Chapter 7 - Why is it difficult to get evidence into use?

Beng Huat See

In the UK, when the new Labour government came into office in 1997 there was considerable talk about evidence-informed policy and practice (Cabinet Office, 1999). Ambitious initiatives were launched and a number of strategies were suggested to encourage the use of evidence (Bullock et al., 2001; Davies et al., 2000; Nutley et al., 2002). This drive has continued to some extent with successive governments. One development was the establishment of the Education Endowment Foundation (EEF) in 2010, a charitable organization funded by the government, whose aim is to generate good quality evidence to support the academic attainment of disadvantaged children in England. Over 20 years, education research has progressed, and we are now seeing more high quality randomised control trials and systematic reviews being conducted in education.

Consequently, there is now a more substantial body of evidence to draw upon to inform classroom teaching. This partial improvement also means we now have more of a mix of research quality, some good and some still poor. Schools wanting to use primary research evidence may therefore have to read scientific literature and make judgements about how much to trust each study. There are three problems with this. First school leaders and teachers generally do not have the time to engage with academic literature. Second, even if they have the time, few school leaders and teachers are trained to assess the quality of research evidence to be able to distinguish trustworthy evidence from unwarranted claims. Third, academic research papers are often not written in a way that are easily accessible to practitioners (See, Gorard and Siddiqui 2016). Of course, it can be argued that it is not necessary for educators to be able to read, assess and understand evaluations of educational programmes, because this can be done for them by others (Slavin 2019).

This chapter sets out some of the practical challenges in getting research evidence into use by schools. It identifies where the sources of good quality evidence for education might be found, outlines the limitations of evidence from these sources, identifies barriers to using research evidence, and finally suggests possible ways to improve the quality of evidence and its take-up.

Where to find good quality evidence?

Several organisations and clearinghouses exist whose job is to synthesise research evidence and make it publicly and freely available in usable forms, clarifying what seems to work in practice and what does not. Examples of these evidence portals include the UK Education Endowment Foundation (EEF) website (https://educationendowmentfoundation.org.uk/), the US Institute of Education Sciences' What Works Clearinghouse (WWC) website (https://ies.ed.gov/ncee/wwc/), the Evidence for ESSA (Every Student Succeeds Act) website (www.evidenceforessa.org) and the Best Evidence Encyclopaedia (BEE). In evidence portals, unlike evidence databases, the evidence has been translated by the portal manager to make it accessible to consumers of evidence, such as practitioners and policy-makers,

For example, the EEF maintains a comprehensive website, with information for schools and teachers looking for programmes they can use to improve the learning outcomes of their pupils, particularly those from disadvantaged backgrounds. On their website are the Teaching and Learning Toolkit (T&L) (https://educationendowmentfoundation.org.uk/public/files/Toolkit/complete/EEF-Teaching-Learning-Toolkit-October-2018.pdf) and its companion Early Years Toolkit (https://educationendowmentfoundation.org.uk/evidence-summaries/early-years-toolkit/). These Toolkits summarise international evidence on teaching, providing summary information on the cost, impact and the strength of evidence for a wide range of education practices, including Arts education, behavioural interventions, parental engagement, one-to-one interventions, metacognition, use of teaching assistants, and small group tuition.

The evidence for the Toolkits is obtained by summarizing or averaging the results of meta-analyses of studies conducted internationally. This leads to an aggregated "effect" size for programmes or interventions, accompanied by an estimate of the security of the finding presented as a number of padlocks. These indicate how trustworthy the overall evidence on any topic is judged to be. This should be useful for schools when deciding which programmes to invest their funding in.

The Toolkit also identifies approaches or practices that are in need of further evidence, and a number of these are being tested in independently evaluated trials. At the time writing, 111 evaluations have been completed, of which 16 are identified as promising, and another 79 projects are still in progress. Schools interested in improving the learning and wider outcomes of their pupils can simply go to the website and look for approaches or practices that are relevant to their needs. For more details, refer to the EEF Toolkit Manual (EEF 2018).

Another source of evidence is the Early Intervention Foundation (EiF) Guidebook (https://www.eif.org.uk/). The EiF is a charity organisation established in 2013 to champion and support the use of effective early interventions to improve the lives of children and young people at risk of poor outcomes. Like the EEF, EiF is one of 7 independent What Works Centres set up to create, share and use high quality evidence in policy and practice. The focus of EiF is on the development of a child from birth to age 18 including physical, cognitive, behavioural, social and emotional development. Unlike the EEF, EiF does not provide funding for research, trials or evaluations. They conduct research and synthesise evidence from trials and evaluations and disseminate the findings. They produce resources to translate research into practical guidance and tools. In other words, their concern is the promotion and translation of evidence into practice and policy.

The Work Works Clearinghouse (WWC) website is another good place to look for programmes that have been tested and evaluated. This is produced by the US Institute of Education Sciences. Unlike the EEF Toolkits, the WWC reviews single studies (as opposed to meta-analyses) of existing research on programmes and practices in education. This is preferable, as it allows separate judgements to be made about the quality of each study. It aims to provide scientific evidence on what works to improve student outcomes. It covers a whole range of topics including attainment and behaviour, and programmes for different phases of education as well as for children with special needs. For each, WWC assesses the overall evidence on the effectiveness of the programme.

WWC also publishes practice guides with programmes for educators to use to address challenges they face in their schools and classrooms (https://ies.ed.gov/ncee/wwc/Publication#/ContentTypeId:3). Each practice guide comes along with instructional tips that teachers can use in their classrooms, a summary of the evidence that supports the instructional tips and a summary of the practice guide recommendations.

An additional feature of WWC is the Study Review Guide (SRG). The SRG is a tool developed by the WWC to be used by trained and certified WWC reviewers to assess studies against WWC design standards. Reviews of studies using the SRG underlie all What Works Clearinghouse (WWC) products. As part of an ongoing effort to increase transparency, promote collaboration, and encourage widespread use of the WWC design standards, IES provides external users with access to a public-use version of the SRG. The public version of the SRG is an online application that guides a user through documenting the characteristics of a study, including features that pertain to a study's eligibility under a WWC protocol. It also assists users in assessing the study design and implementation against the WWC standards, and coding the study findings in a systematic manner consistent with WWC reporting guidelines. The SRG is intended to be used by individuals trained and certified in WWC review policies and procedures, in conjunction with WWC review protocols and the Procedures and Standards Handbook.

The Evidence for ESSA and the Best Evidence Encyclopaedia (BEE) websites are both developed by the Centre for Research and Reform in Education, Johns Hopkins University School of Education. BEE provides summary reviews of a range of education programmes for use with children from kindergarten to primary and secondary schools. Each programme is given a rating according to the strength of evidence of its effectiveness in improving pupils' outcomes. The five levels of rating are:

- No evidence (that is no studies that met the inclusion criteria were found)
- Limited weak evidence with notable effects (limited number of studies or small sample)
- Limited strong evidence with modest effects (sufficient number of studies of adequate size, but mean effect size is +0.10 to +0.19
- Moderate evidence (two large matched comparison studies or many smaller studies with total sample of at least 500 students and mean effect size of +0.20
- Strong evidence (at least one large RCT and multiple small studies with total sample of at least 500 students and mean effect size of +0.20

Evidence for ESSA is a free web-based resource that provides information about reading and maths programmes that meet the ESSA standard of evidence. ESSA classifies evidence as strong, moderate and promising for programmes and practices with at least one significantly positive outcome in an RCT, quasi-experimental/matched study or correlational study respectively. For each progamme on the website a brief description is provided of what the programme looks like, the cost, the phase of education and the level of evidence under ESSA. This website is thus designed for educators looking for 'proven' programmes or evidence about a particular programme.

How good are such portals and how effective are they in getting evidence into use in practice?

EEF Toolkits and syntheses

The EEF Teaching and Learning Toolkits rate each progamme or practice to indicate how reliable or trustworthy the evidence is using a padlock system (the 🔒 icon**).** This is assessed by taking into account the number of research studies or meta-analyses available, the consistency of the impact estimated across the studies synthesized, the strength of causal inference provided by those studies and the outcomes measured. The ratings range from very limited (no evidence reviews available), limited, moderate, extensive to very extensive.

An approach/practice is considered to have very limited evidence if no systematic reviews have been conducted on it. Limited evidence is one where there is at least one review of studies that use methods to allow weak conclusions to be drawn. An approach is considered to have moderate evidence of impact if there were at least two systematic reviews to allow for moderate conclusions to be drawn about the impact. One with extensive evidence would have at least three systematic reviews of studies that use strong research design and where the impact estimates are broadly consistent across studies. An approach is considered to have very extensive evidence of impact if it has been evaluated in at least five reviews where the impact estimates are consistent across all studies, and the studies in the review use research designs that allow strong conclusions to be drawn about impact. Figure 1 summarises the interpretation of the strength of evidence used in the T&L Toolkits.

Figure 1: Interpretation of strength of evidence

🔒 = **Very limited evidence**: No evidence reviews available, only individual research studies.

🔒 🔒 = **Limited evidence**: At least one evidence review. Reviews include studies with relevant outcomes, and studies with methods which enable researchers to draw weak conclusions about impact.

🔒 🔒 🔒 = **Moderate evidence**: At least two evidence reviews. Reviews include studies with relevant outcomes, and studies with methods and analysis which enable researchers to draw moderate conclusions about impact.

🔒 🔒 🔒 🔒 = **Extensive evidence**: At least 3 evidence reviews. Reviews include studies with highly relevant outcomes, and studies with methods and analysis which enable researchers to draw strong conclusions about impact. Impact estimates are broadly consistent across studies.

🔒 🔒 🔒 🔒 🔒 = **Very Extensive evidence**: At least 5 evidence reviews. Reviews are recent, and include studies with highly relevant outcomes, and studies with methods and analysis which enable researchers to draw strong conclusions about impact. Impact estimates are consistent across studies.

Source: Education Endowment Foundation (https://educationendowmentfoundation.org.uk/evidence-summaries/about-the-toolkits/evidence-strength/

The Education Endowment Foundation's Teaching and Learning Toolkit was a revolutionary tool for educators in the UK when it was first devised. It has made access and understanding of findings from evidence synthesis of effective approaches easy for users. As Howard White said in his chapter the EEF Toolkit and the IES What Works Clearinghouses (WWC) are excellent examples of good practice in getting evidence into use.

However, as with all meta-analyses and hyper-analyses, there are a few limitations. Because the EEF T&L Toolkit considers the number of systematic reviews as an indication of the strength of evidence, it s possible that an effective approach may be erroneously rated as having limited evidence simply because there have been no systematic reviews conducted on it, even though there may be many individual robust studies. Therefore, the EEF T&L toolkit now includes individual studies that have not been considered in any of the meta-analyses. But the core evidence so far is still based on summaries of systematic reviews. As such, the Toolkit can only be considered a guide, and the security rating may not accurately reflect the strength of evidence for each approach. Nevertheless, this is a big move forward in evidence-based education. When the T&L Toolkit was first conceived there was practically nothing that teachers could use as a guide. The Toolkit represented the best evidence available at that time. The Toolkits should be treated as a 'live resource' (p. 6) as recommended by the authors of the Toolkits, and as the EEF carries out more robust evaluations of these approaches, new evidence comes in and the security ratings will be revised. The authors are now adding evidence from single studies to the evidence base, but this will take some time to develop (see Chapter 6).

Another concern about the T&L Toolkit is that although it also gives information about how effective an approach is (which is useful), this is measured by averaging the effects from systematic reviews, which are themselves a collection of studies. As such there is the potential for error propagation. Also, different calculations of effect sizes are used in different studies and across systematic reviews. For this reason, T&L used the weighted mean effect size, which assumes a common effect size among the included studies. Where there is insufficient data to calculate a weighted mean and thus the potential to overestimate effect because of the likelihood of positive studies reporting large effects, the median effect size is used. However, meta-analyses that did not include standard errors or sufficient data to calculate standard errors are excluded. In some cases indicative effect sizes are calculated. Therefore, the effect sizes reported in the T&L Toolkit can be quite arbitrary.

Additionally studies included in the systematic reviews may be quite varied in terms of research quality. For example, some may be randomised control trials, some may be correlational/observational studies and others may use one group pre-post comparisons. While experimental and quasi-experiment studies (studies where comparison groups are not randomly allocated) are preferred, correlational studies may also be included. It is not clear if the systematic reviews included in the evidence assessment for the Toolkits weight the studies by their research design, such as giving the effect sizes from experimental studies more weight than correlational studies. This means that the evidence may be skewed towards some types of studies. For example, studies using correlational designs and quasi-experimental designs, and those with small sample size and high attrition tend to show bigger 'effects' than large-scale randomised control studies (Slavin and Smith 2009). Also studies that use bespoke or intervention-related instruments to measure outcomes are likely to show bigger effects than those that use treatment independent or standardized measures (Slavin and Madden 2011). A good example is Hattie's synthesis of over 800 meta-analyses (Hattie 2008) which averages the effects of studies for different age groups using different measures of outcomes. The authors of the Toolkits recognize that the meta-analyses are not always consistent in how they have dealt with these factors that influence effect sizes (EEF 2018) (Higgins 2016). In summary, the effect sizes may not be an accurate reflection of the impact of some of the approaches.

Further, the Toolkit meta-analyses assumed that the studies in the systematic reviews they included in their synthesis had taken into account attrition. It is not always clear how attrition or missing data are dealt with in most studies. Typically, researchers use imputation or assume that missing values are random and ignore cases with key missing values. In reality missing cases and missing data are usually not random. Pupils who are missing from a post-test may be excluded because they are long-term sick, school avoiders, excluded from school for poor behavior, low-performing or have special needs. Thus, excluding them could distort any effect size. On the other hand using the data that is there to try and compensate for the data that is missing can exacerbate this distortion.

Therefore, the Toolkits should be regarded as a guide and should be used together with the evidence from individual EEF trials. The T&L Toolkits are used here only for illustration. Similar comments would apply to any approach based on such “hyper-analyses” (Gorard 2018).

WWC and single-studies

The WWC also provides effectiveness rating and strength of evidence for educational programmes, but unlike the EEF Toolkits, these are based on summaries of individual studies. The effectiveness rating (similar to the padlock rating used by the EEF) appears as a ruler -- - 0 +- + ++. It is based on the quality of the research, ‘statistical significance’ of the findings, and the magnitude and consistency of findings across studies. There are six categories, from positive, potentially positive, mixed, no discernible impact, potentially negative to negative (Figure 2). The number of positive signs (+) indicates the strength of evidence of positive impact. Conversely, the number of negative signs (-) indicates the strength of evidence of a negative effect.

Figure 2: Interpretation of effectiveness rating

-- - 0 +- + ++ *Positive:* There is strong evidence that the intervention had a positive effect on outcomes

-- - 0 +- + ++ *Potentially positive:* There is evidence of positive effect on outcomes with no overriding contradictory evidence.

-- - 0 +- + ++ *Mixed:* The intervention effect on outcomes is inconsistent

-- - 0 +- + ++ *No discernible:* No evidence that the intervention had any effect on outcomes

-- - 0 +- + ++ *Potentially negative:* There is evidence that the intervention had a negative effect on outcomes with no overriding contradictory evidence.

-- - 0 +- + ++ *Negative:* There is strong evidence that the intervention had a negative effect on outcomes

An intervention with "positive effects" means there is strong evidence of a positive effect based on the studies reviewed. As with the T&LToolkits the authors do not mean that the intervention will work in all settings for all students. For example, a one-to-one intervention tested with primary school children may not be effective with secondary school children or delivered as small groups or pairs. Some educational interventions are not given an effectiveness rating. This does not mean that they do not work, but rather that there has been little or no research which meets WWC design standards, so WWC is unable to rate the effectiveness of the intervention.

WWC only includes studies that use an experimental or quasi-experimental design. Literature reviews and meta-analyses are therefore excluded. Each study is then assessed on the credibility of the evidence based on the design, sample, attrition, and the evidence of equivalence or non-equivalence of the intervention and comparison groups prior to the intervention. The three possible ratings are: Meets WWC Group Design Standards without Reservations, Meets WWC Group Design Standards with Reservations, and Does Not Meet WWC Group Design Standards.

The WWC also takes into account confounding factors in their assessment of strength of evidence. For example, an intervention may be offered in addition to other interventions, or the intervention may be offered to all pupils in one school but not in the comparison school. Sometimes all intervention pupils are taught by one teacher. It is therefore not possible to attribute the outcome to the intervention alone. The differences in outcome may be due to the additional intervention, better quality of teaching or both. In such cases, the study is said to have not met the WWC standards. For this reason, quasi-experimental studies, which usually have confounding factors because of unobserved differences between groups, are not given the highest evidence rating.

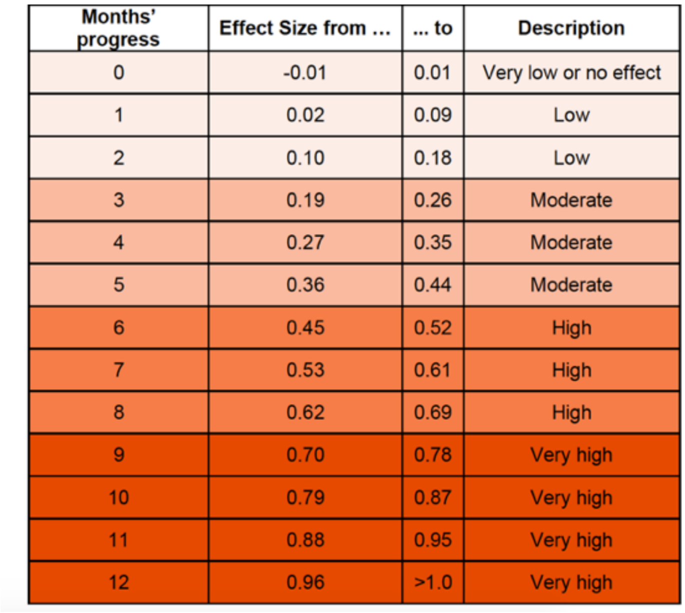
However, the WWC also reports on the statistical significance of the effect size estimates, which they define as one where "the probability of observing such a result by chance is less than one in 20" or $p = 0.05$. This is both confusing for users, and scientifically problematic. The EEF sensibly do not use significance tests or p-values in presenting the uncertainty of results in the Toolkit because it is very difficult to communicate precisely what a p-value (and hence a significance test) actually means. P-values are very commonly construed, even among experts, as the probability that the intervention had no impact, given that you observed a difference as extreme as the one that was actually found. In fact, a p-value means the probability of observing a difference as extreme as (or more extreme than) the one that was actually found, assuming the intervention had no impact. This mistake leads the unwary, including the WCC in their definition above, to conclude that a p-value of 0.05 means that there is only a 5% chance that a positive finding is due to chance, which is not true.

Confusing effect size and research quality

While portals like EEF and WWC report the strength of the evidence and the size of the impact, these cannot always be easily interpreted. First, it is very common (even for academics) to confuse effects with strength of evidence. There is a tendency to relate strong evidence with large impact, whereas studies that have strong evidence may suggest no impact, while studies with large impact may have weak evidence. Second, effect sizes should also be interpreted in context as the rate of progress pupils make varies throughout school, across subjects and age (Baird and Pane 2019; Bloom et al. 2008). Older children, for example, make less progress as they get older (Bloom et al. 2008). According to the DfE data (DfES 2004) children made annual gains of an effect size of 0.8 at the end of Key Stage 1 (age 7) dropping down to 0.4 at the end of Key Stage 3 (age 14). What this means is that a small gain made by older children may be more important than bigger gains made by younger children, or vice versa.

To aid interpretation of what effect sizes mean in terms of children's progress, they are converted to months in progress by EEF (see Figure 3), but it is misleading to use the same conversion for all children and for all subjects. This translation also does not make sense for non-cognitive measures (e.g. children's wellbeing or happiness). Baird and Pane urge caution in using such conversions. As Kvernbekk (2015) explains, what matters is whether a programme or practice leads to improvement in outcomes (whatever this may be).

Figure 3: Conversion of effect size to month's progress



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

All of the issues so far tend to make it harder for users either to simply take results on trust or to make their own (valid) judgements about the strength of any evidence.

Challenges in using published evidence

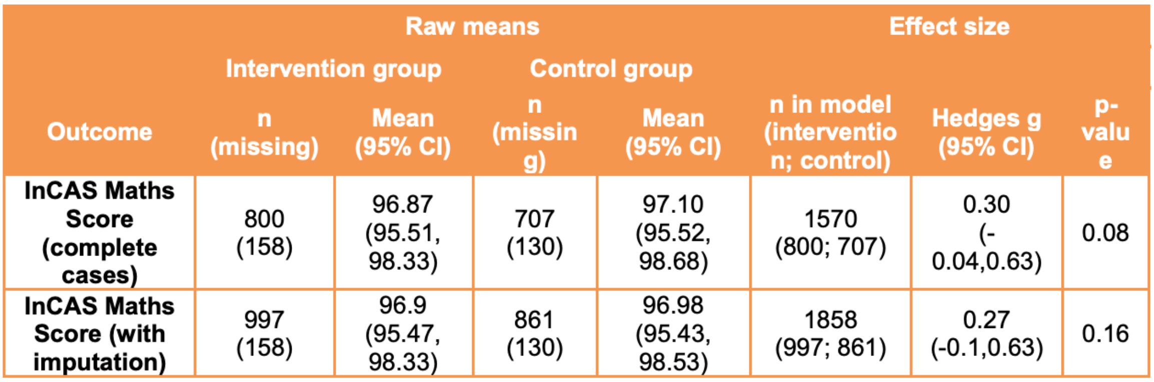
A further challenge for users is that even if we know how to interpret effect sizes there is still the issue of understanding whether the research findings are trustworthy or not. As an illustration of the challenges in using published research, we present a few examples from some popular education programmes that have been "rigorously" tested in randomised control trials and highlight two very common issues: conflicting results and confusing reporting. What do we do when two similarly well-conducted trials show different results and how do we interpret the findings of a study where the data presented is at variance with the results reported?

Confusing reporting

One programme evaluated by the EEF and classified as "promising" is ReflectED (Motteram et al. 2016). The programme aims to improve children's metacognition, defined as the ability to think and manage their own learning. The randomised control trial involved 1,858 Year 5 pupils in 30 schools where teachers within the schools were randomised either to be trained to teach children metacognition strategies or to usual practice. The EEF website showed that the programme improved children's maths performance equivalent to four months' progress (effect size of +0.30) with a 4-padlock rating, suggesting that the finding is very secure. Although the report suggested that there were 30 schools, the website showed that there were only 24 schools (https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/reflected-meta-cognition/). Attrition was reported as 15%.

What is intriguing is that the main headline findings in Table 1 (taken from the report, https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/EEF_Project_Report_ReflectED.pdf) show that in maths the control group (mean score 97.10) actually does better than the intervention group (mean score 96.87). Nowhere does the report explain how this then represents an effect size of +0.30 in favour of the intervention group. No pre-scores are specified. A mistake has been made somewhere.

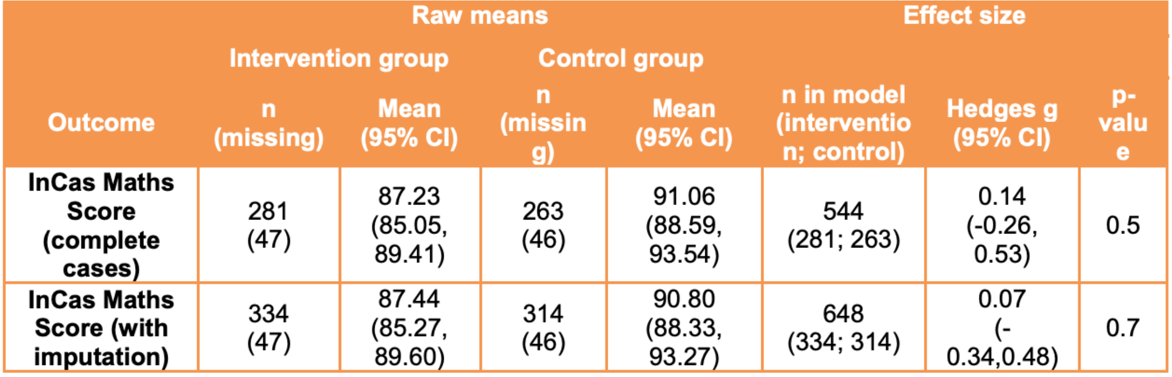
Table 1: Effect sizes for all students



	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
InCAS Maths Score (with imputation)	997 (158)	96.9 (95.47, 98.33)	861 (130)	96.98 (95.43, 98.53)	1858 (997; 861)	0.27 (-0.1,0.63)	0.16

For FSM-eligible children (Table 2), the control group (mean score 91.06) is also ahead of the intervention group (mean score 87.23). Regardless of whether the imputed or non-imputed data are used, the effect sizes would still be negative given that the control group had higher scores than the intervention group. But both tables indicate purportedly positive effects.

Table 2: Effect sizes for free school meals pupils



	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)	281 (47)	87.23 (85.05, 89.41)	263 (46)	91.06 (88.59, 93.54)	544 (281; 263)	0.14 (-0.26, 0.53)	0.5
InCas Maths Score (with imputation)	334 (47)	87.44 (85.27, 89.60)	314 (46)	90.80 (88.33, 93.27)	648 (334; 314)	0.07 (-0.34,0.48)	0.7

These tables are busy, crowded and confusing and yet they still do not include crucial information. They do not show any pre-test data, and they do not include the standard deviations for the mean scores of either group or overall. Yet, the standard deviation is key to converting the difference between means to a Hedges' g effect size. The reader can see that the effect sizes quoted are wrong (because whatever the standard was, the negative difference between means cannot be turned into a positive result), but they cannot work the effect sizes out for themselves. The data presented is therefore unhelpful as it does not tell us how the effect size came about. If the control group had higher scores than the treatment group, how then can the effect be positive?

Each table includes a column for p-values (from significance tests, see above), which are just about impossible for others to comprehend, and which should not be used with non-randomised cases (here around 20% of cases are missing just in cell 1 of Table 1, for example). For the same reasons, specifying a confidence interval (CI) is invalid. Taking such clutter out of the table would make it easier to read, and so easier for teachers to understand and make an appropriate judgement about.

We have asked EEF several times to explain these bizarre results and they have reported asking the evaluators to explain these tables, but otherwise we have no reply. This trial was given a 4-padlock rating by EEF, indicating that the results are very reliable. Based on the data which show negative effects (but reported as positive), EEF identified the programme as promising and have commissioned a larger effectiveness trial. If the EEF and the evaluators cannot explain this confusing finding, what chance do school teachers have in understanding the data? And this work has been extensively peer-reviewed, was conducted by recognised evaluators and must be among the best work available.

This is not an isolated example. It is, in fact, rare and welcome to read an evaluation report in the UK without such needless complexities.

Mixed results

Where different evaluations give mixed or conflicting results this can also be confusing for teachers. The clearest result in both John Hattie's super meta-analyses (Hattie 2008; 2017) and the Toolkit suggests enhanced feedback as one of the most effective classroom approaches. Yet the syntheses include around one third of studies with negative findings suggesting that feedback is ineffective. But interestingly, many schools and teachers are still citing Hattie's meta-analysis as evidence that feedback is effective. As explained above, context is important. It is not simply using feedback, but the kind of feedback, the way it is implemented and the age of children are all important factors to consider (See, Gorard and Siddiqui 2016). School leaders and teachers using such research evidence would have to take the evidence in good faith as it is reported. We would not expect them to scrutinise the academic writing (which is often difficult to read anyway) to examine the context.

On the other hand, there are some programmes which show effects for one age group but not for another. Similarly, effects may be seen in one component of the outcome (e.g. Reading), but not in another (e.g. Writing). Take for example, the EEF evaluation of the Switch-on programme (Gorard, See and Siddiqui 2014). Switch-on (SO) is a literacy programme inspired by Reading Recovery. The efficacy trial involved 308 Year 7 pupils (first year of secondary) from 15 schools. Pupils were individually randomised to SO or usual practice. The results showed positive effects of SO on pupils' reading (effect size +0.24 equivalent to +3 months' progress). Reading outcomes were measured using the GL Assessment New Group Reading Test (NGRT), an independent standardised test. Attrition was 2%. This project was given a security rating of 3-padlocks – lower than the ReflectED trial.

A follow-up effectiveness trial of Switch-on randomised 184 schools rather than pupils, and placed them in three groups rather than two (Patel et al. 2017) – meaning that the trial randomised fewer 'cases' to the smallest group than Gorard et al. (2015). There was a low correlation between the baseline measurement and the post-test, and higher attrition from the control group (11%) than the treatment group. It was given a security rating of 4-padlocks. This study found little or no benefit from Switch-on. The presentation of results again does not show the standard deviation, making it difficult for an interested person to calculate the effect size.

Given the conflicting findings from these two trials, how should schools interpret the evidence. Should schools use SO or not to improve the literacy of their pupils? Is the evidence from the effectiveness trial (rated 4 padlocks) more credible than that of the efficacy trial (rated only 3 padlocks)? It is essential that schools recognise that there are important differences between the two trials. First, the efficacy trial was tested with first year secondary school pupils while the effectiveness trial was tested with primary school children. Second, the efficacy trial used the GL NGRT test to measure reading outcome, while the effectiveness trial tested reading and writing using the Hodder Group Reading Test. Crucially, because the second trial was an effectiveness trial, this meant that there was minimal monitoring by developers. Some schools reported modifying aspects of the programme for delivery. This could have an impact on the results, and may suggest that for any programme to replicate the effects of the trials, schools have to adhere to the programme protocol as closely as possible.

Another programme that was evaluated and showed conflicting results is a programme known as IPEEL, which stands for Introduction, Point, Explain, Ending, Links, and Language. IPEEL is adapted from the American Self-Regulated Strategy Development (SRSD). The efficacy trial (Torgerson et al. 2014) reported positive effects (ES +0.74) with larger benefits for free school meal (FSM) eligible children (ES +1.60). However, it is not possible to recalculate these scores because the report provided no mean scores and standard deviations. A larger follow-up effectiveness trial by the same team (Torgerson et al. 2018) reported mixed results - negative impact on writing after one year (ES -0.09) and positive impact after two years (ES +0.11). It is possible that children need two years of exposure to see any benefit from the programme. However, the authors explained that different writing outcomes were used in the one-year and two-year trials which could account for the difference in results. The one-year trial measured writing outcome on a categorical scale (and was given a 5 padlock rating, the highest security rating possible), while the two-year trial recorded impact on a continuous outcome (and given 3 padlocks).

As with the Switch-on evaluation, the efficacy and effectiveness trials did not use the same age group of children. In the efficacy trial the intervention was tested on Year 5 (age 9-10) and Year 6 (age 10-11) children, while the effectiveness trial was tested on children in the transition phase in Year 6 (age 10-11) and Year 7 (age11-12). Another key difference was that in the efficacy trial the programme was delivered by developers of IPEEL whereas in the effectiveness trial the intervention was delivered by teachers who were trained by newly recruited trainers. This is likely to affect the quality of the delivery and thus the outcomes. In addition, side effects on other outcomes were also noted. While the children improved in their writing after two years, their maths and reading suffered.

It is quite unlikely that teachers, reading the EEF evidence, would pick out such crucial differences in the trial to make sense of the findings. For one thing, they would not have the time to read the whole report. They would probably just look at the headline findings, and the side effects are also rarely highlighted.

Sometimes different evaluations also show up different results depending on the kind of analysis performed. For example, an evaluation of a literacy programme, known as Writing Wings showed contradictory results (Madden et al. 2011). Using hierarchical linear modelling the programme showed no effect, but analysis of covariance showed small positive effect sizes for some outcomes. The overall results are therefore inconclusive.

Even academics are perplexed by how research findings are reported. If I were a teacher, knowing what I know now, I would be very sceptical of almost any research evidence. Therefore, to encourage the use and uptake of research evidence, the research community must first ensure that the quality of research is impeccable, research data are clearly and ethically reported, and finally the findings must be presented simply, clearly and accurately.

Conclusion

This chapter introduces some sources of professional publications that teachers and school leaders can use to improve their practice and their pupils' learning and wider outcomes. The US Institute of Educational Sciences What Works Clearing House and the UK's Education Endowment Foundation websites are two highly regarded avenues that practitioners and researchers can go to for high quality evidence. But, as illustrated, even then there are challenges in utilising such evidence. There is the issue of interpreting the research findings (does it work or does it not) and the reliability of the evidence. The conflicting results, incomplete or confusing reports of results, and inconsistent security ratings, can be quite a minefield for anyone trying to use published research evidence. And even if one understands how to interpret research findings, the findings do not always apply to all conditions or context. Efforts to get teachers and school leaders to use evidence cannot work if research evidence itself is impossibly hard to understand for academics. This is probably the chief factor impeding appropriate evidence use by teachers (Stanovich and Stonovich 2003).

Therefore, one step towards evidence use in the classroom is to equip teachers and school leaders with the tools for evaluating the credibility of these many and varied sources of information. This would require training of teachers, and should start at the point of initial teacher training. Teachers and school leaders need to be appropriately sceptical of research findings and make professional judgements about what works for them and for their pupils. However, this may be too large a task.

More crucially researchers and funders need to improve and simplify their results. Simplifying is often the same thing as improvement in clarity, as illustrated in this chapter. Authors and evaluators have to be scrupulous in presenting the results of their research and making the evidence transparent. Only recently the authors of a paper published in the Journal of American Medical Association (JAMA) publicly retracted their paper (Aboutmatar and Wise 2019) as they had made a mistake in the original paper. The mistake came about when they recoded the variable referring to the study intervention group assignment. In the recoding they reversed the coding of the study groups, meaning that the intervention was wrongly coded as control and the control group as the intervention. The original paper reported that the intervention reduces the number of hospitalisations. When they realised their mistake, they redid the analysis, retracted the paper and republished the paper which totally reversed their original findings. Such mistakes do happen, and in medical science such mistakes can cost lives. This is what ethical research scientists would do – admit their mistakes, be honest about it and rectify the situation. I have total respect for these people. However, in social science this is rare. Even when academics have been shown to be wrong, they continue to obfuscate so that their mistakes are not obvious to all except the most dedicated readers. Most educational research is so poor though that it is 'not even wrong'. And this is what makes it difficult for consumers of research to trust what they read – another barrier to the use of research evidence.

If we want teachers to use evidence-based programmes, the evidence has to be scrupulous so that teachers can be confident that they are using the right tool or programme which will benefit their pupils. As the well-known proverb in the mid-1500s goes, "You can't make a silk purse out of a sow's ear." Therefore, to improve children's learning and wider outcomes, schools need to use programmes that have demonstrable effectiveness. If research evidence is difficult to assess, wrong, or not even wrong, then schools will continue to use classroom programmes of unknown effectiveness or even known to be ineffective. There is a knock-on effect from poor and poorly- reported research.

References

Aboumatar, H. and Wise, R. (2019). Notice of Retraction. Aboumatar et al. Effect of a Program Combining Transitional Care and Long-term Self-management Support on Outcomes of Hospitalized Patients With Chronic Obstructive Pulmonary Disease: A Randomized Clinical Trial, *JAMA*, 322,14, 1417-1418. doi:10.1001/jama.2019.11954

Baird, M.D. and Pane, J.F. (2019). Translating standardised effects of educational programs into more interpretable metrics. *Educational Researcher*, 48,4, 217-228.

Bloom, H.S., Hill, C.J., Black, A.R. and 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 pp 289-328.

EEF (2018). Sutton Trust-EEF teaching and learning toolkit and EEF early years toolkit, working document v.01. London: Education Endowment Foundation.

Gorard, S. (2018) *Education Policy*, Bristol: Policy Press

Hattie, J. (2008). *Visible Learning: A synthesis of over 800 metanalyses relating to achievement*. Oxford: Routledge.

Hattie, J. (2017). *Visiblelearningplus.com.* https://visible-learning.org/backup-hattie-ranking-256-effects-2017/

Higgins, S. (2016). Meta-synthesis and comparative meta-analysis of education research findings: some risks and benefits. *Review of Education* 4, 1, 31–53. http://dx.doi.org/10.1002/rev3.3067

Kvernbekk, Tone. 2015. *Evidence-Based Practice in Education: Functions of Evidence and Causal Presuppositions*. Routledge.

Motteram, G., Choudry, S., Kalambouka, A., Hutcheson, G. and Barton, A. (2016). *ReflectED Evaluation report and executive summary.* London: EEF.

Patel, R. Jabin, N., Bussard, L., Cartagena, J., Haywood, S. and Lumpkin, M. (2017). *Switch-on effectiveness trial: Evaluation report and executive summary.* London: EEF.

See, B.H. (2017). Evaluating the evidence in evidence-based policy and practice: Examples from systematic reviews of literature. *Research in Education*, 102, 1, 37-61.

See, B.H., Gorard, S. and Siddiqui, N. (2016). Teachers' use of research evidence in practice: A pilot study of feedback to enhance learning. *Educational Research*, 58, 1, 56-72.

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.

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

Slavin, R. (2019, May 30). Send us your evaluations [Web log post]. Retrieved from https://robertslavinsblog.wordpress.com/2019/05/30/send-us-your-evaluations/.

Stanovich, P.J. and Stanovich, K.E. (2003). *Using research and reason in education: How teachers can use scientifically-based research to make curricular and instructional decisions*. Jessup, Maryland: National Institute for Literacy.

Torgerson, D., Torgerson, C., Ainsworth, H., Buckley. H., Heaps, C., Hewitt, C. and Mitchell, N. (2014b). *Improving writing quality: Evaluation report and executive summary*. London: Education Endowment Foundation.

Torgerson, C., Ainsworth, H., Bell, K., Elliott, L., Gascoine, L., Hewitt, C., Kasim., Kokotsaki, D., and Torgerson, D. (2018). *Calderdale Excellence Partnership IPEEL: Evaluation report and executive summary.* London: Education Endowment Foundation.