Meta-synthesis and comparative meta-analysis of education research findings: some risks and benefits

Steve Higgins
Durham University

s.e.higgins@durham.ac.uk

Abstract

Meta-analysis, or quantitative synthesis, is the statistical combination of research findings. It can identify whether an intervention or approach, on balance, is effective or not, and can explain variation in findings by identifying patterns associated with larger or smaller effects across studies. It is now more widely applied in medicine and psychology, even though the term was first used in education, and the underpinning statistical ideas date back seventy years or so. This review traces the development of meta-analysis in education and the history of meta-meta-analysis or 'meta-synthesis' in more detail, where the temptation is not just to draw conclusions about similar studies, but to aggregate findings across meta-analyses to understand the relative benefits of different approaches on educational outcomes. A final section presents the rationale for the Sutton-Trust – Education Endowment Foundation Teaching and Learning Toolkit, which aims to present accurate and accessible findings from research studies which are sufficiently applicable to inform professional decision-making and action in schools, as an example of a current 'meta-synthesis' for education.

Keywords: meta-analysis, research synthesis, systematic review

Introduction

Often it is hard to tell from the result of a single piece of research whether or not something is a good idea. This is true in medicine, with one example being the use of different beta-blocker drugs where it took a number of studies to decide that such medication was effective at reducing further heart attacks (Yusuf et al. 1985). This is even more true in social science where it is never likely that one study, however large or robust, will be definitive, due to the variation in contexts and social conditions. We therefore need to be able to identify any trends in research findings by combining them to see what kind of patterns emerge. This review traces the development of meta-analysis in education and the history of meta-meta-analysis or 'meta-synthesis' in more detail, where the temptation is not just to draw conclusions about similar studies using a quantitative

synthesis, but to aggregate or compare findings across meta-analyses to understand the relative benefits of different approaches on educational outcomes. A final section presents the rationale for the Sutton-Trust – Education Endowment Foundation Toolkit which aims to present accurate and accessible findings from research studies which are sufficiently applicable to inform professional decision making and action in schools.

Meta-analysis is a technique used in reviewing and summarising the findings of different research studies and involves the statistical combination of their research findings. In this paper 'meta-analysis' refers to a quantitative synthesis which pools or aggregates findings, using agreed statistical techniques (Chalmers & Altman, 1995; Borenstein et al. 2009), from a series of studies which have been identified explicitly and rigorously, usually by systematic review techniques (Petticrew & Roberts, 2005; Higgins & Green, 2008). One aim of such a technique is to help in drawing conclusions about whether an intervention or approach, on balance, is effective or not. It also seeks to explain variation in research findings by identifying any patterns or significant associations with features of interventions associated with greater or smaller effects ('moderators'). So in understanding whether phonics is an effective approach for early reading there are a number of meta-analyses which have looked at the results of other studies and concluded, on balance that such approaches are effective (e.g. Ehri et al. 2001; Torgerson, Brooks and Hall, 2006; Jeynes, 2008). Each of these meta-analyses identifies an overall or 'pooled' effect (an 'effect size' of 0.41, 0.27 and 0.30 standard deviation units respectively). The precise estimate of effect depends on the detail of the review and meta-analytic procedures with their estimates varying due to the detail of the different research questions and precise inclusion criteria. Each meta-analysis also draws different conclusions about such things as the value of starting phonics at a younger age (Ehri et al. 2001) or the lack of evidence for synthetic over analytic approaches (Torgerson et al. 2006) or the robustness of the findings across studies of different quality (Jeynes, 2008). Different aspects of implementation matter between studies (Durlak & DuPre, 2008). Although meta-analysis is not

restricted to intervention research, this article focuses on studies with experimental designs which are often designed to answer efficacy or effectiveness questions. Meta-analysis is now more widely used in medicine and psychology, but the term was first coined for educational research, and the underpinning statistical ideas date back over a further seventy years or so. This review traces the development of meta-analysis in education and the history of 'meta-synthesis' or 'meta-meta-analysis' in more detail, where the temptation is not just to draw conclusions about similar studies using a quantitative synthesis, but to aggregate and compare findings across meta-analyses to understand the relative benefits of different approaches on educational outcomes.

Meta-analysis combines or 'pools' estimates from a range of studies and can therefore produce more widely applicable and generalisable inferences than would be possible from a single study. In addition, it can show whether the findings from similar studies vary more that would be predicted from their samples so that the causes of this variation ('heterogeneity') can be investigated using moderator analysis to see what features are associated with specific effects, such as the length of time pupils studied, or the importance of training and support, or the use of particular resources, by drawing on data from across the included studies and looking for correlations. 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 or by undertaking a narrative review, you may be misled into thinking that the evidence is less conclusive than if you combine these studies into a single meta-analysis to look at the overall pattern (see Cooper and Rosenthal (1980) for an empirical test of this).

A short history of the origins of meta-analysis

The history of the evolution of meta-analysis as a statistical technique to find more conclusive answers to research questions like this by combining findings is a fascinating one (Morton, 1997; O'Rourke, 2007). It involves a number of academic characters, crosses disciplines and took nearly 60 years from conception in the early 1900s to its birth and naming in the 1970s. Once meta-analysis emerged as fully formed technique, its use expanded rapidly in a number of fields, particularly medicine and psychology. A number of issues in the development of the approach provide some salutary warnings however for contemporary use.

A British mathematician, Karl Pearson, appears to have been the first to think of ways to combine numerical data from different studies. He is considered the pioneer of mathematical statistics and some of his techniques are still used today, such as the 'product-moment' correlation co-efficient and the chi-squared test. In 1904, in the second volume of the British Medical Journal, he published an analysis of the incidence and mortality rates of typhoid fever among soldiers in the British Army in India and South Africa. Results were available for soldiers who had volunteered for inoculation and other soldiers who had not volunteered. He wanted to know whether combining findings from a series of small studies would help answer the question about the effectiveness of inoculation, but was also curious about what was causing variation in findings, as vaccination sometimes appeared useful, but sometimes did not (see also Susser, 1977). These two aims of combining findings for greater certainty but also finding out what accounts for any variation are the core concepts of meta-analysis.

Pearson presented the results of his work in a table where each study was represented by its own row showing the measure of effect, together with a measure of the within-study uncertainty. The last row gives an average correlation as an average or pooled estimate of the effect. Again this

prefigures the development of the 'forest' plot which is now the classic way to present the findings of a meta-analysis. We can certainly recognise Pearson's table as 'meta-analysis', though without specifically being named as such, and it does not contain estimate of the uncertainty of this pooled effect, if we are being pedantic. This was not enough for Pearson however. He also wanted to understand what caused the variation or heterogeneity in the effectiveness of inoculation, so he looked for possible explanations, such as that soldiers who had volunteered for inoculation might have been at lower initial risk of developing the disease. He considered that this uncertainty might be answered by further analysis, but also proposed 'an experimental inquiry' by inoculating every second volunteer as a randomised trial. Although it took another sixty years to develop more formally the techniques that he considered, Pearson set out a visionary approach for the use of randomised controlled trials and to aggregate findings across such trials (see also Fraser et al., 2007).

The following table gives the results of calculating the correlation coefficients of the tables in Appendix B:

INOCULATION AGAINST ENTERIC FEVER:

Correlation between Immunity and Inoculation.

I. Hospital Staffs	+ 0.373	±	0.021
II. Ladysmith Garrison	+ 0.445	±	0.017
III. Methuen's Column	+ 0.191	±	0.026
IV. Single Regiments	+ 0.021	±	0.033
V. Army in India	+ 0.100	±	0.013
Mean value	+ 0.226		

Correlation between Mortality and Inoculation.

VI. Hospital Staffs	+ 0.307	±	0.128
VII. Ladysmith Garrison	− 0.010	±	0.081
VIII. Single Regiments	+ 0.300	±	0.093
IX. Special Hospitals	+ 0.119	±	0.022
X. Various military Hospitals ...	+ 0.194	±	0.022
XI. Army in India	+ 0.248	±	0.050
Mean value	+ 0.193		

If we except IV and VII, the values of the correlations are at least twice (in the very sparse data of VI) and generally four, five, or more times their probable errors. From this standpoint we might say that they are all significant, but we are at once struck with the extreme irregularity and the lowness of the values reached. They are absolutely incomparable with the fairly steady and large values of the vaccination correlations obtained for different epidemics and towns. The effect of enteric inoculation is evidently largely influenced by difference of environment or of treatment.

Figure 1: Pearson's findings (Pearson, 1904: p 1244)

Another pioneer of statistics in scientific inquiry in the early part of the twentieth century was Ronald Fisher. He built on Pearson's work on correlation with the development of techniques like analysis of variance (ANOVA) in his work on agriculture at Rothamsted Experimental Station in Hertfordshire in England. In his 1935 textbook, he gives an example of analysis of multiple studies, identifying the probable and real concern that fertilizer effects will vary by year and by location. Fisher's influence on the development meta-analysis was important. He laid the groundwork for the analysis of multiple studies in his final book, *Statistical Methods and Scientific Inference*, published in 1956. In this he encouraged researchers to summarize their findings clearly and rigorously which would make the comparison and aggregation of cumulative estimates across studies easier, almost the same as if all of the data were available for re-analysis (Box, 1978).

A false start?

One of the major challenges faced by researchers in the last century was the ever-increasing quantity of published research in almost all research fields. This needed new methods to synthesize and summarise the accumulating results. The first systematic attempt at this came in an unlikely area of psychology. In 1940, Joseph Gaither Pratt and Joseph Banks Rhine, based at Duke University in Durham, North Carolina in the USA, published a book with some of their colleagues which reviewed 145 reports of Extra-Sensory Perception (ESP) experiments which had been published between 1882 and 1939. In Chapter 4 they discuss and summarise "the full range of available trials made to test the ESP hypothesis" from experimental studies. They took a critical look at the data from the point of view of the design and conditions of the experiments to address general issues of whether the findings could have happened by chance, or by normal perception of sight, touch and hearing by those involved in the experiment. They also considered experimental errors and clustered results from similar experiments for sub-group analysis. For its time, it was a

rigorous and clearly documented attempt to address the question of whether there was "unequivocal evidence for the occurrence of ESP". One of the approaches included in their review to demonstrate the robustness of the analysis was an estimate of the influence of unpublished papers on the overall pooled effect. Today this is often called the 'file-drawer' problem of publication bias on the basis that non-significant studies were thought to languish in a filing cabinet drawer, resulting in bias in the overall conclusion resulting from the omission of non-significant or negative findings. The effect of published studies, and lack of publication of what are seen as unsuccessful studies remains a problem today. The work of Pratt and Rhine also sets a further challenge for advocates of meta-analysis. They concluded, on the basis of the evidence they summarised, that ESP *did* exist. Today we might criticise their analysis for a series of problems with the underlying studies, for some this was the design and conditions of the experiment, for others there was both publication bias and probably some selective reporting. The most important reason we are sceptical, however, is that, despite attempts, the findings have not been replicable. This suggests that the significant findings in the studies they accumulated had occurred by chance (as you would expect for one in twenty studies at the 95% level), poor design, error, or even cheating. Ben Goldacre, in both Bad Science (2008) and Bad Pharma (2012) argues that some of these issues, particularly selective publication and publication bias, are still a problem in medical research today. Scepticism about the conclusions reached by Pratt and Rhine may well have influenced people's views about the soundness of the methods that they used and this may have slowed the adoption of the systematic accumulation and analysis of evidence more widely.

The reason that this ESP 'meta-analysis' is so important is that it reminds us of three things. First, that the picture created by accumulating research findings over time may systematically present an incomplete or biased view. The reasons for this may be complex and come from a number of sources. Even a rigorous analysis, by the standards of the time, may not uncover this bias. The second challenge is the position of the researchers in relation to the evidence and

argument. The rigour and transparency of Pratt and Rhine's analysis and the quality and the care with which they undertook their experiments makes it unlikely, in my view, that they were knowingly trying to deceive their audience. However they were advocates of ESP and of parapsychology more widely and were trying to convince the scientific community that ESP was a genuine phenomenon. This may have influenced their decisions in the choices they made about their review, introducing researcher bias, at least at the unconscious level. The third issue is the importance of replication, particularly independent replication. This is particularly important in areas like education where the nature of an intervention is rather more difficult to specify and repeat than the contents of an inoculation or formulation of a tablet, or even the design of a card-guessing experiment in ESP. Lack of replication remains a problem today (Open Science Collaboration, 2015).

There is an additional challenge in accumulating research evidence over time which affects some aspects of medicine and some of the social processes and interactions such as those involved in education. Typhoid or enteric fever is usually caused through contamination of drinking water with a particular variety of salmonella bacteria. Better sanitation has reduced the incidence of the disease. We understand considerably more now about the different strains of typhoid and the problem of asymptomatic carriers, like 'Typhoid Mary', who was first identified three years after Pearson's inoculation study was published. We also know through a systematic review and meta-analysis that today's vaccines are between 51% to 55% effective (Fraser et al., 2007; Anwar et al. 2014). Typhoid, however, is treated with antibiotics which influence the evolution of the bacterium, developing resistance and changing the response to vaccination. In education, we know little about how teaching practices evolve and the development of 'pedagogical resistance' to intervention, but we do know that the impact of interventions varies considerably according to context and have only just started to understand the factors that are associated with this variation. Through the 1950s and 1960s, a number of researchers tried parametric statistical techniques to

summarize results from different studies which were reported as correlations or percentages (Kulik & Kulik, 1989). Underwood (1957) adopted percentages to summarise data on interference and lack of retention of information in studies of memory and forgetting. Erlenmeyer-Kimling and Jarvik (1963) applied statistical techniques to 99 correlation coefficients representing degree of similarity in intelligence of individuals who were related. Kulik and Kulik (1989) identified how these approaches to quantitative review are similar to a meta-analysis in three ways. First, a number of studies are summarised from the literature; second, they report findings from similar studies on a common scale, and finally they tried to explain some of the variation in study results according to specific features in the studies they reviewed.

To summarise this section it is clear that by the last quarter of the 20th century researchers were looking for ways to combine findings from similar studies to provide a more secure or more convincing answer to an overarching question. However, in doing this there were a number of challenges. First and foremost is the conceptual question about whether it makes sense to combine the studies. Or, to put it another way, what question can reasonably be answered by blending the results of different research inquiries? Is there likely to be a pattern of bias in the results? How systematically have these studies accumulated or how patchy is the evidence? Are there any gaps in this evidence which might lead to misleading conclusions? Are there any replications or independent evaluations where people have tried explicitly to see if the findings are repeatable?

The birth of meta-analysis

Meta-analysis, in its current form, was not first undertaken in medical research. In 1976, in his presidential address to the American Educational Research Association, Gene Glass coined the term ‘meta-analysis’ to refer to ‘the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings.’ With his colleague Mary Lee Smith, he had aggregated the findings from all of the psychotherapy outcome studies they could

find, in order to challenge the consensus, dominated by Hans Eysenk, that psychotherapy did not work. Eysenk was Professor of Psychology at the Institute of Psychiatry, King's College, University of London and had previously reviewed the effects of psychotherapy by combining percentage scores and computing statistics from this data (Eysenk, 1952). Smith and Glass' analysis showed that "the typical therapy trial raised the treatment group to a level about two-thirds of a standard deviation on average above untreated controls; the average person receiving therapy finished the experiment in a position that exceeded the 75th percentile in the control group on whatever outcome measure happened to be taken". This represents an effect size of 0.6, leaving no real doubts about the value of psychotherapy.

Gene Glass credits Robert Rosenthal for developing the underlying metric of "effect size" or the used of standard deviation units which are the measure used in most educational meta-analyses of impact studies that we now call "meta-analysis." In 1966, Rosenthal had published a book entitled *Experimenter Effects in Behavioral Research*, which contained calculations of a large number of standardized mean differences ("effect sizes") that he then compared across domains or conditions. Glass also acknowledges the impact of Benjamin Bloom on educational thinking more widely, with similarities between his approach in presenting aggregated graphs of correlation coefficients in his 1964 book '*The Stability of Human Characteristics'* though his contribution to meta-analysis is rarely noted. Indeed, his formulation of the search for teaching approaches as effective as one-to-one tuition was formulated in standard deviation units as "the two sigma problem" (Bloom, 1984). From the late 1970s through the 1980s the number of meta-analyses and methodological papers about the statistical techniques burgeoned with a consensus about the best approach emerging only slowly.

Of course not all researchers accepted the approach: Eysenk, for example, called it 'mega-silliness' and expected it to be a short-lived distraction. Others criticised it because of the comparability issue claiming it combined "apples and oranges" (see Slavin 1984 for a discussion

of this). Glass' (2000) riposte is telling "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." This gets to the heart of the issue. If your answer is about apples and oranges, you need to know this and not believe it is just about apples, or oranges, or even all kinds of fruit. Any inference is directly related to what the meta-analysis contains. Understanding this is not always straightforward as it seems. In education we use a large number of general terms. Take 'homework' for example, we are all confident we know what the word means, so identifying studies of homework should not be problematic. However if you want to know if homework helps children to learn better you have to decide what counts as homework. Is a class of five-year olds taking a books home to practice reading with their parents homework? What about learning spellings at home? How about 'homework clubs" where children do their 'homework' at school, *before* they go home? What about reading in preparation for a lesson? Or learning multiplication facts in preparation for a test? Or completing coursework for an examination at home? Are these really all the same thing? Suppose you combine all of these types of studies and conclude that children who are assigned homework compared with similar children who are not given any do better, on average, in tests of their learning, what does this tell you? It does not mean that homework is always effective. It indicates that when people have experimented to find out if being given (and doing) homework helps, the broad answer is yes. It does not tell you that it will work in every instance in the future. If you know what kind of homework studies have been included, you will have a reasonably clear idea about what might be likely to work in your school or for your own children. It would also be helpful to know if there are any other implications in the research. Do reading studies show greater effects? What about the age of the children? What about frequency and regularity of homework set? A general answer is useful, but is then is only the starting point for further questions and investigation.

This combination of studies is also a problem in another way which Eysenk (1978: 515) and others, (such as Slavin, 1984) also took exception to. Eysenk was concerned at the lack of quality control in what he termed the problem of "garbage in – garbage out" (p 517). He thought that only high quality studies should be included as less rigorous studies might introduce bias into the results. The challenge here is in deciding what counts as a high quality study. Do you only include those with a rigorous design, such those with a randomised design which can demonstrate that they have effectively controlled for selection bias and reduced the risk of possible confounding variables, and where the sample size meets a specific threshold and where the quality of the research process is adequately documented? Or do you include data from a wider range of experimental comparisons and check to what extent features of the research design and the quality of the reporting explain variation in the pooled effect? This is not an easy decision. If you have sufficient studies you may be able to afford to reduce the quantity to assure the quality. However you also have to be careful here that this does not introduce other kinds of bias. , Research and reporting quality is not necessarily related to the actual effectiveness of an intervention . A researcher may have had a great idea about how to improve reading, but not been very good at evaluating and describing this. If the findings are consistent with more rigorous studies, additional data on what causes variation may usefully add to moderator analyses, which are often underpowered (Valentine et al. 2009), . Discarding the data without investigating or establishing this seems premature, though it is important to have clear criteria for inclusion and to have a specification or protocol in advance for checking the relationship between study features and quality, particularly in a more inclusive study, so as to prevent data dredging (see, for example, Moher et al. 2009).

The development of meta-analysis as an essential technique to summarise and synthesise research findings across medical studies began a few years after Glass first coined the term in the 1970s. Particularly important here was an innovative randomised trial conducted by Peter Elwood,

Archie Cochrane and others to find out whether taking aspirin lowered the risk of further heart attacks and reduced the mortality rates associated with these (O'Rourke, 2007). The overall results suggested there was a benefit but were not conclusive. So, over the next few years, Elwood and Cochrane collected and combined results using meta-analysis as additional findings from new trials were reported. This aggregation left little room for doubt that taking aspirin after a heart attack was beneficial, and the findings were presented in 1980 in an anonymous Lancet editorial, but penned by the British medical statistician Richard Peto (O'Rourke, 2007).

Peto and his colleagues used a further example with data from randomised trials of beta-blockade following heart attack to encourage medical practitioners to consider aggregation of data from randomised trials systematically, and to combine quantitative estimates of the effects of comparable medical treatments. These developments started a debate, similar to one developing in the social sciences, about the best way mathematically to 'average' or estimate the aggregated findings in a single figure estimating the overall impact of different interventions or treatments, as an effect size. Peto argued for estimating a weighted average of the different effects when the effects were not identical, so treating the meta-analysis as if it was a larger single study: the 'fixed' effect model (based on inverse variance where studies with smaller variance (standard error) contribute more than studies with larger variance). A more conservative approach is to think of each study being a slightly different version of an intervention, with its own random variation which needs to be accounted for and where both the variation *within* studies and *between* studies is taken into account). This kind of 'random effects' model was advocated by meta-analysis pioneers like Larry Hedges (1983) and was developed and promoted to medical researchers by Rebecca DerSimonian and Nan Laird, who also provided simple approximate formulas for Cochran's formal random effects model (DerSimonian & Laird, 1986). For a presentation of the arguments for fixed effect and random effects models and methods of calculation, see Borenstein et al. (2009).

Similar to developments in social sciences a few years earlier, these developments in medical research led to an explosion of research and publication presenting empirical findings, developing the methods and promoting the work to practitioners (O'Rourke, 2007). One key difference in clinical work was the focus on quality assessment compared with work in social sciences. As the technical literature on meta-analysis expanded, the importance of being confident about the data included in a meta-analysis received increasing attention. So using systematic approaches to identify and collect relevant information so as to reduce bias in reviews became more and more important. The greater precision of meta-analysis findings statistically are not of much use if they are very accurate, but misleading! Even today this problem has not been solved in medicine.

The range of terminology used in medical reviews was confusing and towards the end of the last century, Chalmers and Altman (1995) argued that the term 'meta-analysis' should be applied only to quantitative synthesis, as adopted in this article. One of the main reasons for the rapid growth of meta-analysis is that it tackles one of they key challenges in reviewing research in that it can cope with a large number of studies which can overwhelm other approaches (Chan & Arvey, 2012). In addition, 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 the wealth of studies available 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, so techniques to aggregate and build up understanding of a field in terms of the impact of different interventions or approaches and what might explain variation so as to propose further research and test theories and hypotheses are invaluable. In fields like psychology and medicine meta-analysis is relatively uncontroversial as a synthesis method with nearly 40 years of development of the various principles and methods involved, despite its initial origins in education.

There are limitations and perhaps the most important is the assumption that the data from evaluations are equivalent or at least comparable 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 to catch up with their peers is different from asking whether phonics approaches are the best approach for beginning readers (when compared with other approaches to teaching reading). 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 outcomes (such as decoding words or comprehending sentences) would be essential.

Another limitation is the so-called 'file-drawer' problem where studies with null or negative effects are not reported. If a field is systematically missing these 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 which are in schools and both the class they are in the school they attend may influence the impact of different approaches. Analysis needs to take this into account or the effects may be overestimated (Hedges & Olkin, 2014).

However, 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 offer the best source of information to address cumulative questions about effects in

different areas of educational research, as well as in understanding what might explain differences in effects. For these questions the technique is relatively uncontroversial.

Meta-analysis and meta-synthesis

It is also tempting to look at results across different kinds of studies with a common population, so to provide more general or comparative inferences. A comparative meta-analysis, in this sense, compares effects between different kinds of interventions of approaches studies or between meta-analyses. It aims to answer the question "Does X work better than Y?", rather than the more specific "Does X work?". In this sense a comparative meta-analysis is a single meta-analysis where more than one intervention or approach is included to identify which is more effective. A comparative meta-synthesis is defined as where inferences are drawn by comparing findings across meta-analysis. This comparative approach is, of course, vulnerable to the classic 'apples and oranges' criticism, which argues that you can't really make a sensible comparison between different kinds of things. The key point here, as noted above, is that any inferences that you can make are at the level of aggregation of the synthesis. In studying apples and oranges together you can consider what they tell you about fruit, similar characteristics (such as seeds for reproduction, developing from the female parts of the flower, protecting the seeds, developing an edible coating to aid dispersal) and variation (relative size, the nature of the fleshy covering, the detail of the different seeds and the like), but not conclusions specific to oranges (such as segmentation of the fruit, juiciness of the flesh, oily skin and the like). Similarly in combining an analysis of different approaches to improving reading you can draw inferences about the effectiveness of, say, reciprocal teaching, compared with inference training (see for example, Pearson & Dole, 1987), assuming comparable populations and sufficiently similar interventions and designs. However you can't draw conclusions about impact on specific groups, if studies contain a wide range of

different samples of students. Another example is the teaching of writing for primary age pupils (Graham et al. 2012) where the meta-analysis indicates that teaching strategies, adding self-regulation to strategy instruction, teaching text structure, the use of creativity/imagery instruction and the teaching transcription skills are all important features of the effective improvement of writing, but not the specific components of such strategies or which aspects of creativity and use of imagery were beneficial.

A number of researchers have attempted to take meta-analysis a stage further than this, by synthesizing the results from a number of existing meta-analyses and producing what has been called a 'meta-meta-analysis' (Kazrin, Durac & Agteros, 1979), a 'mega-analysis' (Smith 1982), 'super-analysis' (Dillon, 1982), 'super-synthesis' or 'meta-synthesis' (e.g. Sipe & Curlette, 1997) to draw comparative inferences. This remains controversial in educational research and there is a clear separation of types within these studies. 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). Even here, there appears to be a qualitative difference and 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) in effect producing something might best be considered as a series of larger meta-analyses rather than a meta-synthesis. Others, though, make claims about broad and quite distinct educational areas by directly combining results from identified meta-analyses (e.g. Fraser et al. 1987; Hattie, 1992; Sipe & Curlette, 1997). In terms of the apples and oranges analogy, this is a little like asking which fruit is best for you, as a lot depends on what you mean by 'best' and how this is measured.

In the following section a number of these ‘meta-syntheses’ are reviewed to identify some key characteristics and limitations. The first was published just over 10 years after Gene Glass had set out a methodology to aggregate findings across studies, when a team involving Barry Fraser, Herbert Walberg and John Hattie undertook an extensive synthesis of evidence in which they summarised the findings from 226 meta-analyses (Fraser et al. 1987), incidentally indicating the rapid uptake of meta-analysis as a new technique. The main purpose of Fraser and colleagues’ synthesis was to report the findings as an empirical test of Walberg’s own productivity model, based mainly on quantitative syntheses of various bodies of past research. The work was published as a 100 page monograph in the *International Journal of Education Research*. A chapter presents a research syntheses of several thousand individual studies to identify aptitudinal, instructional, and environmental variables which influence student learning consistently and extensively. A further chapter focuses on meta-analyses of individual studies in science teaching and learning to identify the effects of contextual and transactional influences on science learning. Then the various salient features found in what were then current models of student learning were used to structure a synthesis of 134 meta-analyses of achievement outcomes and 92 meta-analyses of attitude outcomes. The results of this impressive synthesis of meta-analyses are then used to identify to what extent the empirical evidence supported Walberg’s nine-factor model of educational productivity.

Hattie has been a pioneer in this field and (1992) took this work a stage further by summarising the 134 meta-analyses which had been identified and reported earlier in Fraser et al. (1987). The synthesis consisted of 22,155 effect sizes computed from 7,827 primary studies which represented between five and 15 million students. His aim was to show how findings from more than 30 years of educational research indicated "the effects of innovation and schooling” but could also be used to “provide insights for future innovation”, as well as to resolve “contrary claims about the effectiveness of schooling, by demonstrating how different points of comparison are used by each

group" (p. 5). As a central focus in this analysis he introduced a "universal continuum" (p. 6) as a basis for this assessment, with a scale expressed in standard deviation units and with results from the meta-synthesis placed on this scale. Although the effect sizes were reported as correlations in Fraser et al. (1987), Hattie (1992) converted these into standardised mean differences, as Bloom had previously done for tutoring (1984). The average effect size across the meta-analyses was 0.40 (with a standard deviation of 0.13). The largest effect sizes were for those interventions providing feedback. The effect sizes for reinforcement, for remediation and feedback, and for mastery learning were 1.13, 0.65, and 0.50, respectively. The lowest effect sizes involved individualization (average effect size of 0.14), while programmed instruction, another approach to personalisation, yielded an average effect size of 0.18. Hattie (1992) identified three overall findings. First, innovation as deliberate attempt to improve the quality of learning can be related to improved achievement and is an underlying theme in the majority of these effects. Second, feedback is the most powerful single influence which improves achievement. Third, the least successful innovations were those that tried to individualize instruction. He saw this as important, as at the time pupils spent about two-thirds their time in school working on their own. In terms of the overall messages, Hattie was concerned that the underlying studies were of variable quality as well as involving different outcome measures. He also indicated that the synthesis did not suggest that effects on achievement are necessarily cumulative. His conclusion was about the value of the comparative information from such an approach and that the "continuum highlights the importance of assessing competing theories - that is, to compare various innovations" (1992, p. 11).

Five years later, Sipe and Curlette (1997) published a review which explicitly aimed to develop the methodology of summarization and meta-synthesis in terms of the evidence related to educational achievement. They undertook a systematic search for meta-analyses, applied rigorous inclusion criteria (excluding 324 of the 427 studies they had identified) and then described the

background, methodological and contextual characteristics of the 103 studies they included. They also undertook an explanatory analysis, similar to Fraser and colleagues (1987), in relation to Walberg's (1984) educational productivity theory. They estimated that fewer than 10% of the meta-analyses in their meta-synthesis overlapped with those in Fraser et al. (1987) and Hattie's (1992) review, because the earlier reviews had included meta-analyses from 1976 to 1985 while theirs included meta-analyses published between 1984 and 1993 (p. 648). (For more details of other early meta-syntheses and a critique of 18 such studies see Sipe & Curlette, 1997, p 597 – 612.)

In 1998 Robert Marzano published "A Theory-Based Meta-Analysis of Research on Instruction" in a report for the US government-funded Mid-Continent Regional Educational Laboratory (McREL). This synthesized research findings from more than 100 meta-analyses and other studies involving over 4,000 comparisons of experimental and control groups (Marzano, 1998). One of the main goals of the synthesis was to identify instructional strategies which should have the greatest likelihood of enhancing achievement for all pupils, across subject areas and age groups. The analysis revealed that there was considerable variability across the studies in terms of how the instructional strategies were defined and how their use in the classroom was described. Marzano is critical of what he describes as the "brand name" approach in meta-synthesis where very broad categories of educational approaches represent the popular labels which are sometimes given to quite complex interventions with a range of salient features or 'active ingredients'. As an example, he cites a meta-analysis conducted by Athappilly, Smidchens, and Kofel (1983) and included in Fraser et al. (1987) where one 'brand name' used is "modern math" but where a number of different components contribute to an aggregated effect size. Features in "modern math" such as the 'use of manipulatives' which had an effect size of .51, could be distinguished from 'direct instruction in concepts and principles' which had an effect size of .35 and was different from 'use of an inquiry approach' which had an effect size of only .04. Marzano argues that aggregating

these into a single 'brand' masks potentially important findings about instructional effectiveness and that more discrete categories are needed which are specific enough to provide guidance for teachers in terms of classroom practice. He describes the meta-synthesis as 'theory driven' in using four high level categories for learning based on *knowledge*, learning which involves the *cognitive* system, or the *metacognitive* system, and the *self-system*. Overall he draws conclusions at the level of specific instructional practices, but also at this more abstract or theoretical level. What makes Marzano's approach important is the overall coherence of the theoretical framework, whilst also striving to be practical at the level of applicability in the classroom.

In what can perhaps be seen as the culmination of his work to date, Hattie (2009) synthesized more than 800 meta-analyses, including those in his 1992 study, and came up with some interesting further findings in his book '*Visible Learning'*. As before, he concluded that most things in education 'work' as the average effect size is about 0.4 (a mean effect which had remained stable since 1992). He then uses this to provide a benchmark for what works above this 'hinge' point. The fact that this figure has remained stable when aggregating quantitative data in education is interesting, but as well as representing the typical difference of bringing about change in education as Hattie argues, it may also be taken to show that differences of just less that half a standard deviation, on average, are of educational interest and worth investigating. It is less clear that the same things have the same effect over time. Older studies of peer tutoring, for example, tend to have larger effect sizes, but whether this is the result of lower evaluation quality, publication bias, allocation bias, researcher bias, or genuinely reflects a change in the counterfactual conditions so larger effects are harder to achieve (see Lemons et al., 2014 for a discussion of the problem of the changing counterfactual).

Whilst Hattie's impressive scholarship is extremely valuable in putting findings together to provide a large-scale landscape of all the quantitative findings from educational attainment amenable to meta-analysis, it is not without its critics (Snook et al, 2009; Terhart, 2011; Higgins

& Simpson, 2011). The key assumption is that the research represented in the meta-synthesis is sufficiently evenly distributed by type and population that any differences which emerge represent differences in the educational themes, rather than differences in the research methods, measurements and target populations. As noted above, combining effect sizes of different kinds is risky. Intervention effects (improvement relative to a comparison or control group) should be distinguished from maturational differences for the students (single group designs). The design of the former takes growth into account between the pre- and post-test, the latter *only* accounts for growth over time. Correlational effects, such as the relationship between, say, homework and school performance (or, to be precise, looking at pupils who do different amounts of homework and comparing this with how well they do on tests of attainment) are rather different from homework intervention effects (where the impact of homework for some is compared with no homework for others). The underlying distributions of educational attainment in these studies are likely to be of different kinds, so that unlike comparing fruit, it may be more like comparing an apple with something other than fruit (Higgins & Simpson, 2011). Effect size estimates in terms of standardised mean differences depend on the distribution of the individual scores as this forms the basis for the comparison.

Overall we can see a trend in educational research which drives the need for aggregation of research findings, partly due to the ever- increasing numbers of studies for review, but also partly to identify more secure messages through aggregation and synthesis. Once this becomes possible quantitatively through meta-analysis, a further level of comparative meta-analysis, or comparative meta-synthesis also become possible. This approach can address issues of the relative benefit of different approaches, though a number of challenges remain to be solved.

Advantages and Limitations of Meta-analysis and Meta-synthesis

All of the limitations of meta-analysis apply to meta-synthesis. Explicit inclusion criteria and a systematic search are essential so that what is included in any aggregation is clear. Whether you choose to be inclusive in identifying studies and then checking how much difference for methodological rigour makes so as to maximise possible data for analysis, or whether you set a high quality threshold for inclusion to ensure rigour and internal validity in the component studies is perhaps less important, though transparency is essential. "Garbage in" is still likely to result in "garbage out" as Eysenk noted. The classic apples and oranges problem is also particularly relevant, but again the answer is similar to the one about comparing fruit. Provided we remember that any findings or "answer" applies at the level of the aggregation of the meta-synthesis, then the approach and any inferences may be warranted. So for example, if we ask whether some approaches to improving children's comprehension are more effective than others, then we can compare the findings from different meta-analyses, assuming they have both included similar studies (such as on typically developing or non-typically developing populations of pupils, with similar sample sizes, using similar outcome measures). If we want to know whether the impact of feedback remains higher than individualised approaches to learning in recent studies, we could look at the findings from different meta-analyses and evaluate whether the inclusion criteria for each have produced a sufficiently comparable set of studies to warrant a clear conclusion.

Although there are limitations to the application of comparative quantitative synthesis (both comparative meta-analysis and comparative meta-synthesis as defined above) in this way, the data from meta-analysis offer the best source of information to try to answer comparative questions between different areas of educational research. It is hard to compare areas 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. Meta-analysis certainly helps to identify which researched approaches have made, on average, the most difference, in terms of effect size, on tested achievement of students in a measure, say, of reading comprehension or another area of educational achievement. 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 more likely to find it beneficial investigate meta-cognitive approaches to improve learning, rather than focus 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 students) 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 teaching) or, indeed, don't work well in both fields (such as individualised instruction). The argument here is that so long as you are aware of the limits of the inferences drawn, then the approach has value. In medicine, developments such as network meta-analysis (see for example, Mills et al. 2013) aim to develop a more systematic comparative framework by analyzing both direct comparisons of interventions within randomised controlled trials and indirect comparisons across trials based on a common comparison (such as a standard treatment or placebo). In education this methodology has yet to be developed so more basic inferences provide the best

evidence we have, especially where we have no studies providing direct comparisons. As Fraser and colleagues (1987) argued nearly 30 years ago 1987:

> "Effect sizes permit a rough calibration of comparisons across tests, contexts, subjects, and other characteristics of studies. The estimates, however, are affected by the variances in the groups, the reliabilities of the outcome measures, the match of curriculum with outcome measures, and a host of other factors whose influences in some cases can be estimated specifically or generally. Although effect sizes are subject to distortions, they are the only explicit means of comparing the sizes of effects in primary research that employ various outcome measures on non-uniform groups."
>
> (pp.151-152)

The Sutton Trust- Education Endowment Foundation Teaching and Learning Toolkit

The assumptions from these early studies which aimed to provide comparative inferences across different areas of education research influenced the thinking about the design of the Sutton Trust- Education Endowment Foundation Teaching and Learning Toolkit ('Toolkit'), a web-based resource for practitioners (https://educationendowmentfoundation.org.uk/toolkit/) which uses meta-synthesis as the basis for its quantitative comparisons of impact on educational attainment. The aim of the synthesis is to provide advice and guidance for practitioners who are often interested in the relative benefit of different educational approaches as well as the detail of how to adopt or implement a specific approach (Cordingley, 2008). The initial work drew on a database of educational meta-analyses of intervention findings in education compiled as part of an ESRC Researcher Development Initiative *(Training in the Quantitative Synthesis of Intervention Research Findings in Education and the Social Science*) between 2008 and 2011. A further collaboration with the Sutton Trust enabled a the development of series of summaries of research which could help schools decide how to allocate any additional funding for the recently announced Pupil Premium initiative in England (Higgins, Kokotsaki and Coe, 2011). The analysis included an estimate of impact which was based on the effect size converted into months' progress,

together with an analysis of cost or the additional financial outlay for schools, together with an evaluation of the extent and robustness of the evidence. This was all summarised in a 'Which?-style' consumer guide to the evidence base (Figure 2). These cost estimates, though crude, were a unique feature of the initial Toolkit. The estimates draw on the most rigorous estimates available in a descending order of priority. First, meta-analyses of randomised trials and well-controlled experiments where any variation in effect, or heterogeneity, is explored and if possible explained. A quality rating ('padlocks' for security) also includes consistency of effects between meta-analyses to achieve the highest rating. If this kind of evidence is not available an estimate is made based on other quantitative data, such as correlational studies or from single studies. The reason for this is that the Toolkit aims to provide the best estimate available in a particular area, with an estimate of the robustness of the evidence, rather than only report where the evidence is robust and secure. The Toolkit is therefore distinctive in that it aims to present a summary of evidence in area where research is sparse or even lacking, at least in terms of causal inference, such as about school uniform or performance pay, but where policy and practice is influenced by people's assumptions about effective practice.

Toolkit to improve learning: summary overview

Approach	*Potential gain*[1]	*Cost*	*Applicability*	*Evidence estimate*	*Overall cost benefit*
Effective feedback	+ 9 months	££	Pri, Sec Maths Eng Sci	☆☆☆	Very high impact for low cost
Meta-cognition and self-regulation strategies.	+ 8 months	££	Pri, Sec, Eng Maths Sci	☆☆☆☆	High impact for low cost
Peer tutoring/ peer-assisted learning	+ 6 months	££	Pri, Sec Maths Eng	☆☆☆☆	High impact for low cost
Early intervention	+ 6 months	£££££	Pri, Maths Eng	☆☆☆☆	High impact for very high cost
One-to-one tutoring	+ 5 months	£££££	Pri, Sec Maths Eng	☆☆☆☆	Moderate impact for very high cost
Homework	+ 5 months	£	Pri, Sec Maths Eng Sci	☆☆☆	Moderate impact for very low cost
ICT	+ 4 months	££££	Pri, 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 2: The 2011 Pupil Premium Toolkit

With support from the recently formed Education Endowment Foundation, these summaries were conceptualised as a series of integrated 'umbrella reviews' (Grant & Booth, 2009) which could provide a rigorous but accessible summary with a common methodology across the different areas of educational policy, practice and research. The accessibility is a key feature as engagement with research evidence is a significant challenge (Cordingley, 2008; Hemsley-Brown & Sharp, 2004). The level of presentation needs to be sufficiently familiar to encourage acceptance, but sufficiently challenging to promote deeper engagement and support changes in practice. The apparent simplicity at the top level can be deceptive as the messages in any specific area are rarely straightforward, so successive levels of detail aim to support deeper engagement. The Toolkit is therefore presented in an accessible summary form at the surface level (https://educationendowmentfoundation.org.uk/toolkit/toolkit-a-z/) but with further detail on each area available, right through to the effect sizes and abstracts of the meta-analyses and other studies used in its compilation and synthesis. A technical appendix also sets out the rationale and detail of the effect size calculations and conversions (for more details about the Toolkit approach, the systematic search criteria, evidence estimates and synthesis see Katsipataki and Higgins, 2016). This aims to ensure the synthesis is accurate, with the methods and assumptions made transparent.

Figure 3: The 2015 Pupil Premium Toolkit

Some of the inspiration for the Toolkit comes from Hattie's (1992/2009) work in producing a comparative map of educational research findings, but some of the methodology from others such as Sipe and Curlette (1996) in having common inclusion criteria (such as a focus on school age pupils and relying where possible on intervention research with warrant for causal inference) and a systematic and transparent search strategy for meta-analyses, combined with Marzano's (1998) goal of practical utility. The overall estimate of effect is still crude (an estimate of months' progress), but this is probably realistic in terms of the comparability of the data and the limitations in the precision currently possible from pooling estimates across meta-analyses. Some areas of the Toolkit have relatively consistent findings, such as meta-cognition and self-regulation or phonics, others have varying estimates from the different meta-analyses, such as parental involvement or behavioural interventions.

Looking forward: what works or what's worked?

If we accept that meta-analysis is an important tool in accumulating research about educational interventions then we need to move forward more systematically than at present. Its potential has been acknowledged for more than 30 years in education (see Carol Fitz-Gibbon's articles in the British Educational Research Journal in 1984 and 1985 for example: Fitz-Gibbon, 1984; 1985). For some questions individual meta-analyses will of course be sufficient, others may require comparative inference.. As indicated above, the next steps in ensuring greater accuracy are to explore methods to improve the rigour of these comparisons such as by developing network meta-analytic approaches (Cipriani et al. 2013), or creating a set of meta-analyses using the same inclusion criteria, to ensure the findings can be more directly compared across meta-analyses, a cumulative meta-synthesis. Although we have made progress in understanding what causes systematic variation (such as the age of pupils (Bloom et al. 2008) or outcome type (Hattie, Biggs and Purdie, 1996) which can be understood in relation to the measurement of effect sizes. Other

kinds of systematic variation have been identified, but the causes are less well understood such as sample size (Slavin & Smith, 2009; Cheung & Slavin, 2015) where larger studies report smaller effects (a correlation of -0.28), though whether this is publication bias (Kühberger et al. 2015) or the impact of trial stage and type with pilot studies reporting higher effect sizes (Wigelsworth et al. 2016) or other aspects of 'super-realization bias' (Cronbach et al. 1980) is not yet well understood.

The variation in impact across all of the areas of the Toolkit and also found in summaries such as *Visible Learning* (Hattie, 2009) and noted by other researchers such as Marzano (1988) suggests that meta-synthesis will never provide a precise prediction of what will be effective in any future application of research findings to a new context. Such synthesis cannot therefore provide definitive claims as to "what works" in education. Rather it is an attempt to provide the best possible estimate of what is likely to be beneficial based on existing evidence. Such studies summarise what has worked as an indication or perhaps a 'best bet' for what might work in the future, but where the spread of findings should be taken into account as well as the average. This indicates that findings from meta-synthesis should be used to inform practice and decisions about how to improve outcomes for learners, but the variation in findings in education and the lack of precision in the aggregation process means that applicability of this information to a new context is going to be a probability rather than a certainty, so is always likely to need active enquiry and evaluation to ensure it succeeds in achieving the desired effects. This requires professional judgement and commitment to engaging with evidence on the part of the practitioner, but also a disposition to interpret, challenge and test particular findings to ensure they are useful. Overall the Toolkit aims to provide accessible and applicable information by summarising research evidence from meta-analysis which is as accurate as current methods allow, but also sufficiently actionable to be useful in the classroom. There are certainly limitations to this approach, particularly in the balancing of the accessibility of the information, with its accuracy in drawing on research evidence, but also in providing sufficient detail and specificity to support action. The main

argument is favour of such an approach is that this is the best option that we currently have to provide such indications. The alternative is to ignore or discard such findings from previous research as incommensurable. The Education Endowment Foundation is using the Toolkit as a guide to commission further research (100 randomised trials between 2011 and 2015 involving over 500,000 pupils) where the evidence will both test the Toolkit findings and feed back into its evidence-based to increase its robustness. This approach may create a database of findings where greater comparability across context and outcomes will help to identify more precisely the causes of variation in research findings, or at least indicate what progress is possible in this area.

References

Athappilly, K., Smidchens, U., & Kofel, J. W. (1983). A computer-based meta-analysis of the effects of modern mathematics in comparison with traditional mathematics. *Educational Evaluation and Policy Analysis*, 5 (4) pp. 485-493. http://www.jstor.org/stable/1164053

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. Educational Researcher, 4-16. http://www.jstor.org/stable/1175554

Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness*, 1(4), 289-328.

Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2009). *Introduction to Meta-analysis* London: John Wiley & Sons, Ltd.

Chalmers, I. & Altman, D.G. (1995) *Systematic Reviews.* London: BMJ Publications.

Chalmers, I., Hedges L.V. & Cooper, H. (2002). A brief history of research synthesis. *Evaluation and the Health Professions* 25: 12-37. http://dx.doi.org/10.1037/0033-2909.87.3.442

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

Cooper, H. M., & Rosenthal, R. (1980). Statistical versus traditional procedures for summarizing research findings. *Psychological Bulletin*, *87*(3), 442. http://dx.doi.org/10.1037/0033-2909.87.3.442

Cordingley, P. (2008). Research and evidence-informed practice: focusing on practice and practitioners. *Cambridge Journal of Education*, 38(1), 37-52. http://dx.doi.org/10.1080/03057640801889964

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.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3), 177-188. http://dx.doi.org/10.1016/0197-2456(86)90046-2

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. American Journal of Community Psychology, 41(3-4), 327-350. http://dx.doi.org/10.1007/s10464-008-9165-0

Egger, M., Smith, G. D., & Phillips, A. N. (1997). Meta-analysis: principles and procedures. BMJ: *British Medical Journal*, 315(7121), 1533. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2127925/pdf/9432252.pdf

Ehri, C.L., Nunes, S.R., Stahl, S.A., & Willows, D.M. (2001). Systematic Phonics Instruction Helps Students Learn to Read: Evidence from the National Reading Panel's Meta-Analysis. *Review of Educational Research*, 71, (3) 393-447. http://dx.doi.org/10.3102/00346543071003393

Fisher, R.A. (1935) *The Design of Experiments*. Edinburgh: Oliver and Boyd.

Fisher, R.A. (1956) *Statistical Methods and Scientific Inference*. Edinburgh: Oliver and Boyd.

Fitz-Gibbon, C. T. (1984). Meta-analysis: an explication. *British Educational Research Journal*, 10(2), 135-144. http://dx.doi.org/10.1080/0141192840100202

Fitz-Gibbon, C. T. (1985). The implications of meta-analysis for educational research. *British Educational Research Journal*, 11(1), 45-49. http://dx.doi.org/10.1080/0141192850110105

Francis, G. (2012). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin & Review*, 19(2), 151-156. http://dx.doi.org/10.3758/s13423-012-0227-9

Anwar, E., Goldberg, E., Fraser, A., Acosta, C.J., Paul, M. & Leibovici L. (2014). *Vaccines for preventing typhoid fever (Review)*. Cochrane Database Systematic Reviews (1): CD001261. http://dx.doi.org/10.1002/14651858.CD001261.pub3 PMID 17636661.

Fraser, A., Paul, M., Goldberg, E., Acosta, C. J., & Leibovici, L. (2007). Typhoid fever vaccines: systematic review and meta-analysis of randomised controlled trials. Vaccine, 25(45), 7848-7857. http://dx.doi.org/10.1016/j.vaccine.2007.08.027

Glass, G. V. (2000). Meta-analysis at 25. http://www.gvglass.info/papers/meta25.html

Graham, S., McKeown, D., Kiuhara, S., & Harris, K. R. (2012). A meta-analysis of writing instruction for students in the elementary grades. *Journal of Educational Psychology*, 104(4), 879. http://dx.doi.org/10.1037/a0029185

Hattie, J. (1992). Measuring the Effects of Schooling. *Australian Journal of Education*, 36(1), 5-13.

Hattie, J., Biggs, J., & Purdie, N. (1996). Effects of learning skills interventions on student learning: A meta-analysis. *Review of Educational Research*, 66(2), 99-136. http://dx.doi.org/10.3102/00346543066002099

Hedges, L. V. (1983). A random effects model for effect sizes. *Psychological Bulletin*, 93(2), 388. http://dx.doi.org/10.1037/0033-2909.93.2.388

Higgins, J. P. & Green, S. (Eds.). (2008). *Cochrane handbook for systematic reviews of interventions*. Chichester, England: Wiley-Blackwell.

Higgins, S., & Simpson, A. (2011). Visible Learning: A Synthesis of over 800 Meta-Analyses Relating to Achievement. By John AC Hattie: Book Review *British Journal of Educational Studies*, 59(2), 197-201. http://dx.doi.org/10.1080/00071005.2011.584660

Hunt, M. (1997). *How science takes stock: The story of meta-analysis*. NY: Russell Sage Foundation.

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. http://dx.doi.org/10.1371/journal.pone.0105825

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 http://dx.doi.org/10.3102/0013189X14539189

Marzano, R. J. (1998). *A Theory-Based Meta-Analysis of Research on Instruction*. Aurora, CO Mid-Continent Regional Educational Lab (McREL). http://files.eric.ed.gov/fulltext/ED427087.pdf

Mills, E. J., Thorlund, K., & Ioannidis, J. P. (2013). Demystifying trial networks and network meta-analysis. *British Medical Journal*, 346: 2914. http://dx.doi.org/10.1136/bmj.f2914

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of Internal Medicine*, 151(4), 264-269. http://dx.doi.org/10.7326/0003-4819-151-4-200908180-00135

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), http://dx.doi.org/10.1126/science.aac4716

O'Rourke, K. (2007). An historical perspective on meta-analysis: dealing quantitatively with varying study results. *Journal of the Royal Society of Medicine*, 100(12), 579-582.

Pearson, K. (1904) Report On Certain Enteric Fever Inoculation Statistics *The British Medical Journal*, 2 (2288) (Nov. 5, 1904), pp. 1243-1246 http://www.jstor.org/stable/20282622

Pearson, P. D., & Dole, J. A. (1987). Explicit comprehension instruction: A review of research and a new conceptualization of instruction. *The Elementary School Journal*, 151-165. http://www.jstor.org/stable/1002039

Petticrew, M., & Roberts, H. (2005). *Systematic reviews in the social sciences: A practical guide*. Oxford: Blackwell.

Pratt, J. G., Smith, B. M., Rhine, J. B., Stuart, C. E., & Greenwood, J. A. (1940). *Extra-sensory perception after sixty years: A critical appraisal of the research in extra-sensory perception*. NY, US: Henry Holt and Company. http://dx.doi.org/10.1037/13598-000

Rosenthal R. (1966) *Experimenter Effects in Behavioral Research*. New York, NY: Appleton-Century-Crofts.

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. http://dx.doi.org/10.3102/0162373709352369

Susser, M. (1977). Judgment and causal inference: criteria in epidemiologic studies. *American Journal of Epidemiology*, 105(1), 1-15.

Terhart, E. (2011). Has John Hattie really found the holy grail of research on teaching? An extended review of Visible Learning. *Journal of Curriculum Studies*, 43(3), 425-438. http://dx.doi.org/10.1080/00220272.2011.576774

Torgerson, C., Brooks, G., & Hall, J. (2006). *A systematic review of the research literature on the use of phonics in the teaching of reading and spelling*. Nottingham: DfES Publications.

Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). How many studies do you need? A primer on statistical power for meta-analysis. *Journal of Educational and Behavioral Statistics*, 35(2), 215-247. http://dx.doi.org/10.3102/1076998609346961

Walberg, H. J. (1984). Improving the productivity of America's schools. *Educational Leadership*, 41(8), 19-27.

Wigelsworth, M., Lendrum, A., Oldfield, J., Scott, A., Ten-Bokkel, I., Tate, K., & Emery, C. (2016/ forthcoming) 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*

Yusuf, S., Peto, R., Lewis, J., Collins, R., & Sleight, P. (1985). Beta blockade during and after myocardial infarction: an overview of the randomised trials. *Progress in Cardiovascular Diseases*, 27(5), 335-371. http://dx.doi.org/10.1016/S0033-0620(85)80003-7

Context and Implications Document for: **Meta-synthesis and comparative meta-analysis of education research findings: some risks and benefits**

Steve Higgins
Durham University

This guide accompanies the following article:

Higgins, S.E. Meta-synthesis and comparative meta-analysis of education research findings: some risks and benefits, *Review of Education*, [DOI will be added by Wiley]

Author's Introduction

Often it is hard to tell from the result of a single piece of research whether or not an intervention or approach is a good idea. This is true in medicine, with one example being the use of different beta-blocker drugs where it took a number of studies to decide that such medication was effective at reducing further heart attacks. This is even more true in social science where it is never likely that one study, however large or robust, will be definitive, due to the variation in contexts and social conditions. We therefore need to be able to identify any trends in research findings by combining them to see what kind of patterns emerge so as to develop understanding of the potential impact of different approaches. This provides 'good bets' which can be tested in practice or evaluated through policy initiatives.

Implications for Policy

- Meta-analysis provides quantitative estimates of the impact of different interventions or approaches on educational outcomes. These may not be precise enough to guide policy definitively in education.
- The costs of interventions also need to be considered as the approach with the largest effect may not be a 'good bet' if it is very expensive or unreliable.
- Evaluation of policy initiatives in education could usefully add to the research evidence by testing the impact achieved to help improve the precision of these estimates over time and to help understand how variation in contexts might influence outcomes.

Implications for Practice

- Comparative estimates of the impact of different interventions can help practitioners decide which programmes or approaches may help in a particular context, but these should be used as a guide to support professional judgement as the variation in impact reported in meta-analyses is broad. A summary of effects can be found here: https://educationendowmentfoundation.org.uk/toolkit/
- Thinking about costs should consider not just the outlay by the school for different interventions or approaches, which can vary widely, but also aspects such as teacher time, or the use of existing resources where is it is the overall efficiency of resource use which should be considered.

- Most meta-analyses focus on tested attainment and pupils performance on standardised tests. Whilst these are a vital part of educational success, they are not the only outcome or achievement schools may consider important.

- In most circumstances a willingness to make new interventions or teaching approaches succeed is needed. It is also important to evaluate impact on learning critically as not all research findings will apply in a new context.