis, attrition reported and method of effect size estimation (Cooper, Hedges, & Valentine, 2009).

All coding and data extraction activities (i.e., abstract screening, full-text review, study features coding, as well as effect size extraction) were carried out by a team of reviewers, each working independently but discussing and resolving queries, and eliciting a third opinion from the core project team when necessary. All coders received training and had to achieve an agreed level of reliability to be included in the coding team. A 10% sample of studies (per coder and per strand) were double coded to assess reliability rates, and an additional data checking and cleaning process was used as a further means to ensure reliability.

Statistical procedures and conventions

The database aims to include and summarize quantifiable school attainment outcomes from primary empirical studies which meet the inclusion criteria and match the Toolkit themes. The key metric used is the Standardised Mean Difference (d-index) or effect size. It should be noted that the use of effect sizes is controversial in education (Simpson, 2018), though the consensus still appears to support their use (Kraft, 2020), particularly when studies adopt similar designs (either experimental or correlational) and compare similar outcomes (such as attainment or attitudes). Whilst they are not ideal, they do provide a metric that can be used across studies using similar but different measures (Higgins, 2018). For studies that report descriptive statistics for continuous measures of pupil attainment outcomes, the post-intervention mean of the control group was subtracted from the post-intervention mean of the intervention group and the resulting difference divided by the pooled standard deviation, adjusted for sample size (Hedges' g). An accompanying standard error was also recorded or calculated which is used to weight the study in the meta-analysis. (It should be noted that this

confidence interval is mathematically related to a p-value threshold of 0.05, such that if the confidence interval in a study includes the zero line, a study would not be considered statistically significant at the 95% level. In practice few educational studies meet all of the requirements for inference to a broader population, particularly that of random sampling. One of the goals of meta-analysis is to avoid some of the issues associated with this practice (Higgins, 2018). If the same group of participants was used more than once (such as the same control group compared with two different treatment groups) the sample size and associated standard error were adjusted so that the study contributed fairly to the overall average (Moeyaert et al., 2017). Where ever possible the descriptive outcome statistics (N, means and standard deviations for control and intervention groups) were collected, even where the study reported an effect size and accompanying standard error, or where an effect size could be calculated from other inferential statistics, so that the effect size could be checked.

All effect sizes were coded either as resulting from a post-test or gain comparison. These effect sizes may sometimes need to be meta-analysed separately as they may represent different metrics (such as when the intervention affects the relative spread of the intervention group (Xiao et al., 2017)). For studies where there was substantial baseline imbalance, a gain score effect size may be preferred (such as in quasi-experimental designs or natural experiments). Chance imbalance is likely to occur in randomized studies (the smaller the study the greater the risk) and can usually be dealt with through an analysis which takes account of baseline measurements. Theoretically, if the sampling of randomized studies in a meta-analysis is unbiased, any imbalance is likely to even out with a large number of studies.

Outcome data were reported in a wide variety of formats. Where possible effect sizes were recalculated from means and standard deviations or other descriptive statistics. For studies that report inferential statistics only such as t, F, or precise p-values, an appropriate conversion formula was applied to calculate the *d*-index as the effect size estimate (Lipsey & Wilson, 2001; Hedges, Shymansky, & Woodworth, 1989; Hedges & Olkin, 1985). These are not always ideal as there are additional distributional assumptions in these conversions which may not always be met and which may introduce error. To ensure appropriate corrections for the small sample size bias, all *d*-indices were converted to the unbiased Hedges' *g* statistic. Wilson's (n.d.) online calculator on the Campbell Collaboration website was the most frequently used conversion tool, though other online resources were identified and bespoke Excel resources were also developed to support the study data extraction team.

This review focuses on academic attainment outcomes. In some studies there were several measures of the same or of similar outcomes from the same sample of learners. When this happened, we selected the most representative measure according to a pre-specified protocol (please see S3 Effect size flow diagram).

Approach to synthesis

Upon completion of data extraction of the studies and a process of data checking and cleaning for each strand, a dataframe was created from the EPPI-Reviewer database. Impact estimates from appropriate analyses in the study reports contributing to a specific Toolkit strand were further synthesised into a single pooled effect using a random effects meta-analysis adopting a restricted maximum likelihood (REML) estimation method following Viechtbauer's (2005) and Langan and colleagues' (2019) recommendations. Analysis was undertaken with the 'R' package 'Metafor' (Viechtbauer, 2010). Wherever relevant, sub-group analyses were also undertaken using a similar approach. The I^2 statistic was used to assess the heterogeneity (Schmid, Stijnen, & White 2020). It is possible that the estimated variability may not only be due to random variation but partially explained by some moderator(s) (Viechtbauer, 2007). To explore the influence of moderators on the amount of heterogeneity, further analysis accounting for moderators were undertaken using mixed-effects meta-regression models and R^2 calculated (Viechtbauer, 2007).

The models for all possible combinations of the available moderators were fitted (number of fitted models equals 2^n). Then, the set of moderators that explain most of the heterogeneity was selected based on the heterogeneity statistics. Using I^2 alone tends to choose the model with too many moderators, which may not necessarily be the optimal model. Therefore, the adjusted I^2 (I^{2d}) was used in the model selection process to avoid the possibility of overfitting. All studies with missing values on any of the selected moderators were ignored while fitting the related mixed-effect model (optimal model). The same studies were excluded in the random effects model and the model without any moderators, to compare its I^2 and that of the optimal model.

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 was undertaken (Hedges & Olkin, 1985). This was to determine whether the removal of a

particular effect size 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 were used, including visual examination of data organized into a forest plots and also performing "one study removed" (Borenstein et al., 2000 - for a more exploratory approach see Baker & Jackson, 2008). Identified outliers were examined with the potential to remove them from the final dataset. Potential sources of bias, such as study design, type of treatment, publication source, missing data, sample size, or attrition, were examined through the corresponding moderator variable analyses.

Publication Bias

Relying on available and published studies may bias or inflate the overall intervention effect, particularly in education, a field with a relatively large proportion of smaller studies. To evaluate potential publication bias across the database, we reviewed the association between publication type and the pooled effect (i.e. journal article, dissertation or thesis, technical report, book or book chapter, conference paper, and other). Thesis completion is not usually influenced by the size of the effect, unlike journal articles and other publications, so this can provide a benchmark for comparison.

Other methods for assessing publication bias were explored, such as a visual inspection of the funnel plots and Duval & Tweedie's (2000) trim and fill routine (Borenstein et al., 2005). Becker (2005) and Banks et al., (2012), however, 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. In education, all of the methods to detect publication bias are problematic due to the small but consistent negative association between sample size and effect size (e.g. Slavin & Smith, 2009) which may relate to the increased quality of implementation possible at smaller scale, sometimes known as 'super-realisation' bias (Cronbach et al., 1980: see also Bell, 2011).

Findings

Both versions of the EEF Toolkit were designed to provide accessible summaries of educational research for teachers and policy makers. They present over a range of approaches to improving learning and teaching, each summarised in terms of its average impact on attainment, its cost and the strength of the evidence supporting it.

The previous version of the Toolkit was a meta-meta-analysis, based on findings from existing meta-analyses. In contrast, the new Toolkit is based on data from single studies. This change allows for more comparable analysis across the Toolkit and more informative exploration of what drives variation within each strand. This variation occurs both at a pedagogical and a methodological level. The former includes variables such as curriculum subject, age of pupils, intervention duration and intensity and the latter includes design features such sample size or outcome and measurement artefacts. This was more challenging and time-consuming to undertake than with the first version of the Tookit.

However, the fact that each of the single studies retrieved were screened against consistent inclusion criteria adds to the transparency and consistency of the Toolkit. In the original version we applied our inclusion criteria to a meta-analyses as a whole, but each meta-analysis had its own inclusion criteria for the specific research questions for each review. This was potentially problematic. By applying the new single study approach to the database the studies included are consistent between Toolkit strands and more comparable both within and between strands.

Impact estimates

We re-estimated the pooled effect for each Toolkit strand based on this new database using the meta-analytic approaches outlined above. A series of variables or 'moderators' were explored consistently across the Toolkit strands as well as additional variables specific to each strand. For example the analysis includes the effects of the different approaches for different pupil age groups and different subject outcomes. For the Peer tutoring strand we looked at the effect of different types of peer tutoring, such as same age, cross age and reciprocal. This more detailed approach can provide more specific recommendations for practice and much better idea of whether an approach that has worked in a particular setting might also work in a different context.

In this section a number of examples are provided, showing the changes between the pooled effect sizes and months' progress from the current Toolkit and the database analysis, such as for Teaching assistant interventions and Feedback. Some areas yielded very similar results (such as Collaborative learning and Small group tuition). A new feature is the 'null' rating, assigned to an area or strand where analysis of individual studies indicate that a pooled effect size is not appropriate for synthesis (such as Aspiration interventions where there were too few studies for a meta-analysis: https://educationendowmentfoundation.org.uk/education-evidence/teaching-learning-toolkit/aspiration-interventions). General patterns can also be

identified, such as a tendency for the pooled effects strands with fewer studies to increase slightly, partly as a result of moving from a fixed effect to a random effects meta-analysis. Table 2 provides a summary of the strands included in the Toolkit comparing the previous versus the current effect sizes, as well as the months gain and the difference between the two Toolkits. This provides an overview of the differences as well as the similarities resulting from the two approaches to research synthesis.

About here: Table 2: Comparison between versions of the Toolkit

The current version consists of 30 strands instead of 34 in the earlier version. Homework was originally split into primary and secondary and others have been removed for the time being and will potentially be included at a later point (e.g. the Built environment). Digital technology has been included in the analysis of individual strands where this is possible. Comparing the effect sizes between versions of Toolkit, ten strands have increased effect sizes, ten have remained the same and five have decreased. There are also three strands where the evidence was insufficient to calculate an overall pooled effect so were assigned a 'null' effect size.

The biggest change is for the Teaching assistant strand where there is an increase of an additional three months progress compared to one month in the previous version. The biggest decrease is for the Feedback strand where it has dropped from eight to six months. The greater granularity allows a look behind this average to see that studies involving spoken feedback tend to have greater impact than written, seven months as opposed to five[1]. The fact that the differences between the Toolkits in months gain were generally between +/-1 month was reassuring, suggesting that the previous version Toolkit was a reasonable approximation, given the limitations of the approach.

Security

In the previous version of the Toolkit the security or 'padlock' rating was based on the number and quality of the underpinning meta-analyses. Features of the meta-analyses evaluated were the ecological validity (such whether studies were undertaken in schools or realistic educational settings), the recency of the meta-analysis and the rigour of the analysis

[1] https://educationendowmentfoundation.org.uk/education-evidence/teaching-learning-toolkit/feedback

(such as whether heterogeneity and publication bias were reported) and whether the meta-analysis conducted a security or risk of bias assessment.

In the current, single study version of the Toolkit the 'padlock' rating is based on the number of included studies in a strand, the proportion of studies that were randomised, the proportion of studies that were independently evaluated, the proportion of studies published since 2000, the extent of attrition in the included studies, the extent of heterogeneity in the meta-analysis and the extent to which this variation is explained by moderators. These indicators are combined to provide an overall estimate of overall security. These features are scaled across the Toolkit strands from 1 to 5. Strands without a quantitative synthesis or pooled effect have a null estimate of impact and are rated zero padlocks. We have always believed it important to summarise evidence even when it is inconclusive or and to report where evidence is lacking as we believe this information is useful for decision-making.

Although the padlock rating approach appears superficially similar, the approaches cannot be directly compared. Overall there has been an increase in the comparability of the evidence summarised for each Toolkit strand, and a more rigorous analyses of what drives variation in each area. The estimates remain approximate and it is still the case that the variation within a Toolkit strand is greater that the variation between strands. However the comparative inferences or 'best bets' about what has 'worked' in research studies are more informative and the moderator analysis provides information about features of the interventions which are associated with smaller and larger effects.

Heterogeneity

Educational meta-analyses typically have high heterogeneity (Higgins, 2018), resulting in part from variations in the context and settings for the studies, as well as the measures used for evaluation and other aspects of the research design and operationalisation. In addition the interventions and approaches themselves tend to be broad categories that are loosely classified and framed (Bernstein, 1973), such as 'peer tutoring' or 'performance pay schemes' (unless a meta-analysis has been undertake of a specific programme or defined intervention). Whilst high heterogeneity is often seen as an issue indicating a lack of suitability for statistical combination in medical fields (Imrey, 2020), in education the focus tends to be on the extent to which this variation can be explained in the analysis (Borenstein et al., 2009). Even then, if a field has a recognisable set of descriptors which are widely used and accepted, then it still seem to be a reasonable undertaking to look at the overall effects of

an approach, answering the questions of the type "Do summer schools tend to have beneficial effects on attainment, on average?". Higher heterogeneity is to be expected in studies involving human interaction as opposed to the tighter classification and framing of drug formulations and trials where the variables can be both defined and controlled more precisely (Higgins, 2018).

The Toolkit operates at a broad level of aggregation with terminology that would be recognised by practitioners and policy makers. The aim is to draw comparative inferences about approaches the teachers recognise to offer 'best bets' or indeed approaches which, on average, have tended to be unsuccessful (such as getting students to repeat a year of their schooling). We therefore expected heterogeneity to be high but hoped that the moderator analysis would explain at least a proportion of this variation. Where this variation related to the pedagogical aspects of the approach, such as the ages of pupils, or the subject which was the focus or even the duration (in weeks) or frequency (in number of days per week) the aim was to refine the guidance on the Toolkit website. Table 3, below contains an overview of the heterogeneity for each of the strands with the R^2 for some of the key moderators.

About here: Table 3: Toolkit strand heterogeneity

As expected overall heterogeneity was high across all of the Toolkit strands, ranging from 72% to 99%. (I^2 is the proportion of the observed variance that reflects the variance in true effect sizes, rather due to than sampling error.) In all cases some variables were identified, either from the common coding categories, or the strand specific variables. This facilitated writing a new section for each Toolkit webpage entitled "Behind the average". A number of features were fairly consistently associated with variation in impact, such as sample size and test type. Studies with smaller samples tended to have larger effect sizes and researcher designed and teachers' tests also tending to have larger effects. These correlations were noted by Slavin & Smith (2008). Overall the extent of the variation associated with these features was small, however. The median R^2 for sample size was about 3%, smaller than Slavin and Smith's finding 7.8% (a correlation of -0.28), but of a similar order. The variation across strands suggests that at least some of this relates to the pedagogical features of the approaches, and perhaps superrealization, as mentioned above. Some design features were not consistently associated with the direction of the effect, such as randomised studies

compared with quasi-experimental designs or setting (US and non-US) where examples of positive and negative associations could be found, though this needs further exploration.

Discussion and conclusions

This shift in the granularity of the use of research evidence in our research was driven by the need to improve the rigour and transparency of our work and to provide more specific implications for practice. Although awareness of research evidence is growing, this is still not easily translated into daily school practice (Hornby et al., 2013). Improving the specificity of the recommendations is one avenue to support greater uptake. This comes at a cost, however in terms of the simplicity of the message as features which explain variation make the summaries more complex. The accessibility of findings from research synthesis tends to be in tension with their applicability.

In identifying the relative benefit of the different educational approaches, the previous version of the Toolkit (as a meta-meta-review or 'super-synthesis') provided a broadly similar overall picture, especially in terms of the patterns of effects and the relative benefit of different approaches, but did not allow for variation within each strand to be explored consistently. This suggests that a review of reviews or a rapid evidence assessment (Varker et al., 2015), based on existing meta-analyses, should give a reasonable overview of the evidence in a particular field. It may not provide sufficiently detailed information in terms of recommendations for practice, and it may not be able to provide an assessment of the robustness of the underlying evidence, unless this is included in the underpinning reviews and is sufficiently comprehensive and consistent to allow comparison and synthesis.

In terms of communicating impact it is clear that practitioners prefer a broad estimate in a familiar metric like months progress (Lortie-Forgues et al., 2021), though this is problematic technically (as mentioned above). In terms of the model introduced at the beginning of the article, this reflects the tension between accuracy and accessibility. However, given the variability within and between strands, this kind of broad estimate may not be such a bad thing compared with a standardised mean difference to two decimal places, which perhaps has a level of spurious accuracy in terms of interpreting the impact of a range of interventions and approaches clustered under a broad educational category or heading (Higgins, 2018).

Generating robust research and synthesis is only one part of the equation, while supporting dissemination, engagement and successful implementation is another: as EEF does with its guidance reports and other activities such as their research schools network. We believe that the nature of evidence in education is such that it will always require some interpretation and contextualisation by those who use it. Evidence is about what has worked in the past and in other contexts, rather than about what will work for a particular school. Even successful interventions may only benefit a relatively small proportion of the intervention group. An effect size of 0.4 is equivalent to about 15% of the intervention group making greater progress than that made by the controls (Uwimpuhwe et al., 2020). Pupils (and teachers) do not all respond to interventions in the same way, so the choices made by schools and teachers about what they think is appropriate to meet their pupils' needs and applicable to their context and experience (as well as acceptable in terms of professional values) is likely to remain central to evidence use in education for the foreseeable future. Using evidence of effectiveness does not guarantee that all pupils will benefit. It may be a 'good bet' but the impact needs monitoring in a new context. External validity in education remains problematic.

Reflections on the challenges of evidence-use in education

The technical improvements to the Toolkit do not necessarily improve evidence-use in education. In fact, the addition of analysis which explores variation within and between the Toolkit strands increases the complexity of the findings and, as a result, potentially reduces the clarity and accessibility of the summaries. It certainly increases the challenge for the user to identify findings which might be applicable to their individual context (see Figure 1, above). The more fine-grained the analysis, the more specific the target audience is likely to be. This indicates the importance of the partnership between the research producer or summariser and the research user in so that each recognises the roles and responsibilities of the other in developing their understanding of their own roles and responsibilities. Creating more complex or more nuanced research summaries will only be successful if this synthesis meets the needs and capabilities of those who might benefit from using them to support their own decision-making in either policy or practice.

We have also not addressed the underlying issues about the use of effect sizes to compare educational interventions. We now have, however, a database which will allow further exploration of this issue. Although the conversion from standardised mean difference to

months progress has its limitations it does maintain an approximation for the impact which may be more appropriate than the apparent precision of an effect size to two decimal places. The extent of the impact may help in identifying how good a bet it is, but knowing more precisely how well an approach has worked 'on average' does not help in identifying whether the approach will be successful in a particular new context.

Researchers sometimes assume that variation in the impact of interventions is related to variation in implementation (e.g. O'Donnell, 2008), so have focused on fidelity of implementation as the solution to improving the evidence-use. However it is not clear yet that we understand what explains variation in impact in terms of the characteristics of the sample, the research design and measurement, before we assume that the differences result from pedagogical variation or differences in what teachers do. Our search for external validity in education is also hampered by lack of random selection (Gorard, 2014) and replication (Mackel & Plucker, 2008). The database can only confirm that these are relevant issues in education research.

Finally, the Toolkit depends on the quality of the underlying studies that it summarises and whilst we can explore to what extent features of study quality are associated with the effects on learning we are limited in identifying a useful 'signal' from the noise that is inevitable in the evidence base.

Future plans for the database

This phase of the development of the EEF Toolkit was a necessary step in developing the overall rigour and comparability of the evidence from experimental research in education. The previous version had reached the limits of what was possible with a meta-meta-review. The next step is to identify potentially missing studies as the meta-analyses we 'unzipped' may not have been comprehensive in their coverage. This should then enable the Toolkit to become a 'living systematic review' (Elliott et al., 2017) which can be updated as more studies become available. It would be valuable for international partners to include studies published in languages other than English, a current limitation.

There is also a host of other technical possibilities, these include multivariate meta-analysis and network approaches, Bayesian meta-analysis (Borenstein et al., 2021, and more systematic identification of what explains variation in the effect sizes which may allow causal modelling of underlying pedagogical mechanisms (such as the particular teaching approach or the intensity of use). In addition, adding new Toolkit strands will be easier and it should

also be possible to add additional outcomes (such as attendance or well-being), although this task becomes increasingly challenging as the number of studies in the Toolkit increases. It should also be possible to undertake meta-analyses of specific programmes or approaches within a Toolkit strand to explore the variation in impact in more detail.

Finally, one of the challenges of research synthesis in education is the range of approaches across the different centres and clearninghouses and the growing number of reviews and meta-analyses that are available, making it difficult for researchers, policy makers and practitioners to keep up to date. Our ambition is that data on individual studies and topics could be exchanged between researchers and clearninghouses, if a common data structure could be agreed to support this.

Acknowledgements

The review team at Durham University would like to acknowledge the initial grant from the Sutton Trust and the ongoing funding from the Education Endowment Foundation which has supported the development of the Toolkit over the last ten years, in addition the support from the EEF team over the years, colleagues at the EPPI-Centre at the UCL Institute of Education, and our team of part-time coders without whom we could not have completed the development of the database.

List of supplementary materials

S1 Study identification flow diagram

S2 PICOS and Spider analysis

S3 Flow diagram for selecting the primary effect size

S4 Main data extraction tool

S5 Effect size data extraction tool

S6 Sample strand specific data extraction tool

References

Ainsworth, H., Hewitt, C. E., Higgins, S., Wiggins, A., Torgerson, D. J., & Torgerson, C. J. (2015). Sources of bias in outcome assessment in randomised controlled trials: a case study. *Educational Research and Evaluation*, *21*(1), 3-14 https://doi.org/10.1080/13803611.2014.985316

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.

Bell, J. F. (2011). The small-study effect in educational trials. *Effective Education*, *3*(1), 35-48. https://doi.org/10.1080/19415532.2011.610642

Bernstein, B. (1973). On the classification and framing of educational knowledge. In B. Brown (ed) *Knowledge, education, and cultural change* (pp. 365-392). London: Routledge.

Best Evidence Encyclopaedia, Center for Research and Reform in Education. John Hopkins University. https://bestevidence.org

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, pp. 59-86.

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

Borenstein, M., Hedges, L.V., Higgins, J., & Rothstein, H. (2021). *Introduction to Meta-Analysis.* (Second edition) London: John Wiley & Sons, Ltd.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5), 365. https://doi.org/10.1038/nrn3475

Cain, T. (ed) (2018). *Becoming a Research-Informed School* (pp. 3-16). London: Routledge.

Carpenter, H., Papps, I., Bragg, J., Dyson, A., Harris, D., Kerr, K., Todd, L., & Laing, K. (2013). *Evaluation of pupil premium*, London: Department for Education.

Cherney, A., Povey, J., Head, B., Boreham, P., & Ferguson, M. (2012). What influences the utilisation of educational research by policy-makers and practitioners?: The perspectives of

academic educational researchers. *International Journal of Educational Research*, *56*, 23-34. https://doi.org/10.1016/j.ijer.2012.08.001

Connolly, P., Keenan, C., & Urbanska, K. (2018). The trials of evidence-based practice in education: a systematic review of randomised controlled trials in education research 1980–2016. *Educational Research*, 60(3), 276-291. https://doi.org/10.1080/00131881.2018.1493353

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: Jossey-Bass.

Deeks, J.J., Douglas, A.G., Bradburn, M.J. (2001) Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis. In: Egger M, Davey Smith G; Altman DG *Systematic Reviews in Health Care: Meta-analysis in Context*. London: BMJ Publishing Group.

Donaldson, S. I. (2009). In search of the blueprint for an evidence-based global society. In S.I Donaldson, C.A. Christie & M.M. Mark *What counts as credible evidence in applied research and evaluation practice*, 2-18. London: Sage.

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

EEF (2018) *Sutton Trust-EEF Teaching and Learning Toolkit & EEF Early Years Toolkit Technical appendix and process manual* (Working document v.01) July 2018 London: Education Endowment Foundation. https://educationendowmentfoundation.org.uk/public/files/Toolkit/Toolkit_Manual_2018.pdf

Education Endowment Foundation (2019). *The EEF guide to Pupil Premium*. https://educationendowmentfoundation.org.uk/public/files/Publications/Pupil_Premium_Guidance_iPDF.pdf

Elliott, J. H., Turner, T., Clavisi, O., Thomas, J., Higgins, J. P., Mavergames, C., & Gruen, R. L. (2014). Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap. *PLoS medicine*, 11(2) https://doi.org/10.1371/journal.pmed.1001603

Foster, D., & Long, R. (2017). *The Pupil Premium. Briefing Paper*, Number 6700, House of Commons Library. London: House of Commons.

Gamoran, A. (2018). Evidence-based policy in the real world: A cautionary view. *The Annals of the American Academy of Political and Social Science*, 678(1), 180-191. https://doi.org/10.1177/0002716218770138

Gorard, S. (2014). The widespread abuse of statistics by researchers: What is the problem and what is the ethical way forward?. *Psychology of Education Review*, 38(1), 3-10.

Gorard, S., Siddiqui, N., & See, B. H. (2019). The difficulties of judging what difference the Pupil Premium has made to school intakes and outcomes in England. *Research Papers in Education*, 36(3), 355-379. https://doi.org/10.1080/02671522.2019.1677759

Gorard, S., Siddiqui, N., & See, B. H. (2021). Assessing the impact of Pupil Premium funding on primary school segregation and attainment. *Research Papers in Education*, 1-28. https://doi.org/10.1080/02671522.2021.1907775

Grant, M. J., and A. Booth. (2009). "A Typology of Reviews: An Analysis of 14 Review Types and Associated Methodologies." *Health Information and Libraries Journal* 26 (2): 91–108. https://doi.org/10.1111/j.1471-1842.2009.00848.x

Hanne, H. F., & Rieper, O. (2009). The evidence movement: the development and consequences of methodologies in review practices. *Evaluation*, *15*(2), 141-163. https://doi.org/10.1177/1356389008101968

Hartling, L., Ospina, M., Liang, Y., Dryden, D. M., Hooton, N., Seida, J. K., & Klassen, T. P. (2009). Risk of bias versus quality assessment of randomised controlled trials: cross sectional study. *British Medical Journal*, 339, b4012. https://doi.org/10.1136/bmj.b4012

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

Hempenstall, K. (2006). What does evidence-based practice in education mean?. *Australian Journal of Learning Difficulties*, 11(2), 83-92. https://doi.org/10.1080/19404150609546811

Hemsley-Brown, J., & Sharp, C. (2003). The use of research to improve professional practice: A systematic review of the literature. *Oxford Review of Education*, 29(4), 449-471. https://doi.org/10.1080/0305498032000153025

Higgins, J.P.T. & Thompson, S.G. (2002) Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*; 21:1539-1558. https://doi.org/10.1002/sim.1186

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

Higgins, S., Kokotsaki, D. & Coe, R. (2011) *Toolkit of Strategies to Improve Learning: Summary for Schools Spending the Pupil Premium: Technical Appendices* London: Sutton Trust.

Higgins, S., Katsipataki, M., Kokotsaki, D., Coleman, R., Major, L.E. & Coe, R. (2013). *The Sutton Trust - Education Endowment Foundation Teaching and Learning Toolkit.* London, Education Endowment Foundation.

Higgins, S. (2016) Meta-synthesis and comparative meta-analysis of education research findings: some risks and benefits *Review of Education* 4.1: 31–53. http://dx.doi.org/10.1002/rev3.3067

Higgins, S. & Katsipataki, M. (2016) Communicating comparative findings from meta-analysis in educational research: some examples and suggestions. *International Journal of Research & Method in Education* 39.3: 237-254 https://dx.doi.org/10.1080/1743727X.2016.1166486

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 https://educationendowmentfoundation.org.uk/public/files/Toolkit/EEF_Evidence_Database_Protocol_and_Analysis_Plan_June2019.pdf

Higgins, S. (2020). The development and worldwide impact of the Teaching and Learning Toolkit. In S. Gorard, (2020). *Getting Evidence into Education. Evaluating the Routes to Policy & Practice*. (pp. 69-83). London: Routledge.

Hornby, G., Gable, R. A., & Evans, W. (2013). Implementing evidence-based practice in education: What international literature reviews tell us and what they don't. *Preventing School Failure: Alternative Education for Children and Youth*, 57(3), 119-123. https://doi.org/10.1080/1045988X.2013.794326

Imrey, P. B. (2020). Limitations of meta-analyses of studies with high heterogeneity. *JAMA network open*, 3(1), e1919325. https://dx.doi:10.1001/jamanetworkopen.2019.19325

Ioannidis, J. P. 2009. "Integration of Evidence from Multiple Meta-Analyses: A Primer on Umbrella Reviews, Treatment Networks and Multiple Treatments Meta-Analyses." *Canadian Medical Association Journal* 181 (8): 488–493. https://doi.org/10. 1503/cmaj.060410 .

Katikireddi, S. V., Egan, M., & Petticrew, M. (2015). How do systematic reviews incorporate risk of bias assessments into the synthesis of evidence? A methodological study. *Journal of Epidemiology and Community Health*, 69(2), 189-195. http://dx.doi.org/10.1136/jech-2014-204711

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, 49(4), 241-253. https://doi.org/10.3102/0013189X20912798

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: a diagnosis based on the correlation between effect size and sample size. *PloS one*, 9(9), e105825. https://doi.org/10.1371/journal.pone.0105825

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

Lortie-Forgues, H., Sio, U. N., & Inglis, M. (2021). How should educational effects be communicated to teachers?. *Educational Researcher*, 50 (6), pp. 345–354 http://doi.org/10.3102/0013189X20987856

Lumley, T. (2002). Network meta-analysis for indirect treatment comparisons. *Statistics in Medicine*, 21(16), 2313-2324. https://doi.org/10.1002/sim.1201

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher*, 43(6), 304-316. https://doi.org/10.3102/0013189X14545513

Martin, S. D., & Shapiro, E. S. (2011). Examining the accuracy of teachers' judgments of DIBELS performance. *Psychology in the Schools*, *48*(4), 343-356. https://doi.org/10.1002/pits.20558

Menter I. (2021) Snake oil or hard struggle? Research to address the reality of social injustice in education. In: Ross A. (eds) *Educational Research for Social Justice. Education Science, Evidence, and the Public Good*, vol 1. Springer, Cham. https://doi.org/10.1007/978-3-030-62572-6_2

Methley, A. M., Campbell, S., Chew-Graham, C., McNally, R., & Cheraghi-Sohi, S. (2014). PICO, PICOS and SPIDER: a comparison study of specificity and sensitivity in three search tools for qualitative systematic reviews. *BMC Health Services Research*, 14(1), 579. https://doi.org/10.1186/s12913-014-0579-0

Moeyaert, M., Ugille, M., Natasha Beretvas, S., Ferron, J., Bunuan, R., & Van den Noortgate, W. (2017). Methods for dealing with multiple outcomes in meta-analysis: a comparison between averaging effect sizes, robust variance estimation and multilevel meta-analysis. *International Journal of Social Research Methodology*, 20(6), 559-572. https://doi.org/10.1080/13645579.2016.1252189

National Audit Office (2015). Funding for Disadvantaged Pupils London: National Audit Office. https://www.nao.org.uk/wp-content/uploads/2015/06/Funding-for-disadvantaged-pupils.pdf

Nelson, J. & Campbell, C. (2017) Evidence-informed practice in education: meanings and applications, Educational Research, 59:2, 127-135, http://doi.org/10.1080/00131881.2017.1314115

O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., & Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic Reviews*, 4(1), 5. https://doi.org/10.1186/2046-4053-4-5

Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C. Mulrow, C.D. et al.(2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *British Medical Journal* 2021;372:n71. http://doi.org/10.1136/bmj.n71

Papaioannou, D., Sutton, A., Carroll, C., Booth, A., & Wong, R. (2010). Literature searching for social science systematic reviews: consideration of a range of search techniques. *Health Information & Libraries Journal*, 27(2), 114-122. https://doi.org/10.1186/2046-4053-4-5

Petersen, K. B., & Reimer, D. (2014). *Evidence and Evidence-based Education in Denmark. The Current Debate*. CURSIV, 14. Denmark: Department of Education, Aarhus University.

Sammons, P., Nuttall, D., Cuttance, P., & Thomas, S. (1995). Continuity of school effects: A longitudinal analysis of primary and secondary school effects on GCSE performance. *School Effectiveness and School Improvement*, 6(4), 285-307. https://doi.org/10.1080/0924345950060401

Scammacca, N., Roberts, G., & Stuebing, K. K. (2014). Meta-analysis with complex research designs: Dealing with dependence from multiple measures and multiple group comparisons. *Review of Educational Research*, 84(3), 328-364. https://doi.org/10.3102/0034654313500826

Schlosser, R. W., Wendt, O., Bhavnani, S., & Nail-Chiwetalu, B. (2006). Use of information-seeking strategies for developing systematic reviews and engaging in evidence-based practice: the application of traditional and comprehensive Pearl Growing. A review. *International Journal of Language & Communication Disorders*, 41(5), 567-582. https://doi.org/10.1080/13682820600742190 .

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

Simpson, A. (2018). Princesses are bigger than elephants: Effect size as a category error in evidence-based education. *British Educational Research Journal*, 44(5), 897-913. https://doi.org/10.1002/berj.3474

Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher*, 31(7), 15-21. https://doi.org/10.3102/0013189X031007015

Slavin, R. E. (2020). How evidence-based reform will transform research and practice in education. *Educational Psychologist*, 55(1), 21-31. https://doi.org/10.1080/00461520.2019.1611432

Slavin, R., & Madden, N. A. (2011). Measures inherent to treatments in program effectiveness reviews. *Journal of Research on Educational Effectiveness*, 4(4), 370-380. https://doi.org/10.1080/19345747.2011.558986

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

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, 9(4), 1-19. http://PAREonline.net/getvn.asp?v=9&n=4 .

Teacher Polling 2017, Sutton Trust, April 2017. https://www.suttontrust.com/wp-content/uploads/2019/12/Pupil-Premium-Polling-2017-data-1.pdf

Thomas, J., Graziosi, S., Brunton, J., Ghouze, Z., O'Driscoll, P., & Bond, M. (2020). *EPPI-Reviewer: advanced software for systematic reviews, maps and evidence synthesis.* EPPI-Centre Software. London: UCL Social Research Institute.

Tymms, P. (1999). Baseline assessment, value-added and the prediction of reading. Journal of Research in Reading, 22(1), 27-36. https://doi.org/10.1111/1467-9817.00066

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2015). Meta-analysis of multiple outcomes: a multilevel approach. *Behavior Research Methods*, 47(4), 1274-1294. https://doi.org/10.3758/s13428-014-0527-2

Van Schaik, P., Volman, M., Admiraal, W., & Schenke, W. (2018). Barriers and conditions for teachers' utilisation of academic knowledge. *International Journal of Educational Research*, 90, 50-63. https://doi.org/10.1016/j.ijer.2018.05.003

Varker, T., Forbes, D., Dell, L., Weston, A., Merlin, T., Hodson, S., & O'Donnell, M. (2015). Rapid evidence assessment: increasing the transparency of an emerging methodology. *Journal of Evaluation in Clinical Practice*, *21*(6), 1199-1204. https://doi.org/10.1111/jep.12405

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3), 1-48. https://doi.org/10.18637/jss.v036.i03

Viechtbauer, W. 2007. "Accounting for Heterogeneity via Random-Effects Models and Moderator Analyses in Meta-Analysis." *Zeitschrift für Psychologie/Journal of Psychology* 215 (2): 104–21. https ://doi.org/10.1027/0044-3409.215.2.104 https://doi.org/10.3102/10769986030003261

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

Voss, P. H., & Rehfuess, E. A. (2013). Quality appraisal in systematic reviews of public health interventions: an empirical study on the impact of choice of tool on meta-analysis. *Journal of Epidemiology and Community Health*, 67(1), 98-104. http://dx.doi.org/10.1136/jech-2011-200940

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

Wallace, B. C., Trikalinos, T. A., Lau, J., Brodley, C., & Schmid, C. H. (2010). Semi-automated screening of biomedical citations for systematic reviews. *BMC bioinformatics*, 11(1), 55. https://doi.org/10.1186/1471-2105-11-55

What Works Network. (2014). *What Works? Evidence for decision makers*. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/378038/What_works_evidence_for_decision_makers.pdf

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

Wilson, D. B., Ph.D. (n.d.). *Practical Meta-Analysis Effect Size Calculator* [Online calculator]. Retrieved 26th August, 2021, from https://campbellcollaboration.org/research-resources/effect-size-calculator.html

Xiao, Z., Higgins, S., & Kasim, A. (2017). An Empirical Unraveling of Lord's Paradox. *The Journal of Experimental Education*, 1-16. https://doi.org/10.1080/00220973.2017.1380591

Inclusion criteria	**Excluded**
The majority of the sample (>50%) on which the analysis is based are learners or pupils aged between 3-18 (further education or junior college students are being included where their study is for school level qualifications).	The majority of the sample are: post-secondary education; in higher education; adults; infants under 3; other students over 18; SEN students only taught in specialist SEN settings. Studies of ESL students only.
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.	Intervention or approach is not classifiable applicable to the current Toolkit strand definition.
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 setting (e.g. an outdoor field centre or museum).	Laboratory studies Specially created environments (both physical and virtual) designed for theoretical research questions, rather than educational benefit.
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.
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.	Attitudinal, affective or motivational outcomes.
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.

Table 1: Database inclusion and exclusion criteria

Final submitted version of the manuscript, please check with the published version if citing or quoting.
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),. https://doi.org/10.1002/rev3.3327

	ES meta-synth	Months meta-synth	ES Db	Months Db	ES diff	Months diff
Arts participation	0.15	+2	0.25	+ 3	+0.10	+1
Aspiration interventions	0.00	0	Null	Null	-	-
Behaviour interventions	0.25	+3	0.28	+ 4	+0.03	+1
Collaborative learning approaches	0.38	+5	0.45	+ 5	+0.07	0
Extending school time	0.11	+2	0.24	+ 3	+0.13	+1
Feedback	0.63	+8	0.48	+ 6	-0.15	-2
Homework	0.10/0.44[2]	+2/+5	0.34	+ 5	-	-
Individualised instruction	0.19	+3	0.27	+ 4	+0.08	+1
Learning styles	0.13	+2	Null	Null	-	-
Mastery learning	0.40	+5	0.45	+ 5	+0.05	0
Mentoring	0.00	0	0.13	+ 2	0.13	+2
Metacognition and self-regulation	0.54	+7	0.58	+ 7	+0.04	0
One to one tuition	0.37	+5	0.41	+ 5	+0.04	0
Oral language interventions	0.37	+5	0.49	+ 6	+0.12	-1
Outdoor adventure learning	0.31	+4	Null	Null	-	-
Parental engagement	0.22	+3	0.34	+ 4	+0.12	+1
Peer tutoring	0.37	+5	0.42	+ 5	+0.05	0
Performance pay	0.04	+1	0.07	+ 1	+0.03	0
Phonics	0.35	+4	0.42	+ 5	+0.07	+1
Physical activity[3]	(0.17)	(+2)	0.08	+ 1	-	-
Reading comprehension strategies	0.45	+6	0.53	+ 6	+0.08	0
Reducing class size	0.19	+3	0.17	+ 2	-0.02	-1
Repeating a year	-0.32	-4	-0.44	- 5	-0.12	-1
School uniform	0.01	0	Null	Null	-	-

[2] Previously split into primary and secondary school effects.

[3] Previously 'Sports participation' with an ES of 0.17, this is a subset of the new category.

Final submitted version of the manuscript, please check with the published version if citing or quoting.
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),. https://doi.org/10.1002/rev3.3327

Setting and streaming	**-0.08**	**-1**	**0.03**	**+ 0**	**+0.11**	**+1**
Small group tuition	**0.21**	**+4**	**0.28**	**+ 4**	**+0.07**	**0**
Social and emotional learning	**0.28**	**+4**	**0.34**	**+ 4**	**+0.06**	**0**
Summer schools	**0.18**	**+2**	**0.25**	**+ 3**	**0.07**	**+1**
Teaching assistant interventions	**0.08**	**+1**	**0.35**	**+ 4**	**+0.27**	**+3**
Within class attainment grouping	**0.21**	**+3**	**0.17**	**+ 2**	**-0.04**	**-1**

Table 2: Comparison between versions of the Toolkit

Final submitted version of the manuscript, please check with the published version if citing or quoting.
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),. https://doi.org/10.1002/rev3.3327

Toolkit strand	No. of studies	Initial I^2	Final I^2	Reduction	R^2	Sample size	Country	Date	ES type	Freq	Rand	Test type	Subject
Arts participation	80	92%	82%	10%	12%	0.00%	0.00%	0.00%	-	0.00%	0.00%	-	1.48%
Behaviour interventions	89	91%	65%	25%	57%	28.10%	-	0.00%	10.40%	1.00%	0.00%	0.00%	-
Collaborative learning	212	97%	94%	3%	19%	1.54%	8.65%	0.00%	0.80%	0.00%	0.53%	6.95%	0.00%
Extending school time	74	98%	86%	12%	12%	2.86%	0.00%	0.00%	0.00%	0.00%	0.00%	0.68%	-
Feedback	155	98%	61%	30%	50%	3.14%	0.00%	2.61%	0.00%	0.00%	0.00%	5.13%	0.00%
Homework	43	93%	76%	17%	37%	0.00%	13.61%	-	0.50%	0.00%	1.65%	0.00%	-
Individualised instruction	198	95%	89%	6%	13%	2.74%	0.17%	0.00%	0.00%	-	0.00%	4.68%	4.00%
Mastery learning	80	92%	85%	7%	40%	18.34%	-	-	5.67%	0.00%	0.00%	6.07%	0.00%
Mentoring	44	90%	68%	22%	37%	0.00%	-	0.00%	0.00%	7.12%	0.93%	0.00%	0.00%
Metacognition and SRL	246	99%	93%	7%	12%	8.27%	0.00%	0.00%	0.00%	0.00%	0.00%	6.10%	5.56%
One to one tuition	123	93%	89%	4%	5%	2.39%	2.30%	0.69%	0.18%	0.03%	-	0.00%	-
Oral language ints.	154	92%	82%	10%	21%	2.93%	7.94%	3.18%	0.29%	0.00%	0.41%	6.30%	-
Parental engagement	97	84%	64%	19%	21%	12.13%	0.00%	5.65%	0.39%	0.53%	0.28%	-	3.22%
Peer tutoring	127	78%	43%	35%	47%	14.12%	2.71%	2.49%	4.75%	2.31%	0.00%	0.00%	0.00%
Phonics	121	87%	75%	12%	39%	13.63%	0.00%	0.00%	0.00%	5.72%	0.00%	9.99%	N/A
Physical activity	61	85%	68%	17%	15%	10.31%	0.00%	-	1.51%	0.00%	4.20%	0.00%	-
Reading comp. strategies	141	89%	77%	13%	29%	4.53%	0.00%	11.73%	0.00%	0.00%	4.96%	0.18%	-
Reducing class size	45	86%	82%	5%	12%	-	-	0.00%	-	N/A	0.00%	1.70%	-
Repeating a year	100	99%	98%	1%	19%	10.20%	-	-	0.00%	N/A	-	-	0.00%
School uniform	7	80%	67%	13%	0%	0.00%		62.00%	-	N/A	-	-	-
Setting or streaming	58	92%	90%	2%	0%	0.00%	-	-	-	N/A	-	-	-
Small group tuition	62	95%	84%	11%	18%	1.37%	3.92%	-	1.94%	0.00%	-	1.75%	-
Social & emotional learning	54	99%	94%	4%	11%	-	-	3.50%	4.37%	0.00%	-	0.005	4.91%
Summer schools	59	80%	67%	13%	0%	7.74%	-	0.70%	-	N/A	0.00%	-	-
Teaching assistants	65	91%	44%	48%	84%	59.69%	2.03%	-	3.85%	0.00%	0.00%	-	-
Within class att grp	23	99%	51%	48%	97%	0.00%	12.00%	-	-	N/A	0.00%	42.40%	44.20%

Table 3: Toolkit strand heterogeneity

Final submitted version of the manuscript, please check with the published version if citing or quoting.
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),. https://doi.org/10.1002/rev3.3327