

Comparative meta-analysis and evidence from research in education

Steve Higgins and Maria Katsipataki

School of Education, Durham University, Durham, UK

Contact details

Professor Steven Higgins

School of Education

Durham University

Leazes Road

Durham

DH1 1TA

s.e.higgins@durham.ac.uk

Tel: 0191 334 8359 (Durham)

0191 334 8310 (Switchboard)

Biographical notes

Steve Higgins is a professor of education and is the lead author of the Sutton Trust-Education Endowment Foundation Teaching and Learning Toolkit. Maria Katsipataki is a researcher working on the search, retrieval and analysis of evidence for this project. Both are based at Durham University, Durham, UK.

Acknowledgements

The Sutton Trust – Education Endowment Foundation (EEF) Teaching and Learning Toolkit has been developed with funding from both the Sutton Trust and the EEF. We are grateful in acknowledging their support for the work on which this article is based.

Communicating comparative findings from meta-analysis in educational research: some examples and suggestions

Abstract

This article reviews some of the strengths and limitations of the comparative use of meta-analysis findings, using examples from the Sutton Trust-Education Endowment Foundation (EEF) Teaching and Learning ‘Toolkit’ which summarises a range of educational approaches to improve pupil attainment in schools. This comparative use of quantitative findings has similar characteristics to umbrella reviews which provide a succinct summary of the current state of evidence to inform practice or policy. Meta-analysis helps to identify which approaches have, on average, made the most difference to tested learning outcomes, in terms of effect size. We suggest that any comparative inferences made should be treated cautiously, but taken seriously. Additionally, we present alternative ways of interpreting effect sizes, security ratings and cost-estimates to make research findings accessible, whilst retaining appropriate accuracy which is discussed using the ‘Toolkit’ as an example. We conclude by arguing that we should consider the available information not as ‘what works’, but ‘what has worked’ to understand its value and limits in terms of supporting the development of research-based practice.

Keywords: evidence-based practice; comparative meta-analysis, ‘what works’, research synthesis

Introduction

One of the challenges in developing evidence-based teaching in schools is to identify what evidence is likely to be useful in a specific context. Individual research studies may have potentially valuable findings, but there are countless studies which might potentially be of value. In this article we present the case for the value of comparative evidence, drawing on meta-analysis, and outline some of the thinking behind the Sutton Trust-Education Endowment Foundation Teaching and Learning Toolkit, which is now consulted by 64% of schools in England, according to the National Audit Office (NAO, 2015). Meta-analysis is used to synthesise or pool results of similar studies to identify, overall, what the cumulative effects are in a particular field, such as typical gender differences in school achievement from correlational studies (Voyer & Voyer, 2014) or the impact of phonics on reading (Torgerson, Brooks and Hall, 2006). In this paper we

are considering results across different studies with a common population in order to provide general or comparative inferences between various areas of educational research which have been used to, or are thought to, improve educational outcomes for children and young people. Our argument here is that teachers and schools need information about the typical relative effects of different approaches to improve learning, with information about the costs of different approaches, as well as information about the nature of that evidence to inform professional decisions about the adoption, evaluation and development and of research-based approaches in schools. We outline the role of meta-analysis as a tool to synthesise findings from intervention research in education, then look at the challenges in interpreting findings across meta-analyses and the tension between presenting messages which are accessible and actionable, but also accurate in terms of the underlying research. This presents a series of challenges in understanding how to apply such findings as the synthesis indicates that this does not provide a guarantee of what will work in relation to a particular practice, but a guide which can be understood as indicating which approaches are likely to be successful and which may require more exceptional effort to have a positive impact on pupils' learning. Furthermore, we discuss different ways of presenting findings from meta-analysis helping educational research to become more accessible and actionable.

Meta-analysis in educational research

Synthesising and summarising research in any field is challenging. In educational research the scope, scale and sheer diversity of research makes this challenge even greater. Over the years a number of techniques have been developed, from narrative literature reviews to systematic reviews, best-evidence syntheses and realist approaches. Sometimes the techniques have been borrowed from other fields; at other times they have been developed to address particular educational challenges. One example of the latter is the development of 'meta-analysis' which, although it has found a home in medical and clinical research, was first announced to the world at the American Educational Research Conference (Glass, 1976). Meta-analysis is a method of combining the findings of similar studies to provide a quantitative synthesis or overall 'pooled estimate of effect' (Borenstein et al. 2011). Meta-analysis has become a well-used research tool, in a wide range of disciplines (Bangert-Drowns, Rudner & Lawrence, 1991; Gough et al. 2012). More specifically, a search on the ERIC database in March 2015 identified more than 4,000 articles written since 1996 that use or discuss meta-analysis. A number of different

aspects of using meta-analysis as a 'tool' to benefit educational research and practice have been addressed in this literature. This paper aims to add to this debate, using examples from the 'Toolkit' in terms of how findings are communicated and appropriate inferences can be drawn amongst a range of diverse perspectives.

Advantages & Limitations

One of the main reasons for the increased popularity of meta-analysis is that it addresses a number of limitations commonly encountered in reviewing research. One key advantage of meta-analysis is that it helps to deal with the quantity of information which can overwhelm other approaches (Chan & Arvey, 2012) to provide an overall summary average in answer to a particular question.

More specifically, it combines or 'pools' estimates from a number of studies and can therefore produce more widely applicable and generalizable inferences when compared with a single study. In addition, it can show whether the findings from similar studies vary more than would be predicted from their samples so that the causes of this variation can be investigated using moderator analysis to see what features influence specific effects, such as the length of time pupils studied, or the importance of training and support, or the use of particular resources, drawing on data from across the included studies. This is an important point, especially for education research where the results from small studies can be combined to provide answers to questions without being so dependent on the statistical significance of each of the individual studies, which is directly related to sample size (Gorard, 2014). Many small studies with moderate or low effects may not reach statistical significance and if you review the field by simply counting how many were statistically significant, one may be misled into thinking that the evidence is less conclusive than if studies are combined into a single meta-analysis to look at the overall pattern. One example of this is Ritter and colleagues (2009) meta-analysis of volunteer tutoring. In the twenty studies included in this review, there are only three statistically significant results and two negative (but non-significant) findings, largely because the nature of the topic makes large samples difficult to research (the average sample size is 41). A pooled effect suggests a more positive interpretation with an effect size of 0.30 (with upper and lower confidence intervals from 0.18 to 0.42, indicating statistical significance at the 95% level). The statistical techniques to undertake meta-analysis form a set of transparent and replicable rules which are open to

scrutiny and which have been accepted across a number of disciplines (Aguinis et al., 2010).

The ability to include a range of studies is particularly important when trying to draw cumulative inferences in a specific area of education research. The number of studies available to review in any area of education can be extensive; therefore techniques to aggregate and build up understanding of a field in terms of the impact of an intervention or approach are invaluable. Meta-analysis has a strong and relatively uncontroversial base, especially in the fields of psychology and medicine, with nearly 40 years of development as a method of synthesis.

On the other hand, there are limitations and perhaps the most important is the assumption that the data from evaluations are equivalent across studies. Here the key issue is a conceptual one (Lipsey & Wilson, 1993). Are the studies being compared the same in terms of the way that they have defined or implemented a particular approach? This also relates to the nature of the question being addressed. Asking whether phonics interventions are effective for beginning readers is different from asking whether phonics approaches are the best approach for beginning readers (when compared with other approaches). Some studies would be included in both reviews, but in one it may be helpful to combine studies in different categories (phonics, whole word, comprehension-led, whole language, etc.) and clarity about definitions and inclusion criteria are essential. These issues are commonly seen in comparative meta-analyses where inferences are drawn between types of intervention and caution is needed from the early stages clearly defining a research question to inclusion criteria for identification of the studies. More specifically, the methods, research questions and terminology used in the studies have to be scrutinised beforehand in order to increase the uniformity and therefore the comparability of the data. Once these characteristics have been examined and categorised then the researchers can include or exclude a study from a collection of meta-analyses for future comparisons, depending on the precise question to be answered. Apart from these considerations for researchers there are additional considerations for practitioners. A meta-analysis can over-emphasise the average effect and blur the focus on what is associated with variation in impact. In particular there may be other characteristics of an intervention that need to be closely attended to in terms of outcomes. As Lendrum and Humphrey (2012) show the implementation of a programme can affect the outcomes. Therefore, it can therefore be very important to study in depth the different stages of a

programme's creation and delivery in order to understand the and replicate the causal mechanisms so as to produce the same positive outcomes (Cartwright & Hardie, 2012). Here we also need to consider the variation in the pupils' outcomes between schools. Using meta-analysis we can further investigate the reasons for the differences found or the observed variance (Sellström & Bremberg, 2006) not only in terms of school effects but other variables that can explain differences in outcomes.

Another limitation is publication bias which arises whenever the probability of a study being published depends on the statistical significance of the results or the belief that positive findings are more worthy of reporting and is often called the 'file-drawer' problem where studies with negative effect are not reported (Scargle, 1999). If a field is systematically missing null or negative studies; then meta-analysis will provide an inflated estimate of the overall effect. Additionally, we have to be cautious with many evaluations of impact in education where the nested or clustered nature of schooling is not taken into effect (Raudenbush, 1997; Campbell et al. 2012). Pupils work in classes in schools and both the class and the school may influence the impact of different approaches (see Sellström & Bremberg, 2006 for an account of the "school effect" on different kinds of pupil outcomes. Analysis needs to take this aspect of variation into account or the effects may be overestimated (Hedges and Olkin, 2014).

However, as mentioned earlier, there are procedures to guard against potential biases through transparent and conceptually clear inclusion and exclusion criteria, careful searching and systematic review, consideration of heterogeneity of effects and publication bias to understand the nature of the data included in a meta-analysis, so as to inform interpretation of the findings. Although there are limitations to the application of quantitative synthesis as described above, the data from meta-analysis offers the best available source of information to address cumulative questions about effects in different areas of educational research and in understanding what might explain differences in effects with statistical techniques which are relatively uncontroversial. This approach is adopted by the US What Works Clearinghouse (see: <http://ies.ed.gov/ncee/wwc/>) in identifying the effectiveness of particular educational programmes.

Comparative meta-analysis or 'super-synthesis'

A further approach drawing wider inference from interpreting results from meta-analyses is to look at findings across different kinds of studies with a common population, so to

provide more general or comparative inferences. This approach is, of course, vulnerable to the classic ‘apples and oranges’ criticism which argues that you cannot really make a sensible comparison between different kinds of things. However as Gene Glass (2000) said, “Of course it mixes apples and oranges; in the study of fruit nothing else is sensible; comparing apples and oranges is the only endeavor worthy of true scientists; comparing apples to apples is trivial.”

A number of researchers have attempted to take meta-analysis this stage further, by synthesising the results from a number of existing meta-analyses. Here there is less consensus in the terminology with result described as a ‘meta-meta-analysis’ (Kazrin, Durac & Agteros, 1979), a ‘mega-analysis’ (Smith 1982), ‘super-analysis’ (Dillon, 1982) or ‘super-synthesis’ (e.g. Sipe & Curlette, 1997). However, one can make a clear separation of types within these approaches. Some use the meta-analyses as the unit of analysis in order to say something about the process of conducting a meta-analysis and identifying statistical commonalities which may be of importance (e.g. Ioannidis & Trikalinos, 2007; Lipsey and Wilson, 1993). Others, however, attempt to combine different meta-analyses into a single message about a more general topic than each individual meta-analysis can achieve (e.g. Bloom, 1984; Walberg, 1984; Hattie, 1992; Sipe & Curlette, 1997; Hattie, 2008). Even here, there appears to be a qualitative difference – some retain a clear focus, either by using meta-analyses as the source for identifying original studies with an overarching theoretical focus (e.g. Marzano, 1998), so in effect producing something might best be considered as a series of larger meta-analyses rather than a meta-meta-analysis. Others, though, make claims about broad and quite distinct educational areas by directly combining results from identified meta-analyses (e.g. Hattie, 1992; Sipe & Curlette, 1997; Hattie, 2008). Given the different current views on comparative meta-analysis which aim to relate effects between meta-analyses we outline the value of meta-analyses as a means to evaluate the overall effects of different school-based approaches which aim to improve attainment and the relative effects when compared with different areas. For example, there are currently numerous high quality meta-analyses exploring the effects of digital technology on pupils’ attainment. In our research we collect these studies, compare their results and calculate an overall estimate of effect of the effect of digital technology-based interventions on school-age pupils attainment., We therefore focus on exploring a broad educational approach rather than a specific programme (for example a software programme such as

Accelerated Reader that can be used in to promote reading, see for example Nunnery et al. 2006)). It is beyond the scope of this paper to present how one can use comparative meta-analyses to select specific interventions with a range of outcomes, but rather we focus on how to use this ‘tool’ to get a general picture of approaches that have worked or have not worked in terms of their impact on tested attainment. We also acknowledge that we have a narrow focus in this paper on curriculum and cognitive outcomes as measured by tests of attainment and there are many other valuable outcomes from schooling which could be assessed in areas, such as health and well-being, self-efficacy or attitudes and dispositions. However one also could argue that academic and cognitive outcomes are the particular goal of education therefore a number of examples from the ‘Toolkit’ are presented to exemplify the potential of comparative meta-analysis.

Issues & challenges

It is hard to compare effects across education research without some kind of benchmark. If you have two narrative reviews, one arguing that, say, parental involvement works and another arguing that digital technology is effective, and both cite studies with statistically significant findings showing they each improve reading comprehension, it is hard to choose between them in terms of which is likely to offer the most benefit if you are a practising teacher. Meta-analysis certainly helps to identify which researched approaches have made, on average, the most difference, in terms of effect size, on the tested attainment of pupils in reading comprehension or other areas of attainment. We suggest that this comparative information should be treated cautiously, but taken seriously. If effect sizes from a series of meta-analyses in one area, such as meta-cognitive interventions for example, all tend to be between 0.6 and 0.8, and all of those in another area, such as individualised instruction, are all between -0.1 and 0.2, then this is persuasive evidence that schools are likely to find more potential in investigating meta-cognitive approaches to improve learning, rather than focusing on individualised instruction. Some underlying assumptions are that the research approaches are sufficiently similar (in terms of design for example), that they compared sufficiently similar samples or populations (of school pupils) with sufficiently similar kinds of interventions (undertaken in schools) and similar outcome measures (standardised tests and curriculum assessments). So, if you think that a meta-analysis of intervention research into improving reading comprehension has a set of broadly similar set of studies, on average, to a meta-analysis investigating the development of understanding in science,

then you might be tempted to see if any approaches work well in both fields (such as reciprocal questioning which appears to do so: Dignath & Büttner, 2008) or, indeed, do not work well in both fields (such as individualised instruction: see Bangert, Kulik & Kulik, 1983 and Willett & Yamashita, 1983).

Our argument is that so long as you are aware of the limits of the inferences drawn, then the approach has value. We suggest that this provides the best cumulative and comparative evidence we have so far, particularly where there are no studies providing direct comparisons. In ‘Visible Learning, Hattie (2009) takes the average of all educational meta-analysis as a comparison point and argues that it is only effects above this ‘hinge’ point which are of interest to policy and practice. We are more cautious (see Higgins & Simpson, 2011 for a critique of ‘Visible Learning’) and argue that small effects may be valuable if they help tackle a difficult educational challenge or if they are cheap and reliable to implement. This argument is developed further in the section on cost-effectiveness below.

Developing a ‘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 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). This included an analysis of cost together with an evaluation of the extent of evidence and is presented in Figure 1 below.

Toolkit to improve learning: summary overview

Approach	Potential gain ¹	Cost	Applicability	Evidence estimate	Overall cost/benefit
Effective feedback	+ 9 months	££	Pr, Sec Maths Eng Sci	⊕ ⊕ ⊕	Very high impact for low cost
Meta-cognition and self-regulation strategies	+ 8 months	££	Pr, Sec, Eng Maths Sci	⊕ ⊕ ⊕ ⊕	High impact for low cost
Peer tutoring/ peer-assisted learning	+ 6 months	££	Pr, Sec Maths Eng	⊕ ⊕ ⊕ ⊕	High impact for low cost
Early intervention	+ 6 months	£££££	Pr, Maths Eng	⊕ ⊕ ⊕ ⊕	High impact for very high cost
One-to-one tutoring	+ 5 months	£££££	Pr, Sec Maths Eng	⊕ ⊕ ⊕ ⊕	Moderate impact for very high cost
Homework	+ 5 months	£	Pr, Sec Maths Eng Sci	⊕ ⊕ ⊕	Moderate impact for very low cost
ICT	+ 8 months	££££	Pr, Sec All subjects	⊕ ⊕ ⊕ ⊕	Moderate impact for high cost

¹ Maximum approximate advantage over the course of a school year that an ‘average’ student might expect if this strategy was adopted – see Appendix 3.

Figure 1: The 2011 Pupil Premium Toolkit

We conceptualised these as a series of related ‘umbrella reviews’ (Grant & Booth, 2009) which would provide a rigorous but accessible summary with a common methodology across the different strands. The feedback at both policy and practice levels convinced us that this was worth developing further and in late 2011 the recently formed Education Endowment Foundation adopted the approach and committed to funding the development of school-based projects across England aiming to raise attainment (see: <https://educationendowmentfoundation.org.uk/>). The Sutton Trust-EEF Teaching and Learning Toolkit is therefore presented in an easily accessible summary form at the surface level (<https://educationendowmentfoundation.org.uk/toolkit/toolkit-a-z/>) but, for transparency, further detail on each area is available, right through to the effect sizes and abstracts of the meta-analyses and other studies used in its compilation provided for each area. These areas have been included based on approaches commonly mentioned in educational policy, school suggestions and areas with a strong evidence of effectiveness not covered by the previous two criteria. The overall accessibility is important as engagement with research evidence is a significant challenge (Cordingley, 2008; Hemsley-Brown & Sharp, 2004). However the apparent simplicity can be deceptive as the messages in any specific area are rarely straightforward so successive levels of detail aim to support deeper engagement. Our intention was to keep the surface level easy to understand, but also to provoke a level of interest or challenge which engages a practitioner in looking deeper. This is the rationale for the way that we have chosen to communicate our findings. Feedback from teachers and practitioners and overall uptake suggests that it has been proven to be helpful. The balance between accessibility and accuracy is a difficult challenge, particularly when it is also important to engage practitioners in the resource. A key limitation, as noted above, is the single focus on tested attainment and not a wider evaluation of the outcomes from schooling.

The next sections provide more detail about the rationale and the main methods and assumptions used in the comparative synthesis of effect sizes in the Sutton Trust-EEF Teaching and Learning Toolkit with reference to other similar approaches where there are key differences. Our emphasis is on identifying comparative messages from research so a number of examples will be presented. We suggest that comparative meta-analysis has an important role to play in educational research; providing information to inform the field, not about ‘what works’ but as a summary of ‘what has worked’. This is an important distinction. The evidence has accumulated over 30 years or so, and derives

from a range of contexts, often drawing heavily on research conducted in the United States, and represents what has been effective compared with what was usual practice at the time (the ‘counterfactual’: see Lemons et al. 2014 for an important discussion of this, as well as earlier arguments from Bracht and Glass, 1968). This affects the warrant for the claims of the Toolkit and argues for professional interpretation and evaluation, rather than a simple ‘application’ of research findings. The US ‘What Works Clearinghouse’ makes a wider claim to external validity in both its name and for the programmes it endorses, based on the internal validity of its analysis and synthesis process and of the underlying studies. However we believe external validity in education is problematic (see, for example, the classic paper by Bracht & Glass, 1968). The lack of replication in education makes this a particular issue (Ahn, Amers & Myers, 2012).

Further Considerations for Practitioners & Researchers

The above approach aims to address two of Biesta’s (2010) objections in the epistemological and practice domains. Two dimensions are important here. First, in understanding the applicability of findings, it is important to know what an intervention or approach was compared with. If your own normal practice is different from the counterfactual in the study, you may see very different results, no matter how rigorous and robust the study. Second, the intervention will have had a specific and particular instantiation, by the researchers, the schools, teachers and pupils involved. As a practitioner you need to know whether your normal practice is similar to that in the control group (which is rarely described), and whether you can apply the intervention in a sufficiently similar way to the study (e.g. without the support from the research team and with similar pupils). With cumulative meta-analysis both the counterfactual or normal practice and the intervention are averaged, identifying what might be thought of as ‘good bets’ for areas where consistent positive findings are found and ‘long odds’, where, on average approaches have been less successful. This also encapsulates to some extent the range of findings included as most approaches have their successes and their failures which can be masked by the overall average. This is a key premise of our argument, suggesting a novel perspective on interpreting findings from randomised trials in general, but also in terms of comparative meta-analysis in particular.

As mentioned earlier, an important aspect of successfully replicating an intervention is where practitioners and researchers need to be aware of is the programme’s implementation in all stages. For example, during the effectiveness stage critical factors

are identified that can influence later stages. Therefore, it is beneficial if these factors are identified before the broader dissemination stage (Lendrum & Humphrey, 2012). On a similar note, implementation should be monitored to ‘agree’ with the original design although various adaptations and deviations from the original implementation seem to be unavoidable in a natural setting. As Lendrum and Humphrey (2012) have shown there are a number of common factors which have been identified that affect implementation in different research areas supporting the argument that a complex mix of variables that link to implementation. Similarly, success of an intervention also seems to be related to the social acceptability and validity of those implementing it (Wolf, 1978). “This argument has been developed by the very nature of the social validity measures that can be manipulated or abused” (Wolf, 1978). Recognising both the possible shortcoming of social validity measures as well as the lack of evaluations involving rigorous effectiveness trials focusing on implementation is a vital aspect that practitioners and researchers should consider when they study meta-analytical results of various interventions. These concerns affect all areas of education research and at present it is difficult to know whether they affect different areas in different ways or whether there are common patterns, such as effect sizes decreasing as interventions are scaled up, as Wigelsworth and colleagues (2016) found for social and emotional interventions.

Toolkit Themes & Inclusion Criteria

The initial themes for the Sutton Trust-EEF Teaching and Learning Toolkit were based on expectations of how schools seemed likely to spend the Pupil Premium when it was first announced. A number of areas were specifically included at the request of teachers who have been consulted at different stages in the development of the Toolkit. The initial source of studies for the Toolkit was a database of meta-analyses of educational interventions developed for the ESRC Researcher Development Initiative (Training in the Quantitative Synthesis of Intervention Research Findings in Education and the Social Science), as mentioned above. Additionally repeated systematic searches have been undertaken for systematic reviews with quantitative data (where effect sizes are reported but not pooled) and meta-analyses (where effect sizes are combined to provide a pooled estimated of effect) of intervention research in education in each of the areas of the Toolkit. These searches have been applied to a number of information gateways including Web of Knowledge, FirstSearch, JSTOR, ERIC, Google Scholar and ProQuest

Dissertations. In addition a number of journals were hand searched (e.g. Review of Educational Research and Education Research Review). Journal publishers' websites offering full-text searching (Elsevier, Sage, Wiley-Blackwell) were also searched for meta-analyses. Relevant references and sources in existing super-syntheses (e.g. Sipe & Curlette, 1997; Marzano, 1998; Hattie, 2008) were identified and obtained where possible. A record of the search strategy used and studies found are kept for each of the Toolkit themes. Other studies found during the search process are also consulted in each area to provide additional contextual information.

Specific inclusion criteria have been developed and standardised during the first stages of this process and are summarised in the figure below:

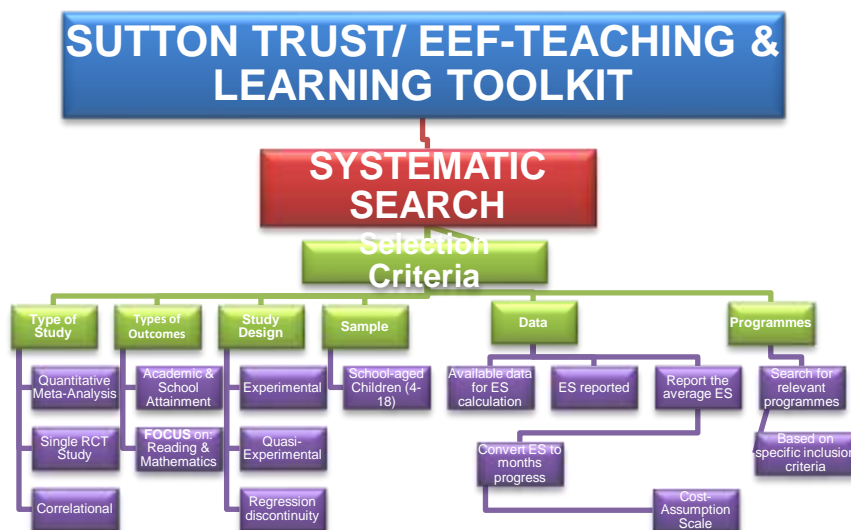


Figure 2: Search and inclusion strategy

These criteria form the basis of our search and guide us through the systematic review process. Once the studies for inclusion have been selected we extract a number of information. These include the pooled effect size, standard error, type of effect size (Hedges' g , Cohen's d or Glass' Δ), confidence intervals, number of studies in the meta-analysis, moderator analysis, and publication bias, amongst others. In each area of the Toolkit an overall estimate of the effects is then identified. Where the data is available and suitable a weighted mean is calculated. This is based on calculating a weight for each meta-analysis according to its variance, based on the reciprocal of the square of the

standard error (Borenstein et al., 2010). Where the data is not available for this an estimate is given based on the available evidence (such as mean and median effects) and a judgement made about the most applicable estimate to use (such as the impact on disadvantaged pupils, or the most rigorous of the available meta-analyses). Where no meta-analyses of educational interventions in a given area can be found an effect size is estimated from correlational studies or large-scale studies investigating the relationship under review. If there is no information from any of these types of research can be found, then individual studies are identified which can provide a broad estimate of effect. The priority during the systematic review is to find rigorous meta-analyses, but in areas where this is not possible (mostly due to lack of available research) there are identified quality criteria that are assigned to each strand to depict the quality of evidence for each topic. Figure 3 represents how different evidence criteria are assigned to different Toolkit topics.






Rating	Description
	<i>Very limited:</i> Quantitative evidence of impact from single studies, but with effect size data reported or calculable. No systematic reviews with quantitative data or meta-analyses located.
	<i>Limited:</i> At least one meta-analysis or systematic review with quantitative evidence of impact on attainment or cognitive or curriculum outcome measures.
	<i>Moderate:</i> Two or more rigorous meta-analyses of experimental studies of school age students with cognitive or curriculum outcome measures.
	<i>Extensive:</i> Three or more meta-analyses from well-controlled experiments mainly undertaken in schools using pupil attainment data with some exploration of causes of any identified heterogeneity.
	<i>Very Extensive:</i> Consistent high quality evidence from at least five robust and recent meta-analyses where the majority of the included studies have good ecological validity and where the outcome measures include curriculum measures or standardised tests in school subject areas.

Figure 3: Evidence security estimates

This classification acts as a useful guideline providing information about how extensive the evidence is, if there are available studies, the quality of the methodology and the reliability of the impact across the reviewed studies in each strand of the Toolkit. Hence, a more complete picture can be drawn regarding the impact of each approach in the Toolkit. For example, aspiration-based interventions currently show to have little to no impact on attainment, but this is only one part of the picture, since the evidence base for this area is weak with no systematic reviews or meta-analyses available, as noted in two recent high quality reviews conducted for the Joseph Rowntree Foundation (Gorard, See & Davis, 2012; Cummings et al. 2013). Therefore, before jumping to conclusions we need to consider all available information. Again this is an important feature that can be incorporated in relevant summaries. By contrast, the US ‘What Works Clearinghouse’ only reports outcomes from studies which meet its rigorous inclusion criteria. This is important when summarising effects of programmes where there is extensive evidence and when the question is whether a particular intervention works or not. But by excluding areas with little rigorous evidence there is a danger that areas of interest to teachers and schools may be overlooked. Whilst the level of confidence in less rigorous research is lower, it is often a useful indicator and can provide more indicative evidence. Another consideration is that correlational and descriptive research is often essential to build new ideas and theories that can be further investigated using experiments or explore areas where experimental designs are not possible (Slavin, 2002). We do recognise the value of well-designed research but we further suggest that areas that have educational importance should not be overlooked only because there are no or limited rigorous evidence from randomised controlled trials. On the contrary, a more complete perspective should be presented for the practitioners to use this information to get a better view of different research areas.

All the stages involved in the development of the Toolkit serve as examples that can be used to inform similar approaches to synthesise and communicate research findings. They have proven to be useful tools and processes that have produced summaries with a level of rigour and transparency to defend their accuracy while introducing accessible and user-friendly ways of communicating knowledge with to practitioners from differing backgrounds.

Effect sizes: presentation, interpretation & implications

An effect size (standardised mean difference) is a key measure in intervention research and an important concept in the methodology of the *Toolkit* as well as a common metric used to present meta-analytic findings. It is basically a way of measuring the *extent* of the difference between two groups (Lipsey & Wilson, 1993; Vacha-Haase & Thompson, 2002). It is fairly easy to calculate, and can be applied to any measured outcome for groups in education or in other areas of research more broadly.

The value of using an effect size is that it quantifies the effectiveness of a particular intervention, relative to a comparison group. It allows us to move beyond the simplistic, 'Did it work (or not)?' to the far more important, 'How *well* did it work across a *range* of contexts?' It therefore supports a more critical and rigorous approach to the accumulation of knowledge, by placing the emphasis on the most important aspect of the intervention – the size of the effect – rather than its statistical significance, which conflates the effect size and specific sample size. For these reasons, effect size is the most important tool in reporting and interpreting effectiveness, particularly when drawing comparisons about *relative* effectiveness of different approaches. A perhaps under-used benefit of using effect size is that findings can also be converted back to the original scale, whether that is reading progress or national examination scores. We have not attempted to do this in the *Toolkit*, as the focus is on emphasising the relative benefits of different approaches. However it would be possible to do so and provide a more specific estimate of impact on a test or outcome of interest to a practitioner.

In the *Toolkit* we were keen to promote the understanding of the effect size to different audiences and to make the findings more accessible. Therefore, we equated effect size to a single scale of school progress in months to as a crude but meaningful equivalent. The use of a single scale which was intuitively easy to understand was an important aspect of the accessibility of the entry level of the overall *Toolkit* summary. We have estimated that a year of progress is about equivalent to one standard deviation per year and corresponds with Glass' observation that "the standard deviation of most achievement tests in elementary school is 1.0 grade equivalent units; hence the effect size of one year's instruction at the elementary school level is about +1" (Glass, 1981: 103). However, it is important to note that the correspondence of one standard deviation to one year's progress can vary considerably for different ages and types of test. It is also the case that effect size difference reduces with age. Hill and colleagues (2008) estimate

annual progress on tests drops from 1.52 to 0.06 for reading and from 1.14 to 0.01 for mathematics in the US from Kindergarten to Grade 12. Wiliam (2010) estimates “apart from the earliest and latest grades, the typical annual increase in achievement is between 0.3 and 0.4 standard deviations”. One implication of this is that our estimates of improvement may underestimate the gains achievable for older pupils. By the end of secondary school age, the difference between the attainments of successive age groups is relatively small, especially compared with the spread within each. For these older pupils it may be a bit misleading to convert an effect size into typical months’ gain: one month’s gain is typically such a small amount that even quite a modest effect appears to equate to what would be gained in a long period of teaching.

There are other reasons for preferring a more conservative estimate of what it likely to be achievable in practice. One problem is that estimates of the effects of interventions come from research studies that may optimise rather than typify their effects. For these reasons it may be unrealistic to expect schools to achieve the gains reported in research whose impact may be inflated (this is what Cronbach and colleagues (1980) called ‘super-realisation bias’). Other evidence suggests that effect sizes will also be smaller as interventions are scaled up or rolled out (Wigelsworth et al. 2016). Slavin and Smith (2009) report that there is a relationship between sample size and effect size in education research, with smaller studies tending to have larger effect sizes. This may be due to the stage of intervention (pilot, efficacy and effectiveness) as Wigelsworth and colleagues have shown

A further problem is that part of the learning gain typically achieved in a year of schooling may be a result of maturational gains that are entirely independent of any learning experiences that are, or could be, provided by the school (Luyten et al., 2006).

Researchers and practitioners should therefore take into consideration the potential variables that can affect the observed gains in pupils’ progress. For these reasons we have also selected what we see as a more conservative estimate, based on effect size estimates for younger learners, which can be improved or refined as more data becomes available about effect size transfer from research studies to practice. Figure 4 shows the different categories that fall under the estimates of the effect sizes translated into month’s gain.

Months’ progress	Effect Size from to	Description
-------------------------	---------------------------------	---------------	--------------------

0	-0.01	0.01	Very low or no effect
1	0.02	0.09	Low
2	0.10	0.18	Low
3	0.19	0.26	Moderate
4	0.27	0.35	Moderate
5	0.36	0.44	Moderate
6	0.45	0.52	High
7	0.53	0.61	High
8	0.62	0.69	High
9	0.70	0.78	Very high
10	0.79	0.87	Very high
11	0.88	0.95	Very high
12	0.96	>1.0	Very high

Figure 4: Effect size conversion

Presenting effect sizes as month’s additional gain has been proven to have had a significant role in the communication of our findings to practitioners, schools and people who do not have (and may not want) a full understanding of effect sizes. Presented in a simple but meaningful way research findings are disseminated to diverse audiences. We believe that this is an important aspect of our approach, especially for educational research, though there is a trade-off here again between accuracy and accessibility. One of the main criticisms that we have encountered during the development of the Toolkit was that frequently educational research lacks clarity in presenting research findings to wider non-academic audiences. Thus translating the effect size into month’s gain made results from meta-analyses and quantitative studies easy to comprehend, but at the cost of some of the precision in these estimates.

There are some further notes of caution in comparing effect sizes across different kinds of interventions and evaluations (see Cheung and Slavin, 2015). Effect size as a measure assumes a normal distribution of scores. If this is not the case then an effect size might provide a misleading comparison. If the standard deviation of a sample is decreased (for example, if the sample does not contain the full range of a population) or inflated (for example, if an unreliable test is used), the effect size is affected. A smaller standard deviation will increase the effect size whereas a larger will reduce it. Another key issue is

which standard deviation is chosen (Hill, Bloom, Black & Lipsey, 2008) as this primarily determines the comparability of the effect size (Coe, 2002). This choice can explain the variation in methods advocated above. For example, a decision has to be made as to whether to use the control group's standard deviation or a 'pooled estimate' of both the experimental and control group. There are different implications involved in the above choice described in detail by Olejnik and Algina (2000).

There is also some variation associated with the type of outcome measure with larger effect sizes typically reported in mathematics and science compared with English (e.g. Higgins et al., 2005) and for researcher designed tests and teacher assessments compared with standardised tests and examinations (e.g. Hill et al., 2007, p. 7).

Finally, studies reporting effect sizes with groups from either end of the distribution (high attaining or low attaining learners) are likely to be affected by regression to the mean if they do not compare similar learning levels (Shagen & Hogden, 2009). This would inflate effect sizes for low attaining pupils (who are more likely to get higher marks on re-test) and depress effect sizes for high performing students when they are compared with 'average' pupils. If the correlation between pre-test and post-test is 0.8, regression to the mean may account for as much as 20% of the variation in the difference between test and retest scores when comparing low and average students.

In summary, considerable caution is needed in making comparisons where a number of factors need to be taken into account in understanding what influences effect size estimations. One of the assumptions in the Toolkit is that the overall distribution of studies in educational research is sufficiently similar that the patterns which emerge from this analysis represent our best indicator of the relative effects of different approaches. An important part of our argument is that we think this is an empirical question. The current synthesis is the best that we have, but this can be refined in two ways, first by undertaking meta-analyses which are designed to be more comparable, such as by having common inclusion criteria and comparing impact on similar groups of pupils with similar kinds of outcome measures. The second is by looking at the findings from future studies to see whether they follow a similar pattern. Between 2011 and 2014, the EEF commissioned 87 new experimental trials involving over half a million pupils in England, with one aim being to feed the results back into the Toolkit (EEF, 2014) to test how useful the findings from previous research are as predictions of impact in subsequent interventions.

Broader implications: proximal and distal effects

Looking at the findings across the Toolkit or other similar syntheses also offers possibilities for further interpretation and broader implications. One of these is the value of approaches which focus on the quality of interactions in teaching and learning processes. Proximal interventions which directly influence teaching and learning interactions, such as feedback, meta-cognition and self-regulation, peer tutoring, one-to-one tuition and collaborative learning are grouped as the upper end of effects (0.4 to 0.6), whereas more distal influences on teaching and learning such as performance pay, class size, ability grouping or school uniforms have much lower average effects. This suggests that if schools are interested in improving outcomes directly, then focusing on teaching and learning is likely to be more productive (see Figure 5). This is similar to Seidel and Shavelson's (2007) analysis of teacher effectiveness research where similar distal and proximal effects were noted. However it is also possible that some of the greater impact is related to how closely the intervention was related to the outcome measures.

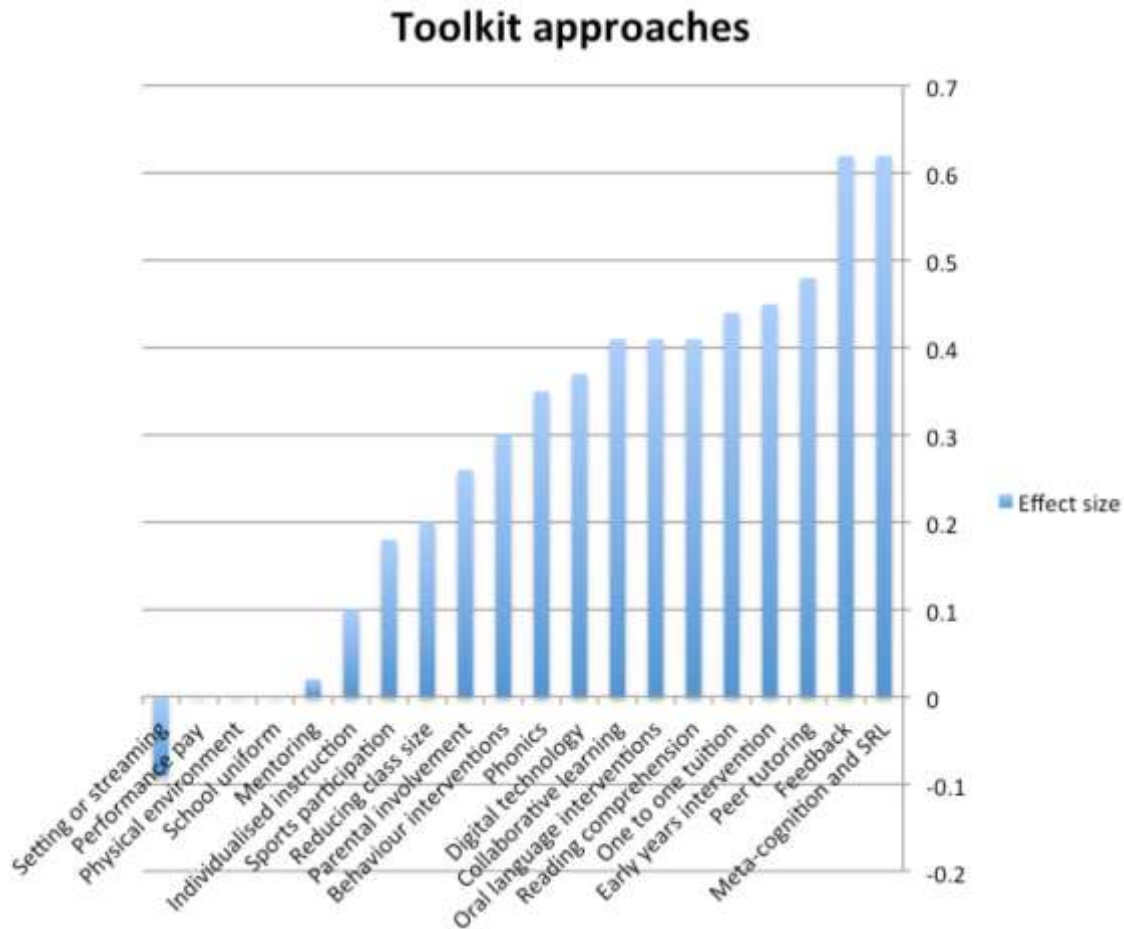


Figure 5: Toolkit approaches ranked by effect size

Cost-effectiveness estimates

Another key element of the Toolkit is the cost-benefit analysis that takes place for each intervention. Our argument here is that an approach which is cheap and easy to adopt, but has a relatively low impact, may still be better than one which is very expensive. More specifically, cost estimates are calculated based on the additional likely costs of adopting an approach with a class of 25 students. In cases where an approach does not require any additional resources, estimates are based on the cost of training or professional development to establish new practices. In terms of cost-effectiveness and the available options for schools, they need to consider the investment they are making. For example, reducing class sizes only lasts for as long as the funding maintains smaller classes. Technology equipment typically lasts for three to five years. On the other hand, developing teachers' skills through professional development can be more valuable since it may make more 'permanent' and long lasting changes within the school environment. Hence, having this information for all presented interventions schools can compare the costs along the potential 'duration' that a specific intervention approach can have. Figure

6 shows the overall costing assumptions that have been used to give an estimate for each approach presented in the Toolkit.

£	<i>Very low:</i> up to about £2,000 per year per class of 25 pupils, or less than £80 per pupil per year.
££	<i>Low:</i> £2,001-£5,000 per year per class of 25 pupils, or up to about £170 per pupil per year.
£££	<i>Moderate:</i> £5,001 to £18,000 per year per class of 25 pupils, or up to about £700 per pupil per year. This represents the 2012/13 Pupil Premium allocation (£623).
££££	<i>High:</i> £18,001 to £30,000 per year per class of 25 pupils, or up to £1,200 per pupil.
£££££	<i>Very High:</i> over £30,000 per year per class of 25 pupils, or over £1,200 per pupil. By 2014/5, the Pupil Premium is projected to rise to approximately £1,200 per pupil.

Figure 6: Cost estimates

This feature of the Toolkit has also been proven to be valued, informing school leaders and teachers on this important dimension upon which, among others, they base their decisions on how to spend the Pupil Premium (in England) or other discretionary spending. Educational research rarely reports the costs of different approaches, so from both a policy and practice perspective this is usually seen as important. This makes the Toolkit distinctive when compared with other approaches in education. Similarly with the translation of effect sizes into month's gain, the cost is a simplification in that the focus is on the extra cost of an approach or intervention to develop and implement, but does not include the overall school running costs including teachers' salaries, unless an approach requires additional investment in staffing. This again reflects the tension between accessibility and accuracy when summarising findings from research.

Example of comparative meta-analysis from the Toolkit

The Toolkit is currently comprised of 34 topics summarising different approaches to intervention having as a main goal the improvement of school attainment. Each topic has information about how much does the intervention cost to implement, the breadth of evidence, and how many months gain can be achieved. Additionally, detailed information is provided about the approach along with any suggested programmes that match our eligibility criteria which have evidence of impact. Therefore, the online presentation of the Toolkit provides both a brief and quick overview of the topics as well as a more thorough presentation for those who seek additional material. The tension here is in

presenting comprehensive information or limiting the analysis to areas of particular interest to practitioners.

All information on which the analysis has been made is available and can be viewed in a comparative way. For example mentoring might have a low impact on attainment but it is cheap to administer so compared with one-to-one tutoring which is more expensive teachers might decide to use the first intervention. Another example having different comparisons is the teaching assistants and one-to-one tutoring. Both have high costs but, on average, the former has one month gain and the latter five months. Another variable that can be considered is the evidence rating. Small group tuition has a limited evidence rating providing 4 months additional gain. On the other hand, summer schools have two months additional gain in attainment but more extensive evidence of effectiveness.

Therefore, teachers might consider using an approach which appears more promising based on the extensiveness of the evidence surrounding this approach. As argued above, synthesis of evidence in terms on impact, cost and evidence strength should inform schools' decision-making, not as guarantees of 'what works', but as indicating the likely chances of success. If a school chooses an area where effects tend to be lower and the evidence is robust then we suggest that they need a clear understanding of what they should do to ensure they are more effective than the average approach described in the studies. If they choose an area of high effect, they will only need to be as successful as schools typically were in the studies included in the synthesis: good bet, as opposed to a riskier one. This perspective also encourages schools to take responsibility to ensure that new approaches are successful, rather than assuming that it is something which 'works' without this commitment. Taken together all the aforementioned characteristics using the Toolkit as an example we suggest that these are important elements that need to be both presented in educational research and carefully considered by the practitioners and researchers.

Summary

The overall aim of the Toolkit is not to provide definitive claims as to what will work to improve learning as a guarantee of future success. Rather it is an attempt to provide the best possible typical estimate of what is likely to be beneficial based on existing evidence which includes the changing nature of the counterfactual in this estimate. It exemplifies how we think meta-meta-analysis or super-synthesis should be used to inform practice. More specifically, it currently serves as an example of different ways of communicating

research findings in the educational area while making evidence more accessible for practitioners and decision-makers.

The variation in findings in education and aggregation process means that applicability of this information to a new context is going to be a probability rather than a certainty as it emphasises the average effects, rather than the range and what might cause variation. We think interpretation and application is always likely to need active enquiry and evaluation to ensure it helps to achieve the desired effects. This requires professional judgement and commitment to engaging with evidence, but also a disposition to interpret, challenge and test particular findings to ensure they are helpful. It also addresses one part of Biesta's (2007) criticism of education research as restricting opportunities for participation in educational decision-making, as the emphasis is on the professional choices that practitioners could (and should) make on the basis of the accumulated evidence. Overall, the present article aimed to present the 'Toolkit' as an example of communicating research to a wider audience as well as re-considering the ways of using findings from comparative meta-analysis to inform future approaches. We believe it balances some of the key tensions between accessibility, accuracy and applicability. We also believe that it makes a case for summaries derived from comparative meta-analysis which can help practitioners and others interested in educational outcomes to understand the relative effects of different educational approaches.

References

- Aguinis, H., Pierce, C. A., Bosco, F. A., Dalton, D. R., & Dalton, C. M. (2011). Debunking myths and urban legends about meta-analysis. *Organizational Research Methods*, 14(2), 306-331.
- Ahn, S., Ames, A. J., & Myers, N. D. (2012). A review of meta-analyses in education: Methodological strengths and weaknesses. *Review of Educational Research*, 82 (4), 436-476 <http://dx.doi.org/10.3102/0034654312458162>
- Bangert, R.L., Kulik, J.A., Kulik, C.C. (1983). Individualized Systems of Instruction in Secondary Schools. *Review of Educational Research*, 53.2. pp. 143-158.
- Bassey, M. (2001). A solution to the problem of generalisation in educational research: fuzzy prediction. *Oxford Review of Education*, 27(1), 5-22.

- Bell, M., Cordingley, P., Isham., C. & Davis., R. (2010) Report of Professional Practitioner Use of Research Review: Practitioner engagement in and/or with research. Coventry: CUREE, GTCE, LSIS & NTRP. Available at: <http://www.curee-paccts.com/node/2303>
- Biesta, G. (2007). Why “what works” won’t work: Evidence-based practice and the democratic deficit in educational research. *Educational Theory*, 57(1), 1-22.
- Biesta, G. J. (2010). Why ‘what works’ still won’t work: From evidence-based education to value-based education. *Studies in Philosophy and Education*, 29(5), 491-503.
- Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13(6), 4-16.
- Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to meta-analysis*. John Wiley & Sons.
- Campbell, M. K., Piaggio, G., Elbourne, D. R., & Altman, D. G. (2012). Consort 2010 statement: extension to cluster randomised trials. *British Medical Journal*, 345, 5661.
- Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: a practical guide to doing it better*. Oxford University Press, USA.
- Chan, M. E., & Arvey, R. D. (2012). Meta-Analysis and the Development of Knowledge. *Perspectives on Psychological Science*, 7(1), 79-92.
- Cheung, A. C., & Slavin, R. E. (2015). How Methodological Features Affect Effect Sizes in Education *Best Evidence Encyclopedia* Baltimore: Johns Hopkins University. http://www.bestevidence.org/word/methodological_Sept_21_2015.pdf
- Coe, R. (2002). It’s the effect size stupid; what effect size is and why is it important. Paper presented at the Annual Conference of the British Educational Research Association, University of Exeter, England, 12-14 September 2002.
- Cordingley, P. (2008). Research and evidence-informed practice: focusing on practice and practitioners. *Cambridge Journal of Education*, 38(1), 37-52.

- Cronbach, L.J., Ambron, S.R., Dornbusch, S.M., Hess, R.O., Hornik, R.C., Phillips, D.C., Walker, D.F. & Weiner, S.S. (1980). *Toward reform of program evaluation: Aims, methods, and institutional arrangements*. San Francisco, Ca.: Jossey-Bass.
- Cummings, C., Laing, K., Law, J., McLaughlin, J., Papps, I., Todd, L., & Woolner, P. (2012). *Can changing aspirations and attitudes impact on educational attainment? A review of interventions* York: Joseph Rowntree Foundation.
- Dignath, C., & Büttner, G. (2008). Components of fostering self-regulated learning among students. A meta-analysis on intervention studies at primary and secondary school level. *Metacognition and Learning*, 3(3), 231-264.
- Dillon, J.T., (1982). Superanalysis. *American Journal of Evaluation* 3(4) pp 35-43.
- Glass, G. V. (2000). *Meta-analysis at 25*. Available at: <http://glass.ed.asu.edu/gene/papers/meta25.html>.
- Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational researcher*, 3-8.
- Education Endowment Foundation (2014) *Annual Report 2013-4* London: EEF
- 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., See, B. H., & Davies, P. (2012). *The impact of attitudes and aspirations on educational attainment and participation*. York: Joseph Rowntree Foundation.
- Gough, D., Thomas, J., & Oliver, S. (2012). Clarifying differences between review designs and methods. *Systematic Reviews*, 1(1), 28.
- Grant, M. J., & Booth, A. (2009). A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal*, 26(2), 91-108.
- Hattie, J.A. (1992). Measuring the effects of schooling. *Australian Journal of Education*, 36, 5-13.
- Hattie, J.A. (2008). *Visible Learning*. London: Routledge.
- Hedges, L. V., & Olkin, I. (2014). *Statistical method for meta-analysis*. Orlando, FLA: Academic Press.

- Hemsley-Brown, J.V. and Sharp, C. (2004). 'The use of research to improve professional practice: a systematic review of the literature', *Oxford Review of Education*, 29 (4) 449-470.
- Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, 2(3), 172-177.
- 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.
- Ioannidis, J.P.A. & Trikalinos, T.A. (2007). The appropriateness of asymmetry tests for publication bias in meta-analyses: a large survey. *Canadian Medical Association Journal* 176 p 8.
- Kazrin, A., Durac, J., & Agteros, T. (1979). Meta-meta analysis: A new method for evaluating therapy outcome. *Behaviour research and therapy*, 17(4), 397-399.
- Lemons, C.J., Fuchs, D., Gilbert, J.K., & Fuchs, L.S. (2014), Evidence-Based Practices in a Changing World: Reconsidering the Counterfactual in Education Research. *Educational Researcher*, 43(5), 242–252.
- Lendrum, A., & Humphrey, N. (2012). The importance of studying the implementation of interventions in school settings. *Oxford Review of Education*, 38(5), 635-652.
- Lipsey, M.W. & Wilson, D. B. (1993). *Practical meta-analysis* Thousand Oaks, CA: Sage publications.
- Luyten, H. (2006). 'An empirical assessment of the absolute effect of schooling: regression/ discontinuity applied to TIMSS-95'. *Oxford Review of Education*, 32: 3, 397-429
- Marzano, R.J. (1998). *A Theory-Based Meta-Analysis of Research on Instruction*. Aurora, Colorado, Mid-continent Regional Educational Laboratory. Available at: <http://www.mcrel.org:80/topics/products/83/> (viewed 31/05/11).

- National Audit Office (2015) Funding for disadvantaged pupils London: National Audit Office <http://www.nao.org.uk/wp-content/uploads/2015/06/Funding-for-disadvantaged-pupils.pdf>
- Nunnery, J. A., Ross, S. M., & McDonald, A. (2006). A randomized experimental evaluation of the impact of Accelerated Reader/Reading Renaissance implementation on reading achievement in grades 3 to 6. *Journal of Education for Students Placed at Risk*, 11(1), 1-18.
- Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary educational psychology*, 25(3), 241-286.
- Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, 2(2), 173.
- Scargle, J. D. (1999). Publication Bias (The " File-Drawer Problem") in Scientific Inference. *arXiv preprint physics/9909033*.
- Sellström, E., & Bremberg, S. (2006). Is there a “school effect” on pupil outcomes? A review of multilevel studies. *Journal of Epidemiology and Community Health*, 60(2), 149-155.
- Sipe, T. & Curlette, W.L. (1997). A Meta-Synthesis Of Factors Related To Educational Achievement: A Methodological Approach To Summarizing And Synthesizing Meta-Analyses. *International Journal of Educational Research* 25 (7), pp 583-698.
- Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational researcher*, 31(7), 15-21.
- 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.
- Smith, N.L. (1982). Evaluative Applications of Meta- and Mega-Analysis. *American Journal of Evaluation* 3(4) pp 43.

- Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin*, 140(4), 1174-1204.
<http://dx.doi.org/10.1037/a0036620>
- Walberg, H. J. (1984). Improving the productivity of America's schools. *Educational Leadership*, 41(8), 19-27.
- White, I. R., & Thomas, J. (2005). Standardized mean differences in individually-randomized and cluster-randomized trials, with applications to meta-analysis. *Clinical Trials*, 2(2), 141-151.
- Wigelsworth, M., Lendrum, A., Oldfield, J., Scott, A., Ten-Bokkel, I., Tate, K., & Emery, C. (2016) The impact of trial stage, developer involvement and international transferability on universal social and emotional learning programmes's outcomes: A meta-analysis *Cambridge Journal of Education* (Online Early).
- Willett, J.B., Yamashita, J.J. & R.D. Anderson (1983). A Meta-Analysis of Instructional Systems Applied in Science Teaching. *Journal of Research in Science Teaching* 20(5):405-17.
- Wiliam, D. (2010). 'Standardized Testing and School Accountability'. *Educational Psychologist*, 45: 2, 107-122.
- Wolf, M. M. (1978). Social validity: The case for subjective measurement or how applied behavior analysis is finding its heart. *Journal of applied behavior analysis*, 11(2), 203.