Part I

Understanding Meta-analysis and Meta-synthesis

This part forms the first half of the book and looks at meta-analysis, its rationale (Chapter 1) and history (Chapter 2). It moves on to describe the origins and development of meta-meta-analysis, or 'meta-synthesis' (Chapter 3), which sets the background to an overview of Sutton Trust – Education Endowment Foundation Teaching and Learning Toolkit (Chapter 4). The Toolkit is an online summary of the intervention research in education for teaching 3- to 18-year-olds, designed initially to make the findings from intervention research accessible to schools and teachers in England, but increasingly used internationally.

Each chapter is headed by a number of key questions which I aim to answer in the course of the respective chapter. The writing style is a combination of academic (with citations and references) and a more reflective approach. My aim is to present key issues in evidence synthesis in education as I see them and to explain how I understand them from a personal perspective.

Chapter 1: Why Meta-analysis?
What is meta-analysis?
Why do we need it?
What can we learn from it?
What are its limitations?
Chapter 2: A Brief History of Meta-analysis
What are the origins of meta-analysis?
What can we learn from its evolution and development?
Chapter 3: Meta-synthesis in Education: What Can We Compare?
What is reasonable to compare?
How much difference do interventions make?
Chapter 4: The Teaching and Learning Toolkit
What patterns of effects appear when comparing meta-analyses?
Are there general implications for teaching and learning?

What can we infer from specific areas such as feedback and metacognition?
What do meta-analyses of digital technology and learning styles tell us?
What challenges arise in using evidence in this way?

Chapter 2 draws heavily on the *Review of Education* (Volume 4.1: 31–53[1]) article, 'Meta-synthesis and Comparative Meta-analysis of Education Research Findings: Some Risks and Benefits'. Chapter 4 has been developed from the 2016 article, 'Communicating Comparative Findings from Meta-analysis in Educational Research: Some Examples and Suggestions', which I wrote with Maria Katsipataki for the *International Journal of Research & Method in Education* (Volume 39.3: 237–254[2]), and I am grateful to the publishers John Wiley & Sons and Taylor & Francis (Informa UK Limited), respectively, for permission to reuse and develop this material.

[1] http://dx.doi.org/10.1002/rev3.3067
[2] http://dx.doi.org/10.1080/1743727X.2016.1166486

1 Why Meta-analysis?

Key questions
What is meta-analysis?
Why do we need it?
What can we learn from it?
What are its limitations?

Nothing Like Leather
A town fear'd a seige and held consultation
Which was the best method of fortification;
A grave skilful mason said, in his opinion,
That nothing but stone could secure the dominion;
A carpenter said, though that was well spoke,
It was better by far to defend it with oak;
A currier, wiser than both these together,
Said, 'try what you please – there's nothing like leather'.

Daniel Fenning (1771). *The Improved Universal Spelling-Book* (16th Edition). London: Crowder, Baldwin & Collins, p. 36.

Introduction

This chapter presents an overview of the importance of meta-analysis in relation to our knowledge about effective teaching and learning in education. Meta-analyses of interventions about the use of phonics in reading, an area of enduring interest to policy and practice, are used to show where findings from meta-analysis of intervention research can provide a clearer picture of the complexity of findings in this particular field. These are presented through a selection of research findings from educational meta-analyses.

One of the problems that practitioners face is being presented with a vast array of research studies in education which all seem to claim they have 'significant' findings. Sometimes these are contradictory findings, such as about the effects of different ways of deploying teaching assistants; sometimes the findings are just about very different aspects of education, such as benefits of homework or of parental involvement

or the effects of peer-tutoring and the impact of breakfast clubs. How are busy teachers and head teachers to make sense of the conflicting information and advice? Which should they take notice of and which should be ignored? It sometimes feels like researchers are like the leather-maker in the fable at the beginning of the chapter, with each one advocating for their own particular area of interest as a solution to educational problems. This was what drew me to meta-analysis in the first place. I saw it as a way of getting an overview, of hovering above a research landscape and making sense of studies in a particular field to draw an overall conclusion about the average or typical benefits of a particular approach. It also offers a way to find out which features of different approaches are linked with the more and less successful research studies.

What Is Meta-analysis?

Meta-analysis is a statistical procedure used to combine the data or the findings from multiple studies. When the effect or the impact on learning outcomes from experimental research studies is consistent from one study to the next, meta-analysis can be used to identify this common effect by using appropriate statistical procedures to calculate a fair or unbiased average overall impact. When the effect or impact varies from one study to the next, meta-analysis can also be used to identify features of the various studies that explain the variation.

Suppose you have searched carefully and systematically and identified eight studies of the impact of peer feedback on the quality of students' writing (see Table 1.1: these are taken from a meta-analysis by Graham and colleagues published in 2015). The studies have used different measures of writing quality; they also vary in terms of design (quasi-experimental, or randomised experiments) and scale (sample size). How can we make sense of the findings overall? First, it is possible to estimate the impact using a common measure, called effect size (see the section later in this chapter for a further discussion of the strengths and weaknesses of this approach). Then we can weight each study so that it contributes fairly to an overall average (smaller studies will count less and larger studies more: there are different ways to do this, and more details are provided in Chapter 2, but the aim is to produce a fair average). This lets us summarise or synthesise the findings as shown in Table 1.1. Overall the impact across all the studies is an effect size of 0.58, which means that approximately 72% of the pupils in the peer feedback classes were above the mean of the comparison classes at the end of the different experiments (though this varied from study to study). This small

Table 1.1: *Graham's meta-analysis of peer feedback on writing quality*

Study	Design	Sample size	Effect size
Benson (1979)	QED	288	0.36
Boscolo and Ascorti (2004)	QED	122	0.97
Holliway (2004)	RCT	55	0.58
MacArthur et al. (1991)	QED	29	1.33
Olson (1990)	QED	42	0.71
Philippakos (2012)	RCT	97	0.31
Prater and Bermudez (1993)	RCT	46	0.15
Wise (1992)	QED	88	0.62
Overall		**767**	**0.58**

(given in Graham et al. 2015)

meta-analysis suggests that overall the impact of peer feedback approach has been beneficial in these research studies.

The important feature of meta-analysis is that it focuses on two questions at the same time. These are 'does it *work* (on average)?' and '*how well* does it work (on average)?' It converts findings from different studies to a common scale so that they can be combined or 'synthesised'. The value of meta-analysis becomes even more apparent the more studies you have to combine.

Statistical significance focuses on just one dimension of one of these questions in terms of whether the effect reaches a threshold relative to the scale of the experiment. Does it work in terms of the chance variability you might expect, given the sample size? I have always found the focus on statistical significance confusing, partly because of the logic of null hypothesis testing and partly because there are other things to consider in judging how convincing the findings from a single study are, which are often as, if not more, important. The use of effect sizes and meta-analysis offers a way through this confusion by identifying patterns of findings and effects across studies, providing a bigger picture than one offered by a single research study, however impressive the design and analysis.

A Common Scale to Compare Findings: Effect Size

Effect size is such an important metric that we need to examine the idea more closely. Effect size is a key measure in research generally and meta-analysis in particular. It is a way of measuring the *extent* of the difference between two groups. It is easy to calculate, easy to grasp

and can be applied to any measured outcome for groups in education or in 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 systematic 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 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.

The basic idea is to compare groups, relative to the distribution of scores. This is the standardised mean difference between two groups. There has been some debate over the years about exactly how to calculate the effect size (see Chapter 2), but in practice most of the differences in approaches are small in the majority of contexts where effect sizes are calculated using data on pupils' learning (Xiao et al., 2016). It is important to remember that, as with many other statistics, the effect size is based on the *average* difference between two groups. It does not mean that all of the pupils will show the same improvement.

An Historical Aside: Statistical Significance

We have reached a crossroads in the history of the use of statistics in research studies. Significance testing has been used in experimental studies for nearly 100 years, but there is a growing consensus that its misuse is problematic. In 1925, Ronald Fisher (1890–1962) advocated for the idea of testing a hypothesis statistically, and named the technique 'tests of significance' in his monograph *Statistical Methods for Research Workers*. He suggested a probability of one in twenty (0.05) as a convenient cut-off level. In Fisher's approach, set out in more detail in 1935 in his book *The Design of Experiments*, you start by hypothesising the opposite of what you believe to be true, because of the problem of moving from deductive to inductive inference. This is called the *null hypothesis*: that there is no difference between the groups. You then conduct your experiment and consider how likely it is that the data that you have painstakingly collected could have occurred on the assumption that there would be no difference (which you probably didn't believe, or you wouldn't have bothered to run an experiment to find out; the double negatives add to the challenge of

understanding this approach). This always seemed to me a slightly dishonest philosophical sleight of hand to avoid the problem of induction. You then set an arbitrary level of probability of between one in ten and one in a hundred, or even one in a thousand, but usually one in twenty (0.05 or 95%, as Fisher originally recommended), that you think is reasonable to draw conclusions. Depending on this level of probability, you conclude you have disproved the null. I was never convinced that one in twenty was that unlikely. The probability of getting two pairs in five-card poker or throwing a total of six with three dice in a single throw is about one in twenty.

You then make the abductive leap that the *opposite* of the null hypothesis must therefore be true and you accept the 'alternative hypothesis', which is that there really is a difference between the groups. However, it is clear that neither the null hypothesis nor its alternative has ever been proved. You have just decided that the data you have suggests that it is unlikely that there is no difference and then flipped this to conclude there really is a difference, on the basis that it (probably) won't work out like that more than one in twenty times. This decision is based on an arbitrary threshold about the likelihood of the reverse of what you actually believe. When this was first explained to me (well, actually not the first time, or even the second; it was perhaps the first time I understood what was being explained), I thought I was entering Terry Pratchett's Discworld and saw the null and its alternative balanced on the back of an infinite regress of hypothesis-bearing turtles. The use of tests of significance and null hypothesis testing certainly moved scientific inference forward in the last century. However, in education we need greater precision than the way the approach has evolved now allows. It now constrains more than it enables. This scepticism about formal hypothesis testing forms part of my advocacy for meta-analysis. I am more interested in the extent of the difference between the groups and the regularity of the extent of this difference across studies than I am about the particular level of probability of the opposite of what I really think.

Types of Effect Sizes

Effect sizes can be thought of in two broad categories. First are those that compare the extent of the average differences between two groups, relative to the spread of the scores. These are often referred to as standardised mean differences.[1] Second are those which report the relationship between two variables or the extent to which their overlap is explained,

[1] Such as Cohen's d, Glass's Δ or Hedges' g.

sometimes called variance-accounted for effect sizes.[2] Different effect sizes can be converted mathematically into others (a standardised mean difference (d) to a variance-accounted for effect size (r), for example). This can be done either directly, where they have a mathematical equivalence, or by estimating, based on assumptions about the distributions. In this book, the focus is on standardised mean differences, as these are most commonly used in intervention research in education. However, it is important to bear in mind that the research design from which the data is analysed, the sample or population studied, the measures used and the precise calculation method used all affect the comparability of particular effect size measures (see Appendix B for more information).

What Does an Effect Size Mean?

As an example, to help understand what an effect size means, suppose we have two classes of 25 pupils: one class is taught using an effective feedback intervention, and the other is taught as normal. The classes are performing at the same level and so are equivalent before the intervention. For this illustration, let's say that the intervention is effective with an effect size of 0.8 (which is about the average for feedback interventions). This is typically considered a 'large' effect size (Cohen, 1988). This means that the average pupil in the class receiving the feedback intervention (i.e., the one who would have been ranked 12th or 13th in their class) would now score about the same as the person ranked sixth in a control class which had not received the intervention. The rankings of the classes have moved relative to each other. This may result from a change in the outcomes for as few as 7 pupils in a class of 25.

You can also picture it this way. There are two classes of pupils walking across a field (see Figure 1.1). The pupils are walking as a mixed group, but each class are wearing matching hats, different from the other. One group are wearing striped hats, the other hats with spots. From above, the crowd is roughly circular in shape, and everyone is moving across the field in the same overall direction. A few are walking a bit faster than others and are at the front. A few stragglers are dawdling a bit behind, but the overall shape is roughly symmetrical. You can't tell if one group is further ahead than the other, but if you took a picture from above and measured the position of each child, the average would be a line across the middle of the crowd for both classes.

[2] Such as R-squared (R^2), eta-squared (η^2) or omega-squared (ω^2).

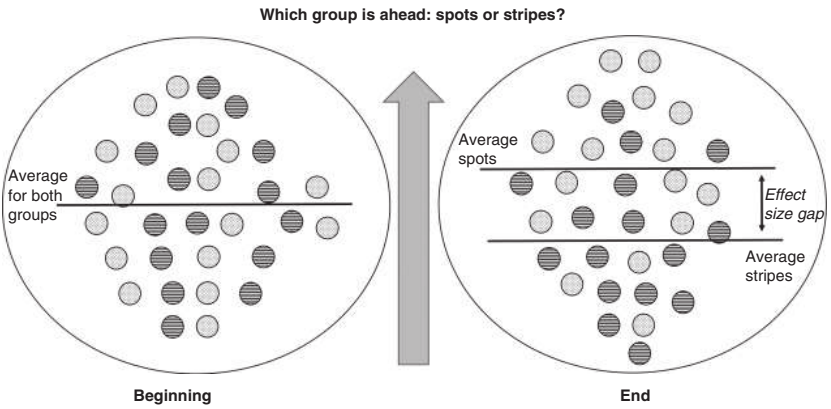

Figure 1.1: Visualising differences as effect sizes

The group with spotted hats are then given an instruction to speed up (at this point you have to imagine that the pupils in the striped-hats group don't hear this). The spotted-hats class mostly start walking a bit quicker, but this varies, as not all children react to the instruction in the same way. After ten minutes, you take another picture from above and work out the centre of each group (the average position). The spotted group have moved ahead, on average, and you can measure the distance between the two average lines (see Figure 1.1).

The effect size is a way of summarising the gap between the two classes and adjusting for how spread out each class is (a standardised effect size). Visualising the different positions of these two classes provides a helpful way of thinking about effect size. It is a measure of how far apart the two groups are as a result of the intervention (see Figure 1.2).

We can also think about what this means in a range of other ways. These can be confusing, but it is important to remember that they are all trying to express the same idea, the extent of the difference relative to the spread. Figure 1.2 displays this for an effect size, or a standardised mean difference, of 0.8.[3] For this extent of difference, 79%of the intervention group will be above the mean of the control group. This calculation is called Cohen's U3 or the percentage of pupils' scores in the lower-mean group that are exceeded by the average score in the higher-mean group. However, it is also important to note that 69% of the two groups will still overlap. There is a 71% chance that a person picked at random from the

[3] I am grateful to Kristoffer Magnusson for permission to use this image from his helpful website: http://rpsychologist.com/d3/cohend/.

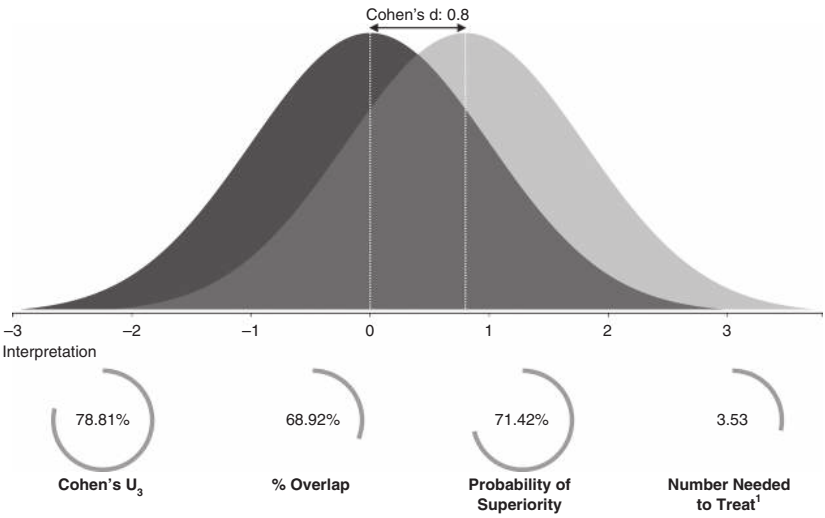

Figure 1.2: An effect size of 0.8

treatment group will have a higher score than a person picked at random from the control group (rather than 50–50 if they were the same). This is sometimes called *probability of superiority*. In terms of likely benefit, to have one more favourable outcome in the intervention group compared to the control group we would need to treat at least four pupils (technically 3.53, but I struggle with half measures here[4]). This means that if 100 pupils go through a similarly effective intervention, 28 more pupils are likely to experience a favourable outcome, compared with if they had been in the control group. This effect is typically called large, following Cohen (1988). But as Cohen himself noted, it is important to be cautious about labels like this with an arbitrary cut-off. It is worth reproducing Cohen's point in full here:

> The terms 'small,' 'medium,' and 'large' are relative, not only to each other, but to the area of behavioral science or even more particularly to the specific content and research method being employed in any given investigation. In the face of this relativity, there is a certain risk inherent in offering conventional operational definitions for these terms for use in power analysis in as diverse a field of inquiry as behavioural science. This risk is nevertheless accepted in the belief that more is to be gained than lost by supplying a common conventional frame of reference

[4] This is the 'number needed to treat', or NNT; for those interested, see Furukawa and Leucht's (2011) paper on converting a standardised mean difference effect size (d) into the NNT.