How To Assist Research-Informed Education?

Stephen Gorard, Wenqing Chen

Durham University Evidence Centre for Education

DOI: https://dx.doi.org/10.47772/IJRISS.2025.910000338

Received: 12 October 2025; Accepted: 20 October 2025; Published: 12 November 2025

ABSTRACT

This paper looks at some of the key decisions to be faced in promoting the use of research evidence in policy and practice – the quality of the evidence to be used, and how that quality can be judged. Once the policy/school context has determined the issues to be faced, and before the professional judgement of policy-makers/teachers is used to help carry this out, evidence should presumably inform the decision on how to address those issues. However, there is a large amount of apparent research evidence on most topics, and much of that research is very weak. Who decides which evidence is appropriate to use, and how is that judgement made? This paper illustrates the issues with reference to examples drawn from what works summaries, using a generic approach to judging quality, and shows that deciding on the relative quality of research evidence is paramount in whether evidence use will be effective or not. The paper continues by considering the difficulties for users and others in judging the quality of research, but argues that this problem must be solved. The only real alternative would be for practitioners to have purported evidence-led approaches thrust upon them, while policy-makers would presumably continue to evade the imperative to use evidence appropriately themselves.

INTRODUCTION

It sounds sensible to say that education should be informed by research evidence. This is not to say that evidence should dominate. But that once the policy/school context has determined the issues to be faced, and before the professional judgement of policy-makers/practitioners is used to help carry out any plan, good evidence should presumably inform the decision. However, plausible as this sounds, there is as yet little convincing evidence that using research evidence does improve real-life educational outcomes. This is partly just because the idea has not been tested much. But it is also partly because it is not clear which bodies of evidence are trustworthy enough to be used, or even how such a judgement of trustworthiness could be made.

There is a lot of "research" on education, but much of it is very weak – lacking a coherent design, small-scale, with lots of unacknowledged missing data, and/or using inappropriate measures. This makes it difficult to identify the best and most robust evidence or bodies of evidence to use, taking time and skill that most research users currently do not possess. But using any evidence, without consideration of its quality, would mean that educational outcomes might be unchanged or even harmed by doing something inappropriate. This would represent a damaging opportunity cost for each generation of students and their one-chance of formal education. It would then also be difficult to demonstrate the assumed benefits of using evidence for education, encouraging those stakeholders who are resisting any use of evidence, rather than working to make evidence use more appropriate, feasible and fruitful.

This paper starts by summarising the current situation for knowledge about evidence use in education. There is little strong evidence, and this so far shows no impact on student outcomes from evidence use by teachers. In at least some studies this could be because the primary evidence used by teachers may not actually be robust enough for real-life use. The paper continues with a discussion of how the robustness of any evidence in education can be assessed, and then illustrates, with the example of evidence on using enhanced feedback in the classroom, how important it is that the robustness of primary evidence is assessed before it can be safely used. This section illustrates the dangers of synthesising bodies of work without consideration of quality, as had been done by authors like Hattie (1992). The paper then shows via secondary data and an example how hard it is for users like teachers to be able to judge the quality of existing evidence. It ask what kind of conduit can be trusted to collate, judge and synthesise evidence for users. The final substantive section suggests a

somewhat unwelcome alternative, which is that teachers could simply be forced to use and evidence-led approach. The conclusion considers some more nuanced ways forward.

The push for evidence use

In England, as in many other areas, there is considerable pressure for educators to be research-informed, and for their practice to be evidence-led. However, even 20 or more years since the inception of a "what works" revolution, the actual evidence is sparse on how to implement evidence use successfully (Flynn 2019, Nuttley et al. 2019). A recent very large-scale review of how to get evidence into use found that there was no clear answer on how evidence transfer from research to use could be best achieved (Gorard et al. 2020a). It has not even been convincingly demonstrated that using research evidence in education does actually improve educational outcomes.

Much of the research writing on this topic is conceptual or advocacy work. Finnigan (2023), for example, suggests strategies and models for research-practice partnerships without testing them. Cherney et al. (2013) also propose suitable mechanisms but do not test them. Chapman and Ainscow (2019) present a conceptual framework for evidence use, and Coburn and Talbert (2006) discuss institutional theory.

Only a small proportion of research writing about evidence use is empirical - only a subset of which has tried to evaluate the impact of evidence-use by education practitioners. And most of that small proportion is very weak in terms of design or scale.

For example, Washburn et al. (2023) report success from a summer literacy intervention, based on prior evidence. However, there was no comparator. The improvement was only from pre- to post-intervention testing, when these Grade 1 to 3 students were also older. Greenwood et al. (2003) claimed success from school-wide use of an evidence-based literacy approach. But the study was in only one school, and compared the evidence-based approach with outcomes for students from a different grade level anyway. Dieker et al. (2009) reported developing Web-based video models of effective instructional practices, as defined by research evidence, to help train teachers. They discuss the "influence" of such modelling but never tested it, even with the 11 participants that they mention.

Tortorelli and Bruner (2022) based their work on 17 US elementary teachers trying to use an evidence-based approach to improve student formative spelling assessments over one year. There was no comparison group. Taylor et al. (2005) found that US schools (out of 13) with more success in implementing an evidence-based reform had slightly better improvement in reading. But the study is only correlational, and the result could well be due to pre-existing differences between schools and teachers.

The list of such small, weakly designed projects is long in this field, as in all fields of education research. Of course, there is nothing remiss with such small and possible indicative projects in themselves as long as they are not over-interpreted by their authors or others. The danger lies in them confusing the field, and suggesting that the answers to how evidence can best be used are known, when they are clearly not.

There is a small amount of research on the impact of evidence use by teachers that is somewhat better than the examples above. Such research would usually have a counterfactual of a reasonable size, and focus on a restricted pre-defined range of outcomes. The overall pattern of this work is incomplete, but it is a reasonable generalisation to say that this strongest work has been the least likely to find any impact of evidence use by practitioners.

For example, Rose et al. (2017) used two evidence champions in each of 60 schools, trained over four workshops by academic researchers, to promote evidence use in their schools. Compared to a control group of 59 schools, there was no improvement in student Key Stage 2 reading results after two years, and not even an improvement in teacher attitudes to research use. This may mean that the approach of using champions is not effective. Or it may mean that the purportedly evidence-led intervention presented to teachers was ineffective. In this kind of meta-work it is difficult to distinguish between these two situations, and this emphasises the importance of only using the best available evidence. Here the intervention used was growth mindset. If the

complete evidence base for growth mindset is not as strong as Rose et al. imagined then this otherwise innovative study could not in fact test the impact of research champions (Li and Bates 2020).

Wiggins et al. (2019) used two cohorts of students in 40 schools. The research leaders from each intervention school were trained and assisted by academic researchers to promote evidence use, via eight CPD sessions, termly follow-up meetings over two years, a newsletter, website with resources, peer network, and school visits by the academic team. Each school had to decide on their priorities for improvement, consult the EEF Toolkit of evidence (see below), implement the most promising approach, monitor progress and change the approach if it appeared to be not working. The results were minimal. After two years the intervention students were ahead in maths (effect size 0.04) and English (0.03) by a tiny amount. Here the research evidence to be used was not specified but its source was. Perhaps the way that the Toolkit is created does not actually yield the most promising approaches as clearly as it could. And therefore again we cannot tell whether evidence use does not work, or whether the evidence used was not strong enough.

Lord et al. (2017) looked at the impact of providing schools with research summaries and other evidence-based resources. The 12,500 primary schools were randomised to four treatment groups and a clean control. Treatment schools were sent the resources, with or without light touch support, and their subsequent KS2 results monitored. This passive transfer of research evidence made no difference to student outcomes in English. Schools probably receive too much mail for this approach to work, and schools may anyway need more support to use the material. Given that the evidence summaries provided were of four different types from four different expert groups it is less likely that the lack of impact was due largely to a choice of inappropriate underlying evidence. But this explanation is still possible.

Griggs et al. (2016) evaluated an intervention based on a designated research champion covering five schools (2,075 pupils), auditing schools' needs, and providing a variety of generic and targeted development activities. This did not even increase teachers' reported use of research, let alone improve student attainment. Speight et al. (2016) found no evidence of impact on teacher behaviour from encouraging teachers in 10 self-selected primary schools (teaching 280 pupils) to use modified or summarised research evidence on issues such as metacognition.

Helmberger (2014) reported that the frequency of teachers' use of research-based strategies is not linked to student test success. The nature of the students and their SES was more important in determining outcomes. Similarly, Edmonds (2007) looked at the results of teachers using research-based comprehension approaches to teaching social studies. How well the approach was implemented was not related to student outcomes. The key predictor was students' reading test scores on entry to the fourth grade. Jacob (2017) found no difference in the attainment of US students in grades 2 to 5 when teachers in seven schools were randomised to Evidence-Based Literacy Instruction or a control. One of the problems faced was the difficulty for the teachers of implementing the approach. Similar issues arose in Andreassen and Braten (2011). This raises the issue of whether an intervention or policy can ever be considered appropriate if it is not feasible for teachers to implement.

Related to this, See et al. (2016) compared the Key Stage 4 outcomes of nine schools which had all used a policy of research evidence to inform their own practice, with the progress in 49 other secondary schools in the same area. The research school headteachers used Hattie's evidence on enhanced feedback (see below). There was no impact. It may be feasible for practitioners to use research evidence to inform their own practice. However, to do it well would require clearer guidance, professional development and modelling of any strategies suggested. The headteachers all struggled with the primary evidence despite the presence and assistance of academic researchers. Anyway, like growth mindset perhaps the evidence for feedback is not as strong as has usually been portrayed.

Assessing the quality of individual studies

Many different schemes have been proposed for assessing the relative trustworthiness of research findings – of the kind that could be used in education practice. Some are merely checklists of what is in research reports, and are therefore only a precursor to judging quality (Logullo et al. 2020). Some schemes are intended for users of research, like That's a Claim (https://thatsaclaim.org/). Some like the padlock ratings used by the Educational

Endowment Foundation (EEF) in England help users to judge individual studies, before their results are synthesised (https://educationendowmentfoundation.org.uk/help/projects/the-eef-security-rating/). Others are intended more for researchers judging the security of a larger body of work consisting of many individual studies (Gough 2007), such as systematic reviews of evidence (Madaleno and Waights n.d.). All such procedures have much in common, with an emphasis on full reporting of large-scale, well-designed studies, and would presumably lead to similar conclusions.

An especially simple approach for assessing the internal validity of individual studies appears in Gorard (2021). It is based on the design of each study and how well that fits the research question, the scale or size of the study, and the completeness and accuracy of the data. There are other issues that could be considered such as conflicts of interest, or appropriateness of analyses. However, once the first four are decided then the other factors either generally fall into line as well, or at least cannot make our judgement worse.

Whatever approach is used to judge quality relies on full. comprehensible reporting of that research. For example, it is not possible to judge whether the research design fits the research question. unless the question and design are both clearly stated in the report. Judging this is summarised in the first column of Table 1. A suggested procedure for using the table would be to start with the first column, reading down the design descriptions until the research you are reading is at least as good as the descriptor in that row. In this the row, move to the next column and read down the descriptions, if needed, until the study is at least as good as the descriptor in that row. Repeat this process for each column. For any column, if it is not possible to discern the quality of the study from the available report(s) then the rating must be placed in the lowest category. Each study sinks to its lowest level on these four key factors. The final column in the table gives the estimated security rating for that study. A much fuller description appears in Gorard (2021).

Table 1 - A 'sieve' to assist in the estimation of trustworthiness of any research study

Design	Scale	Missing data	Measurement quality	Rating
Strong design for research question	Large number of cases per comparison group	Minimal missing data, no impact on findings	Standardised, independent, accurate	4□
Good design for research question	Medium number of cases per comparison group	Some missing data, possible impact on findings	Standardised, independent, some errors	3□
Weak design for research question	Small number of cases per comparison group	Moderate missing data, likely impact on findings	Not standardised/independent, errors	2□
Very weak design for research question	Very small number of cases per group	High level of missing data, clear impact on findings	Weak measures, high level of error, or many outcomes	1□
No consideration of design	A trivial scale of study	Huge amount of missing data, or not reported	Very weak measures	0□

The overall rating suggests a research finding whose trustworthiness is at least at the level of the descriptions in that row. So 4□ suggests a study that is as secure as could reasonably be expected (because no research will be perfect), and 0□ represents a study that is so insecure that it adds nothing safe to our knowledge (the quality of most actual research in education). Of course, there are no objective criteria for deciding on any rating, or even on how many categories of ratings there should be. This is true of any such approach to judging quality. However, this table was the basis for the EEF scheme, has been used by a large number of reviewers (including teachers) who have found it useful in sifting thousands of studies by quality, and found considerable agreement between themselves when allocating security ratings independently.

ISSN No. 2454-6186 | DOI: 10.47772/IJRISS | Volume IX Issue X October 2025

Why the quality of research matters

Using this sieve approach for illustration, the next section shows why judging the quality of individual studies matters, and how not doing so can lead to misleading and perhaps harmful use of evidence, especially when bodies of evidence are summarised.

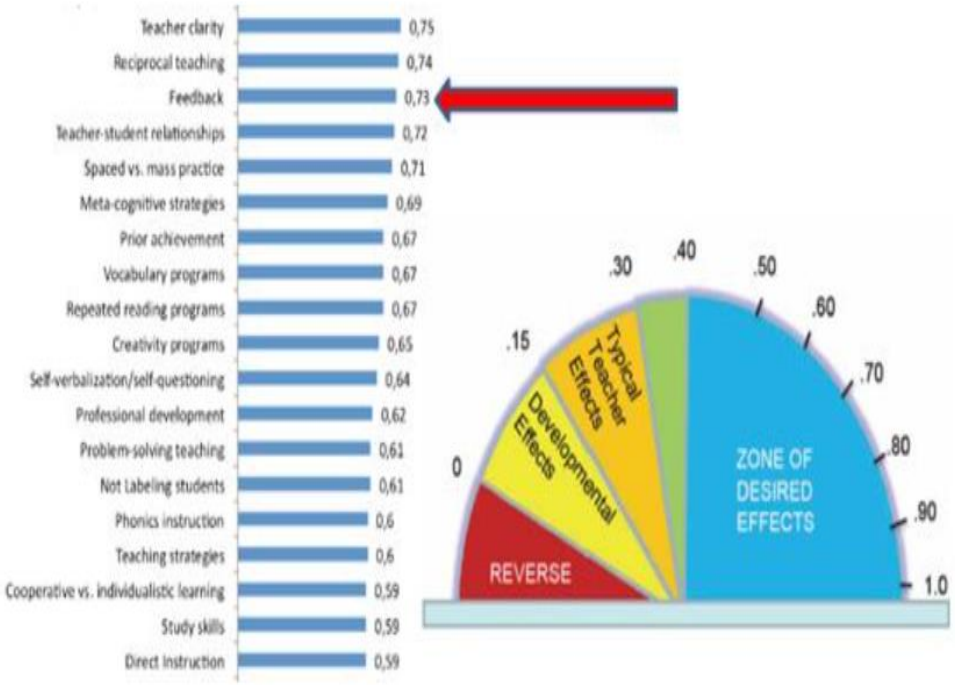
Enhanced feedback as an example

A widespread and pioneering example of engineering bodies of evidence into greater use for education comes from Hattie (2008), and subsequent publications. This was perhaps the first real attempt to summarise a huge body of evidence on key topics for education and to present it in a way that teachers and other users could understand. In outline, it involved a meta-analysis of all available prior meta-analyses on each topic, estimated an overall "effect" size for each, and portrayed these in a standard and simple diagrammatic form. The impact of this work was considerable.

Other researchers and organisations have now produced similar summaries including the What Works Centres, and the Educational Endowment Foundation Toolkit, in England. The Toolkit aggregates available meta-analyses on each topic, and portrays the resulting "effect" size as an estimated average number of months progress for students. It also shows how much any intervention or programme might cost, and gives an indication of the strength of the evidence so far. The latter is a crucial addition to any summary of evidence.

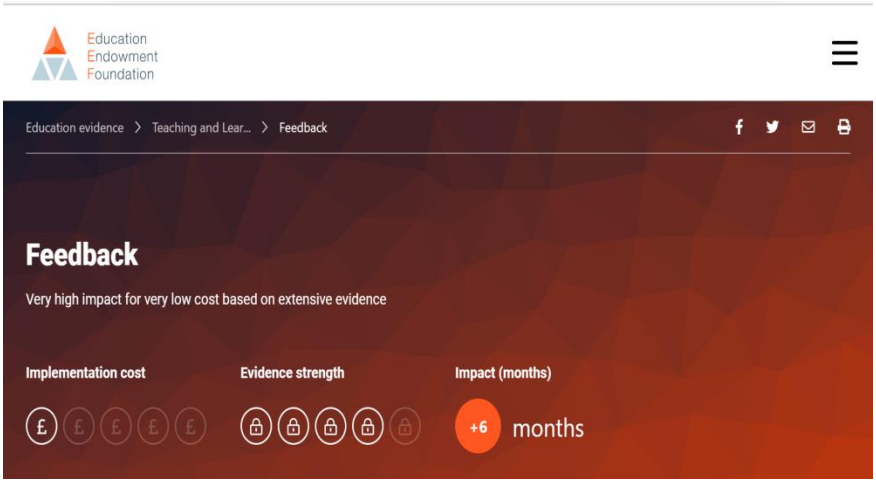
Hattie (2008) produced an aggregated "effect" size for using enhanced feedback in the classroom of +0.73 (Figure 1). Improving the use of immediate, interactive, formative responses in classrooms would require some extra development for teachers, and perhaps a period of adjustment for students. It can otherwise be achieved for no additional costs to schools. Feedback is therefore presented as a feasible and powerful approach to improved teaching.

Figure 1 – "Effect" sizes from Hattie (2008)



A similar result appears independently in the EEF Toolkit (Figure 2). Here the result is presented as an average additional six months progress for students, where teachers use enhanced feedback. This is "high impact" for "very low cost". In addition, the Toolkit provides an estimate of how strong the body of evidence that lies behind this finding is. Here, it is four padlocks out of a possible five (where five is a rare and ideal possibility).

Figure 2 – "Effect" size for feedback from EEF Toolkit



We have chosen to consider feedback in more detail because it has been portrayed as being usefully effective, relatively easy to implement, and backed up by a sound evidence base. It is therefore a strong example of an approach that might sensibly be adopted by schools and teachers enthusiastic to use an evidence-based approach. What follows is an illustration of the difficulties of such an adoption by users. It is not really about the underlying substantive value of enhanced feedback in the classroom, or its merits relative to any other approach.

As noted above, both estimates of the value of feedback are based on meta-analyses of existing meta-analyses of prior studies. This means that the "effect" sizes used in the underlying meta-analyses are aggregated in the over-arching meta-analyses, weighted by the number of cases in each study. Otherwise, no account is taken of the research quality of the underlying meta-analyses, and no account at all is taken of the research quality of the individual studies in those meta-analyses, or their appropriateness to be combined in this way.

This means that the "effect" size of a poor quality study is treated as equivalent to the "effect" size of a very good study. For example, Hattie (2008) synthesised the results of 74 meta-analyses of feedback, that totalled 4,157 individual studies (reporting 5,755 different effect sizes). This sounds very powerful. However, some of these studies were randomised control trial designs, appropriate for making clausal claims. Others, indeed most studies, were not. Some were based on convenience comparisons – teachers that simply were or were not observed to be using enhanced feedback. It is clear that good teachers (however judged) are more likely to use feedback. But what is not clear is whether other teachers can improve their teaching by being made to use feedback. Convenience comparisons cannot address this, and should not have been included in the synthesis. Or if included they should somehow be weighted to count for less in the outcome than stronger studies do.

One study, for example, had used a MANOVA analysis based on those students who agreed to participate compared to those who refused, presented as evidence of the impact of the intervention. This is incorrect in many ways. Comparing the results of volunteers with those of refusers is not appropriate. The groups will differ in many ways, including motivation, and so any subsequent difference in outcomes cannot be attributed to feedback. In addition of course, MANOVA like all tests of significance has as a necessary underlying assumption that all cases have been randomised (i.e. not volunteers).

This latter error is especially important for meta-analyses because many of the underlying studies did not report "effect" sizes, nor the fundamental findings needed to compute "effect" sizes (such as the mean and standard deviation of each group). In this large number of reports Hattie (2008) and the EEF Toolkit rely on the p-values created by significance tests. They did this whether the study met the requirements for the test or not (the vast majority will not, if only through dropout). Added to this is the issue that p-value thresholds (like $p<0.05$) are not "effect" sizes, and that any attempt to convert them to one will create an inaccurate estimate. P-values themselves cannot tell the reader whether the impact was positive or negative, and if the means are provided so that the direction of impact is known, then the p-values are not needed.

Neither the underlying studies, nor the meta-analyses at either level, take missing data into account. This is crucial because missing data cannot be assumed to arise by chance and always creates the potential for bias (Gorard 2020). Again, in both summaries, the studies (and meta-analyses) with high levels of missing data are incorrectly treated as equivalent in the aggregation to studies with little or no missing data.

The "effect" sizes aggregated are not just based on different study designs, and different and often missing ways of computing them. They are based on different samples and populations. Some studies were conducted with very young students, others with university students or adults. Some involved only students with special needs or behavioural disorders. Further, the measures used as outcomes differed widely. Some studies used standardised tests of academic outcomes (across a wide range of subjects), some bespoke tests, some teacher assessed, some laboratory-based, and others were observational. Many were not about academic but emotional or behavioural outcomes. All were simply aggregated regardless, to get that overall +0.73 result. This does not make any kind of sense.

One of the underlying meta-analyses used by Hattie (2008) was by Kluger and DeNisi (1996). This was described as "the most systematic" and "included studies that had at least a control group, measured performance, and included at least 10 participants". Measuring performance of at least 10 cases, with some kind of counterfactual sounds like a minimal requirement. But the horrifying implication is that other meta-analyses in the 74 included by Hattie (2008) had studies with fewer than 10 cases, no comparator, or did not measure performance.

Additionally, as in any large-scale endeavour some mistakes can creep in. For example, not all of the studies or "effect" sizes cited in Hattie (2008, Table 1, p,83) can now be found. Hattie (2008) states that the overall effect size of 54 studies in Lysakowski and Walberg (1982) was +1.13 whereas the original paper reports it as being +0.97. Hattie (1992) reported that "Skiba, Casey and Center (1986) used 315 effect-sizes (35 studies) to investigate the effects of some form of reinforcement or feedback and found an effect-size of 1.88", but it is later reported as having 35 effect sizes (not studies), and an effect size of +1.24. These differences may be small, but this kind of thing is a concern in a summary of evidence that is intended for widespread use.

The main point made here is that these hyper-analyses, and many others, are conducted without consideration of the quality or trustworthiness of the underlying evidence.

An even bigger consideration is whether meta-analyses such as these are even possible. Because an effect size is a "standard" score, some commentators have suggested that this makes effect sizes comparable between studies using different measures and approaches. This is not true, and assuming it to be true can be misleading (Morris 2019). The value of a standardised difference between scores relies on a number of factors such as the research design, scale of the study, missing data, and the nature and quality of the underlying measurements (Gorard 2021). This means that effect sizes cannot reasonably be aggregated or averaged, except where all of the studies involved were of the same design and quality. Effect sizes are also open to publication bias (Chowl and Ekholm 2018). Effect sizes give no indication of the quality of the study from which they emerged, although there is evidence that smaller and weaker studies tend to yield larger-seeming effect sizes (Wolf et al. 2020). As ever, research quality matters.

Evidence on studies of feedback

Gorard et al. (2017) report a structured review of single studies related to the use of feedback in primary age classrooms. The number of studies (19) is less than in Hattie (2008), for example, because the search was limited to the year 2000 and beyond, and only studies of primary age, with a clear comparator, and academic outcomes, were included. As shown in Table 2, the majority of the remaining studies show positive impacts from the use of enhanced feedback. Around a third showed no benefit or worse. Treating all of these studies as equivalent, apart perhaps from their scale, would lead an analyst to conclude that, on balance, feedback was an effective and promising approach to improving teaching. And aggregating their results in a meta-analysis, even taking into account their sample size, would lead to a substantial positive "effect" size.

Table 2 - Impact of enhanced feedback in primary schools, as assessed by 19 studies found in a systematic review

Number of studies reporting positive effect	Number of studies reporting no effect or negative effect
13	6

Source: Gorard et al. (2017)

However, if the quality of each study is taken into account, using the sieve above or a similar process, then the picture changes considerably (Table 3). There are still 13 positive outcomes and six negative or neutral, of course. Studies rated as zero for quality (adding nothing to our knowledge of the casual impact of feedback) were excluded. Most studies (14 out of 19) are of the lowest quality. These might be convenience comparisons, very small, have high levels of missing data or be based on measures tied to the intervention. Of these, 11 reported positive outcomes. This is a common finding of bias in reviews – weak studies tend to have larger and more positive "effect" sizes. Of the remaining five studies, three have negative results and only two report a positive outcome. There are no studies with a causal design, large-scale, low dropout, and standardised outcomes (4□) on this topic. Of the two best quality studies, one is positive and one negative. These are summarised in Table 4.

Table 3 - Quality and impact : studies of feedback

	Number of studies reporting positive effect	Number of studies reporting no effect or negative effect
4□ Highest quality	-	-
3□ High quality	1	1
2□ Medium quality	1	2
1□ Low quality	11	3

Table 4 – Highest quality studies of feedback

Reference	Intervention	Smallest cell	Attrition	Effect size
Lang et al. 2014	Formative Assessment	15 schools, 2,000+ pupils	1 school, unknown number of pupils	0.20
Phelan et al. 2011	Feedback (Year 7)	2,045 pupils	Not reported	0.03

Using the quality of individual studies, in a way that a meta-analysis does not permit, changes the conclusion that a reasonable reader might otherwise draw from any summary of evidence. Based on quality judgements, it is not clear that feedback is such a promising and well-evidenced approach to improving teaching. This does not apply only to enhanced feedback but to any claim based on summarising prior evidence. Others may rate these studies somewhat differently. There has to be an element of judgement. But the key point here is not a substantive one about these studies, or of the benefits or not of feedback. The main point is to illustrate how taking study quality into account can transform an apparently evidence-based picture. And as noted above, if we want to know whether using evidence can improve policy or practice then we must be sure that the evidence used is solid.

Who will judge the quality of evidence?

It is clear that any synthesis of the primary evidence on any topic must take the quality of each study into account. Not to do so is misleading and dangerous. However, it is less clear who can or should make those judgements, and how they would do so.

Teachers report lack of support for their use of evidence (Lysenko et al. 2016), often find research inaccessible (McCartney et al. 2018, Booher et al. 2020), are not equipped by their training to handle research (Mai and Brundrett 2022), and are even sometimes individually hostile to evidence use. Teachers may imagine that using evidence removes their autonomy, and jeopardies their vision of what education should be (Holloway and Hedegaard 2023). In fact, many teachers are sceptical of the value of evidence-led improvement, or to believe that good teachers are born that way, and this makes them unlikely to use evidence-led approaches (Nägel et al. 2023). This occurs especially where the teachers have lower academic attainment themselves. In this context, it is noteworthy that teaching is a profession entered disproportionately by graduates with lower academic qualifications than many other professions (Gorard et al. 2021).

Teacher confidence in using evidence

We gained access, via the UK Data Service, to the data from an NFER 2010 Survey of Teachers with a focus on the 3,936 teachers in England with experience in Key Stages 1 to 4. Only 58% of responding teachers said that using other people's research was useful, while 24% had not tried it. Of six suggested approaches to improving their teaching, using research evidence was regarded as the least useful. Instead, reflecting on their own practice was the most popular (97%). The usefulness of reviewing practice as part of school self-evaluation was next (73%), followed by using teacher assessment data (71%), and working towards performance management targets (64%). Even conducting their own research was regarded as slightly more useful than using external research evidence (61%).

Around 60% of teachers claimed they knew where to find research evidence relative to their teaching. The biggest barrier to use of external evidence reported by these teachers was lack of time (44%), but only 24% said that their school encouraged the use of research evidence, and only 16% said that they had regular opportunities to discuss research with colleagues. However, it is then curious that 34% said they had conducted their own research in the last year, and 52% were confident in their ability to conduct their own research to improve their teaching. It is curious because doing research must take more time (and skill to do it right) than simply engaging with the research of others. Perhaps this means that teachers conducing their own research are using a very limited range of skills or are not conducting the research to a high and safe standard.

From our own work, it is clear that teachers are not generally ready to make accurate judgements about the quality of prior research, even when they are enthusiastic about evidence use (Gorard et al. 2020b). Many believed that research published in a peer-reviewed journal or conducted by experts must be trustworthy. For example:

All research published in reputable journals is trustworthy

Research published by experts in the field is trustworthy because they have done extensive research in the area

Similarly, many reported that widely-used research or approaches approved by government must be of high quality. The naivete is, in some ways, astonishing:

Education programmes that are used by many schools must be good otherwise schools would not have used them

We can trust education programmes recommended by the DfE because they would have been carefully evaluated

But at heart the problem is that most teachers do not have the time or skills to understand research, and so tend to rate the importance of evidence lower than their own judgement. For example:

I find it difficult to understand research papers

The most effective teaching method is based on experience

An example of the difficulty of assessing research

The following example illustrates how hard it is for teachers and other users of evidence to decide which evidence to trust. It comes from an EEF trial of metacognition. Metacognition, like feedback, is suggested to be both cheap and effective by EEF, and by Hattie (2008) with an aggregated "effect" size of +0.69. Here the example comes from one study – a large randomised control trial (over 1,500 students) with a focus on improving maths attainment. The Executive Summary (p.5) says that "In this trial, pupils who participated in ReflectED (metacognition) made an average of four months' additional progress in maths compared to pupils who did not. For a cheap intervention (as with feedback there is no expense except teacher development) this is a substantial average gain (Figure 3).

Figure 3 – Summary table for ReflectED

Table 1: Executive Summary Table

Outcome	Number of schools	Effect Size (95% confidence)	Estimated months' progress	EEF security rating	EEF cost rating
Mathematics	30	0.30 (-0.04,0.63)	4 months	🔒🔒🔒🔒🔓	££££££

Source:https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/EEF_Project_Report_ReflectED.pdf

Figure 3 is the overall summary which teachers and school leaders might be expected to use, without looking any further. It is relatively simple and contains scale, "effect" size, padlock rating (EEF version), and cost graphics. Readers would have a right to expect that the rest of the report explains and supports this headline. The main finding is repeated in Figure 4. The "effect" size in column 7 is +0.3. This is perhaps more realistic than the higher aggregated figures for feedback or metacognition based on hyper-analyses, but still promising.

Figure 4 – Main results table for ReflectED

Table 6: Effect sizes for Primary Outcome – with and without imputation

	Raw means				Effect size		
	Intervention group		Control group				
Outcome	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	n in model (intervention; control)	Hedges g (95% CI)	p-value
InCAS Maths Score (complete cases)	800 (158)	96.87 (95.51, 98.33)	707 (130)	97.10 (95.52, 98.68)	1570 (800; 707)	0.30 (-0.04,0.63)	0.08

This is promising until more attention is paid. The figures in brackets in columns 2 and 4 are the number of cases missing. These are usually students randomised to a treatment but then not providing an outcome score. While high (over 16% of the original total), these figures are not unreasonable for a large trial. However, they do mean that the cases are no longer fully randomised – a mathematical necessity in order to compute p-values or confidence intervals. Therefore, the table can be clarified for users by ignoring or deleting the erroneous column 8, and the impossible figures in brackets in columns 3 and 5.

The two remaining bits of data are key. The mean for the intervention group (column 3) on the outcome measurement is 96.87, and the mean for the control is 97.10. Students in the control performed slightly better than the students given the treatment (training in metacognition). These figures cannot possibly be the basis for an "effect" size of 0.3. A simple randomised control trial involves randomising a large group of cases to two groups, so that they are unbiased. If all else remains constant apart from the treatment of the two groups, then it is warranted to attribute (some of) any difference in outcomes to the differential treatment. The analysis is simple because the design carries the weight of the warrant.

Here, however, the researchers have not accepted their headline findings (nor provided standard deviations so that the basic "effect" size can be checked). Instead they have done some modelling – an approach used more commonly with datasets from passive designs such as cross-sectional, with non-randomised cases. The researchers have clearly not done this because they accept that their cases are no longer randomised due to attrition, because they have presented p-values. But they have managed to convert a slightly negative result

into a strongly positive one, through an analysis that is not clear and that the EEF was not able to explain. They may, perhaps, be correct in doing so but how can users of evidence be expected to understand what has happened, and how much should we trust future meta-analyses that simply add +0.3 as the true "effect" size to their aggregated outcomes, when the control group actually did better in the trial? Note that this study was given a very high rating of 4 by EEF for its quality, even though it is incomprehensible to its intended audience.

At the outset, the EEF was a huge improvement on what had gone before, and was known for its relatively easy to understand reports accessible to users like practitioners. This was possible because, as outlined above, the power of RCTs makes both analysis and reporting simple. Now though the reports of individual studies like the one above are so complex that the original audience cannot understand them. So, it is not clear who they are intended for. In general, teachers will not have the time, inclination or skills to be able to make a suitable judgement about which studies and which syntheses of evidence to trust.

Conduits between evidence and users

A plausible-sounding solution is that teachers, and other users, are not required to select and interpret robust research. Instead, there would be a trusted intermediary who would select and translate the most trustworthy evidence (Brants and Ariel 2023). The model might be like the National Institute for Health and Care Excellence (NICE) in the UK (Perry et al. 2010), or the "Office for Educational Research" proposed by the Royal Society/British Academy (2018). Teachers might prefer teacher-related conduits such as their professional bodies (Miller et al. 2010), or education designers (McKenney and Schunn 2018). Other conduits might include knowledge brokers, think tanks, and What Works Centres or clearing houses (Edwards 2010).

However, the evidence available does not find that knowledge brokers like this are any more effective at getting good evidence into use than simple access to research (Dobbins et al. 2009). Their selection of the evidence to be promoted can be ad hoc, or even biased (Massell et al. 2012). Intermediaries are subject to the same pressures as users, and so may distort or cherry-pick evidence to suit their user clients, or pursue an agenda not clearly related to the overall evidence (Malin and Lubienski 2015). Even when the assessment of core evidence is conducted in good faith, studies have shown that different approaches and expert panels will reach different decisions about the overall evidence on any topic (Phillips et al. 2020).

The prospect of enforced evidence use

Another way forward would be for users simply to be forced to use evidence-led approaches. This is already happening in health with interventions implemented at a population level such as the fluoridation of water to prevent dental caries. This has a lower treatment cost and is more reliable since it requires no additional actions by the users (Kansagra and Farley 2012). It could be the model for curriculum and lesson resources, and for the content of textbooks (Prediger et al. 2021). Teachers could then use evidence-based approaches successfully without knowing what the evidence is (Doabler et al. 2014). This is how much successful impact from research takes place (Cain and Allan 2017). In terms of conscious teacher actions, an analogue could be the clinical guidelines for nurses (Thomson et al. 2000), which are required to be evidence-based (Oman et al. 2008), and their implementation is enforced via audit and the use of report cards (Valentine et al. 2014).

It is likely that if evidence is to improve education successfully then more of such enforced or population-level measures need to be developed, validated, and implemented. However, this approach still does not solve the central problem either. It still depends on the highest quality evidence being identified and summarised appropriately and disinterestedly by someone or some organisation. The examples in this chapter concerning enhanced feedback, metacognition and others show that even for these widespread approaches the evidence is not really that clear.

The DfE in England requires that schools show that they are using their Pupil Premium and catch-up funding effectively, by demonstrating on their websites and through inspections that they "use the wealth of evidence of 'what works', evaluated by the Education Endowment Foundation (EEF)" (Pupil premium: overview - GOV.UK (www.gov.uk)). This seems appropriate. However, as shown in this chapter, it is not clear that

schools can totally rely on that kind of evidence, or understand what it means. Nor is it clear that schools do use evidence in the way intended, rather than searching for evidence to justify their existing preferences and then putting that on their websites (Gorard 2020b).

A stronger version of the same idea appears in the guidance to providers for initial teacher training in England (Gov.UK 2022). This guidance is said to be evidence informed, and checked by EEF (again). Teacher trainers can only use evidence if it is "coherent within the framework".

A somewhat older example, also from DfE, is the requirement that all primary schools teaching literacy via systematic synthetic phonics, and only through an agreed provider (Choosing a phonics teaching programme - GOV.UK (www.gov.uk). A phonics test for children effectively enforces this approach. As with the guidance for teacher trainers, and the use of the EEF Toolkit, this enforced use of evidence is suitable only if the evidence for it is really secure.

Suggestions for improvement

In the UK, there is a strong and financially dominating push for all research, especially academic research, to have "impact" – to make a difference to those it is intended to benefit. This is apparent in appointment, progression and promotion criteria for university staff, in the demands from UKRI when funding projects, and in the rules for REF impact case studies that underpin the future QR funding of universities. However, it makes no sense for all research to have impact. Most education research is of very poor quality and could produce worse outcomes if actually acted upon. The push for evidence use in real-life must be for the use of high quality evidence only.

This then is the central question raised by this chapter – who can we trust to judge the evidence that is available to help improve education? Simply making evidence available does not appear to work (Gorard et al. 2020a). But once it is synthesised, simplified and promoted by others it may no longer be trustworthy, as the examples in this chapter suggest.

There are several steps that could be taken to help. The evidence base on how best to get evidence into use is very weak. There is a clear double standard here. Stakeholders have been demanding higher quality evidence of "what works" but, with the exception of those studies funded by EEF and cited at the start, have not generally been prepared to fund robust studies of how to translate "what works" into use. Others need to follow this valuable EEF lead and fund more robust evaluations of evidence into use. We have enough concepts, maps, theories, barriers and views of stakeholders already. It is worrying that research on this topic is so poor given that any impact relies on recognising good primary evidence.

In terms of the underlying research that provides evidence to be used in real-life, there is plenty of funding. What funders could do to help would be to stop funding so much work that is weak and does not lead to warranted conclusions. It is not clear why so much public money is wasted. Publishers and others could work harder to make sure that primary research reports are as simple and complete as possible. Users need full reporting and easy comprehension. Most education research currently has neither. This gives users a plausible (but incorrect) reason to reject the use of all evidence.

All summaries of evidence would be better if based on explicit judgements of the quality of each individual study. And meantime, aggregations of evidence that are not based on quality can be safely ignored.

REFERENCES

1. Andreassen, R. and Bråten, I. (2011) Implementation and effects of explicit reading comprehension instruction in fifth-grade classrooms, Learning and Instruction, 21(4), 520-537
2. Booher, L., Nadelson, L. and Nadelson S. (2020) What about research and evidence? Teachers' perceptions and uses of education research to inform STEM teaching, Journal of Educational Research, 113, 4, 1-13

3. Brants, H. and Ariel, B. (2023) Building bridges in place of barriers between school practitioners and researchers: on the role of embedded intermediaries in promoting evidence-based policy, Evidence & Policy, https://doi.org/10.1332/174426421X16793289365262
4. Cain, T. and Allan, D. (2017) The invisible impact of educational research, Oxford Review of Education, 43, 6, 718-732
5. Chapman, C. and Ainscow, M. (2019) Using research to promote equity within education systems: possibilities and barriers. British Educational Research Journal, 45(5), 899-917
6. Cherney, A., Head, B., Povey, J., Boreham, P. and Ferguson. M. (2013) The utilisation of social science research – the perspectives of academic researchers in Australia, Journal of Sociology, 51, 2
7. Chowl, J. and Ekholm, E. (2018) Do published studies yield larger effect sizes than unpublished studies in education and special education?, Educational Psychology Review, 30:727–744, https://doi.org/10.1007/s10648-018-9437-7
8. Coburn, C. and Talbert, J. (2006). Conceptions of evidence use in school districts: Mapping the terrain, American Journal of Education, 112(4), 469-495
9. Dieker, L., Lane, H., Allsopp, D., O'Brien, C., Butler, T., Kyger, M., ... and Fenty, N. (2009) Evaluating video models of evidence-based instructional practices to enhance teacher learning, Teacher Education and Special Education, 32(2), 180-196
10. Doabler, C., Nelson, N., Kosty, D., Fien, H., Baker, S., Smolkowski, K., and Clark, B. (2014) Examining teachers' use of evidence-based practices during core mathematics instruction, Assessment for Effective Intervention, 39, 2, 99 –11
11. Dobbins M., Hanna S., Ciliska D., Manske S., Cameron R., Mercer SL., O'Mara L., DeCorby K. and Robeson P. (2009) A randomized controlled trial evaluating the impact of knowledge translation and exchange strategies, Implementation Science, 23, 4, 61
12. Edmonds, M. (2007) Utilizing implementation data to explain outcomes within a theory-driven evaluation model, The University of Texas at Austin, Utilizing implementation data to explain outcomes within a theory -driven evaluation model - ProQuest
13. Edwards, M. (2010) Making research more relevant to policy evidence and suggestions, in Bammer, G., Michaux, A. and Sanson, A. (Eds.) Bridging the 'Know-Do' Gap (pp. 55-64), ANU Press
14. Finnigan, K. (2023) The Political and Social Contexts of Research Evidence Use in Partnerships. Educational Policy, 37 (1), 147–169
15. Flynn, N. (2019) Facilitating evidence-informed practice, Teacher Development, 23(1), 64-82
16. Gorard, S. (2020) Handling missing data in numeric analyses, International Journal of Social Research Methods, 23, 6, 651-660
17. Gorard, S. (2021) How to make sense of statistics: Everything you need to know about using numbers in social science, London: SAGE
18. Gorard, S., See, BH and Siddiqui, N. (2017) The trials of evidence-based education, London: Routledge
19. Gorard, S., See, BH and Siddiqui, N. (2020a) What is the evidence on the best way to get evidence into use in education?, Review of Education, DOI: 10.1002/REV3.3200
20. Gorard, S., Wardle, L., Siddiqui, N. and See, BH (2020b) Engagement and impact in addressing and overcoming disadvantage, pp. 136-165 in Gorard, S. (Ed.) Getting evidence into education: Evaluating the routes to policy and practice, London: Routledge
21. Gorard, S., Ventista, O., Morris, R. and See, B. (2021) Who wants to be a teacher? Findings from a survey of undergraduates in England, Educational Studies, https://www.tandfonline.com/doi/full/10.1080/03055698.2021.1915751
22. Gough, D. (2007) Weight of Evidence: a framework for the appraisal of the quality and relevance of evidence, Research Papers in Education, 22, 2, 213-228
23. Gov.UK (2022) Initial teacher training (ITT) provider guidance on stage 2, Initial teacher training (ITT) provider guidance on stage 2 (publishing.service.gov.uk)
24. Greenwood, C., Tapia, Y., Abbott, M. and Walton, C. (2003) A building-based case study of evidence-based literacy practices: implementation, reading behavior, and growth in reading fluency, K—4, The Journal of Special Education, 37(2), 95-110
25. Griggs, J., Speight, S. and Javiera, C. (2016) Ashford Teaching Alliance Research Champion: Evaluation Report, London: Education Endowment Foundation
26. Hattie, J. (1992). Measuring the Effects of Schooling. Australian Journal of Education, 36(1), 5–13

27. Hattie, J. (2008) Visible learning, London: Routledge
28. Helmberger, T. (2014) The balanced approach to literacy instruction in middle schools, Doctoral dissertation, Indiana State University, The balanced approach to literacy instruction in middle schools - ProQuest
29. Holloway, J. and Hedegaard, M. (2023) Democracy and teachers: the im/possibilities for pluralisation in evidence-based practice, Journal of Education Policy, 38:3, 432-451
30. Jacob, B. (2017) When evidence is not enough: Findings from a randomized evaluation of Evidence-Based Literacy Instruction (EBLI), Labour Economics, 45, 5-16
31. Kansagra, S. and Farley, T. (2012) Translating research into evidence-based practice, American Journal of Public Health, 102, 8, Letters
32. Kluger, A. and DeNisi, A. (1998) Feedback interventions: Towards the understanding of a double-edge sword, Current Directions in Psychological Science, 7, 67–72
33. Li, Y. and Bates, T. (2020) Testing the association of growth mindset and grades across a challenging transition: Is growth mindset associated with grades?, Intelligence, 81
34. Lord, P., Rabiasz, A., Roy, P., Harland, J., Styles, B. and Fowler, K. (2017) Evidence-Based Literacy Support: The" Literacy Octopus" Trial: Evaluation Report, Education Endowment Foundation
35. Logullo, P., MacCarthy, A., Kirtley, S. and Collins. G. (2020) Reporting guideline checklists are not quality evaluation forms, Health Science Reports, 3, 2, e16
36. Lysakowski, R., and Walberg, H. (1982) Instructional effects of cues, participation, and corrective feedback: A quantitative synthesis. American Educational Research Journal, 19: 559–578
37. Lysenko, L., Abrami, P., Bernard, R. and Dagenais, C. (2016) Research use in education: An online survey of school practitioners, Brock Education Journal, 25, 1
38. Madaleno, M. and Waights, S. (n.d.) Guide to scoring methods using the Maryland Scientific Methods Scale, https://whatworksgrowth.org/public/files/Scoring-Guide.pdf
39. Mai D, and Brundrett M. (2022) The importance of developing teachers as researchers in the new general education curriculum of Vietnam, Management in Education, doi:10.1177/08920206221093001
40. Malin, J. and Lubienski, C. (2015) Educational expertise, advocacy, and media influence, Education Policy Analysis Archives, 23, 6
41. Massell, D., Goertz, M. and Barnes, C. (2012) State Education Agencies' Acquisition and Use of Research Knowledge for School Improvement, Peabody Journal of Education, 87:5, 609-626
42. McCartney, E., Marwick, H., Hendry, G. and Ferguson, E. (2018) Eliciting student teacher's views on educational research to support practice in the modern diverse classroom: a workshop approach, Higher Education Pedagogies, 3,1, 342-372
43. McKenney, S. and Schunn, C. (2018) How can educational research support practice at scale? Attending to educational designer needs, British Educational Research Journal, 44, 6, 1084-1100
44. Miller, S., Drill, K. and Behrstock, E. (2010) Meeting Teachers Half Way: Making Educational Research Relevant to Teachers, Phi Delta Kappan, 91(7), 31–34
45. Morris, P. (2019) Misunderstandings and omissions in textbook accounts of effect sizes, British Journal of Psychology, https://doi.org/10.1111/bjop.12401
46. Nägel, L., Bleck, V. and Lipowsky, F. (2023) Research findings and daily teaching practice are worlds apart, Teaching and Teacher Education, 121, 103911
47. Nutley, S., Boaz, A., Davies, H. and Fraser, A. (2019) New development: What works now? Continuity and change in the use of evidence to improve public policy and service delivery, Public Money & Management, 39:4, 310-316
48. Oman, K., Duran, C. and Fink, R. (2008) Evidence-based policy and procedures: an algorithm for success, The Journal of Nursing Administration, 38, 1, 47-51
49. Perry, A., Amadeo, C., Fletcher, M. and Walker, E. (2010) Instinct or Reason: How education policy is made and how we might make it better, CfBT, https://www.educationdevelopmenttrust.com/~/media/EDT/Reports/Research/2010/r-instinct-or-reason-2010.pdf
50. Phillips, P., Castle, D. and Smyth, S. (2020) Evidence-based policy making: determining what is evidence, Heliyon, 6(7), e04519

51. Prediger, S., Barzel, B., Hußmann, S., and Leuders, T. (2021) Towards a research base for textbooks as teacher support: The case of engaging students in active knowledge organization in the KOSIMA project, ZDM–Mathematics Education, 53, 1233–124
52. Rose, J., Thomas, S., Zhang, L., Edwards, A., Augero, A. & Roney, P. (2017) Research Learning Communities: Evaluation Report, Research_Learning_Communities.pdf (educationendowmentfoundation.org.uk)
53. Royal Society/British Academy (2018) Harnessing educational research, https://royalsociety.org/~/media/policy/projects/rs-ba-educational-research/educational-research-report.pdf
54. See, BH, Gorard, S. and Siddiqui, N. (2016) Can teachers use research evidence in practice?: A pilot study of the use of feedback to enhance learning, Educational Research, 58, 1, 56-72
55. Skiba, R., Casey, A. and Center, B. (1985) Nonaversive Procedures in the Treatment of Classroom Behavior Problems, The Journal of Special Education, 19(4), 459–481
56. Speight, S., Callahan, M., Griggs, J. and Javiera, C. (2016) Rochdale research into practice: evaluation Report, London: Education Endowment Foundation
57. Taylor, B., Pearson P., Peterson D. and Rodriguez M. (2005) The CIERA School Change Framework: An evidence-based approach to professional development and school reading improvement, Reading Research Quarterly, 40, 1, 40-69
58. Thomson, P., Angus, N. and Scott, J. (2000) Building a framework for getting evidence into critical care education and practice, Intensive and Critical Care Nursing, 16, 3, 164-174
59. Tortorelli, L. and Bruner, L. (2022) The Word Nerds project: Findings from a research–practice, Journal of Research in Reading,
60. Valentine, A., DeAngelo, D., Alegría, M. and Cook, B. (2014) Translating disparities research to policy: A qualitative study of state mental health policymakers' perceptions of mental health care disparities report cards, Psychological Services, 11, 4, 377-387
61. Washburn, E., Gesel, S., Fitzgerald, M., Beach, K. and Kingsbery, C. (2023) The Impact of a Comprehensive, Evidence-Based Approach to Summer Literacy Intervention on the K-3 Reading Skills of Economically and Culturally Diverse Students, Reading and Writing Quarterly, 10.1080/10573569.2022.2147463
62. Wiggins, M., Jerrim, J., Tripney, J., Khatwa, M. and Gough, D. (2019) The RISE project: evidence-informed school improvement, Evaluation report, The RISE Project: Evidence-informed school improvement | EEF (educationendowmentfoundation.org.uk)
63. Wolf, R., Morrison, J., Inns, A., Slavin, R. and Risman,K. (2020) Average effect sizes in developer-commissioned and independent evaluations, Journal of Research on Educational Effectiveness, 10.1080/19345747.2020.1726537