Evaluating the evidence in evidence-based policy and practice: Examples from systematic reviews of literature

*Beng Huat See
School of Education
Durham University

*Corresponding author's email: b.h.see@durham.ac.uk

Abstract
With the push for evidence-informed policy and practice, schools and policy makers are now increasingly encouraged and supported to use and enagage with research evidence. This means that consumers of research will now need to be discerning in judging the quality of research evidence that will inform their decisions. This paper evaluates the quality of evidence behind some well-known education programmes using examples from previous reviews of over 5,000 studies on a range of topics. It shows that much of the evidence is weak, and fundamental flaws in research are not uncommon. This is a serious problem if teaching practices and important policy decisions are made based on such flawed evidence. Lives may be damaged and opportunities missed. The aim of this paper is to show how widespread this problem is and to suggest ways by which the quality of education research may be improved. For example, funders of research and research bodies need to insist on quality research and fund only those that meet the minimum quality criteria. Journal editors and reviewers need to be cognizant of fundamental flaws in research and reject such submissions. One way to do this is to encourage submission of the research design and research protocol prior to acceptance, so acceptance or rejection is based on the design and not on the outcomes. This helps prevent publication bias and biased reporting. Individual researchers can improve quality by making it their moral responsibility to be truthful and transparent.

Keywords: Evidence-based policy and practice, assessing trustworthiness of research evidence, systematic reviews

INTRODUCTION

This paper addresses the question of the validity of evidence that informs policy and practice and argues that more needs to be done to ensure better quality control of published research.

When the new Labour government in UK came into office in 1997 there was an increasing push for evidence-informed policy and practice to ensure that policies are relevant and effective (Cabinet Office 1999). A wide range of ambitious initiatives were launched, and a number of strategies suggested to encourage the use of evidence in public policy and practice (Davies et al. 2000; Nutley et al. 2002; Bullock et al. 2001). But this push for evidence-based education (EBE) is not just the concern of the labour party. In 2013, Michael Gove, the then Tory Education Secretary, asked Dr Ben Goldacre (author of the famous book, Bad Science) to look into how research evidence could be used to inform teaching and learning in schools (Goldare 2013). In the US they went further. They pass a law known as the Every Student Succeeds

Act (ESSA), which encourages and, in some states, requires schools to adopt 'proven' programmes. Federal funding requires local education agencies to include evidence-based interventions in their school improvement plans. The ESSA categorises three levels of evidence: strong evidence, that is supported by at least one randomized controlled trial; promising evidence with at least one-correlational with pretests as covariates and programmes with high quality research based on strong theory (Slavin 2016). As Slavin pointed out in his blog, there are potential problems in relying on evidence based on correlational studies since these studies cannot prove causation. Therefore programmes that have only ever been tested using correlational studies should not be classified as 'proven' programmes. In this paper you will find plenty of examples of such studies and you will see why relying on such evidence is not useful.

What all this means is that there is now a need for consumers of research (e.g. teachers, school leaders and policy makers) to be educated in understanding and judging the quality of research so they can make informed decisions about what is good evidence and what is not. A common conclusion made in most systematic reviews in social science is the poor quality research. Weak methodology and biased reporting, for example, are widespread (Gorard et al. 2007). But not everyone who uses research evidence is trained to spot these weaknesses or understand the research enough to apply it appropriately. Currently many policies and teaching approaches have been implemented based on incomplete or flawed evidence, and in some cases based on misinterpretation of the research findings. For example, there are a range of computer software for teaching literacy and numeracy being used in schools in England but which have not been tested in robust evaluations. Evidence of their effectiveness is often based on the developer-sponsored research (see Khan and Gorard 2012; Gorard et al. 2016). But users are not necessarily aware of this nor they do they have the time or knowledge to check the validity of such research evidence.

With the UK government's push for teachers to engage in and with research, and the need to make research accessible to practitioners, published research evidence becomes even more significant as schools are turning to the strategies described in such research for guidance. However, there is currently very little guidance on how to judge the quality of evidence in education research. In medicine and in health care, there are a number of protocols developed to do precisely that (Rychetnik et al. 2002). Examples include the Cochrane Reviews (http://www.cochrane.org/what-is-cochrane-evidence), the York University Centre for Reviews and Dissemination (https://www.york.ac.uk/crd/research/) and the Kauffman Best Practices Project approach (2004). In 2011, on the recommendation of Michael Gove, the Education Endowment Foundation (EEF) was established to test the evidence base on what works in raising the attainment of children in England using large scale randomised controlled trials RCTs).

While evidence from these RCTs is being generated, as a guide, the EEF developed the Teaching and Learning Toolkit (previously known as the Pupil Premium Toolkit) (https://educationendowmentfoundation.org.uk/resources/teaching-learning-toolkit). The Toolkit (Higgins et al. 2012) summarises the evidence of widely used interventions, including an estimation of the cost of each of the approach, the strength of evidence (padlock rating) and the impact in terms of months progress). The aim of the Toolkit was to help schools and teachers make informed decisions on which of these programmes to adopt.

This is a a big step forward in evidence-based education. However, it has to be mentioned that the evidence in the Teaching and Learning Toolkit is based on meta-analyses of meta-analyses and systematic reviews – often lumping together the effects of different studies of varying qualities (Simpson 2017). As the authors of the Toolkit cautioned, the evidence in the Toolkit is based on the average from different research studies and may not work in all contexts and should be used with professional advice. Therefore the Toolkit can only be considered a menu, and the security rating does not actually reflect the strength of evidence for each approach. The authors also explained that the Toolkit is a 'live resource' (p.6) and as the EEF carries out more robust evaluations of these approaches, new evidence comes in and the security ratings will be revised. So for now, this can only be a guide and is therefore not prescriptive.

In the meantime while the EEF and the other research clearinghouses evaluate the research evidence in education, many more weak studies are being funded and produced and whose evidence is being used in policy and practice. This paper looks at the evidence behind some well-known education programmes using examples from previous reviews of over 5,000 studies on a range of topics to illustrate that much of the evidence on which policies and practices is based is weak. And methodological flaws are common. Studies reporting big effects do not necessarily mean that the approach is effective and vice versa. This is a serious problem if teaching practices and important policy decisions are made based on such flawed evidence. Lives may be damaged and opportunities missed. The aim of this paper is to show how widespread this problem is and to suggest ways by which the quality of education research may be improved.

For reason of space, we have selected the more influential and interesting programmes. These include those on the EEF Teaching and Learning Toolkit: the Summer School programme, Enhanced Feedback and Assessment for Learning (AfL), parental involvement interventions and arts education interventions. These are programmes that have been introduced in schools in England to raise the attainment of children from disadvantaged backgrounds. For each of these programmes or initiatives, we first present the policy background and then the evidence base for these policies.

To judge the validity of the evidence we apply a set of simple criteria (see Chapter 4 Gorard et al. 2017). These include a consideration of the:

- Sampling strategy (how participants were selected, i.e. volunteer vs non-volunteers)
- Threats to validity (e.g. sample size, are the comparison groups similar)
- How outcomes are measured (e.g. researcher-developed tests or intervention-related instruments)
- Attrition (is the attrition large enough to alter the results of the findings)
- Conflict of interest (research conducted by the developers)
- Biased reporting
- Whether the conclusions are warranted by the findings

Summer School Programme

In September 2011, the coalition government announced that £50 million would be made available for a summer schools programme in England. The scheme was intended to support disadvantaged pupils in the transition phase from primary to secondary school. In 2012 over 1,700 schools were involved in summer school programmes, and in the summer of 2013 the number increased to 1,900, all sponsored by the Department for Education. An evaluation of the impact of summer school involving around 21,000 pupils in UK was conducted comparing the outcomes of these pupils with pupils in comparator schools that did not participate (Martine et al. 2013). The 'evaluation' did not look at the impact on attainment, and was solely concerned with survey responses to items about pupils' confidence, social skill and readiness to attend their secondary school. Positive effects of participation in summer school were reported. Comparing participating schools with non-participating schools can lead to misleading conclusions since those involved in summer schools may be self-selected and are therefore different (perhaps more motivated and progressive) than those not involved. Moreover, the outcomes were based on pupils' self-report, which is often not very reliable as pupils more often than not report success even if the impact evaluation shows no effect (Gorard et al. 2017)

What is the evidence?

There has been a lot of research on the effect of summer schools on children's academic outcomes, most of which claimed positive effects. A systematic review of 93 studies of summer schools suggested that they are more effective with maths than literacy (Cooper et al. 2000). Many of these studies were non-experimental comparing outcomes of participants before and after summer school with no comparison group. This makes it difficult to conclude if children would have done equally well if they had not participated in the summer school. Other studies using a comparison group also reported positive impacts on test scores (e.g. Sunmonu et al. 2002; Schacter and Jon 2005).

All these sound very promising, but these studies compared participants with non-participants. Since participation in summer school is voluntary, those who volunteered are likely to be different from those who did not. Another study using a form of regression discontinuity analysis estimated the impact of summer school on both maths and reading as about +0.12 for 3rd graders but not for the 6th graders (Jacob and Legfren 2002). While this is a stronger evidence participants are not randomly assigned to conditions. There may be differences between participants and non-participants that are unobserved.

Other studies purported to provide stronger evidence using random assignment (the gold standard of evaluations) of participants also suggested positive effects of summer school on academic outcomes. However, these studies also suffer from issues of validity due to high attrition. Since attendance in summer school is voluntary the level of dropout is often high. Borman and Dowling (2006), for example compared 438 students randomised to summer school with 248 receiving no intervention. The study reports success for the intervention, but because of the high dropout they compared the post-test scores of only those who continued to attend with all of the control students. This is likely to inflate the perceived impact of the summer school since those who turn up may be very different from those who do not.

One well-known summer school programme that has been evaluated several times using randomised controlled trials is the BELL (Building Educated Leaders for Life) summer school. There have been several

evaluations of the BELL summer schools in the US (BELL 2001, 2002, 2003). Results suggest that there was no benefit for writing (Harvard Family Research Project 2006), but a positive effect was reported and no effect size was provided. Most of these evaluations were by the developers themselves who were more concerned about finding an impact than in what that impact is.

A famous study by Chaplin and Capizzano (2006) of the BELL summer school using robust RCT and standardised tests looked at BELL summer schools in Boston and NY City (Chaplin and Capizzano 2006). In this study participants were randomised to summer school or not. The overall 'effect' size for reading calculated here (not by the original authors) was +0.02 which is negligible. As in Borman and Dowling's study this study also suffered from high attrition after randomisation. 46% of those randomised dropped out or refused to continue with the study, and the results are available for only 44% of the initial randomised students. Chaplin and Capizzano claimed substantial positive impacts.

In fact, the data in the paper suggests that the programme had no impact on all measures (reading and academic self-concept). There was a negative impact on vocabulary score of being on the programme compared to the control group. The only statistically significant impact was for increasing parental reading to children, but even then effect sizes were small. Chaplin and Cappizano then claimed that because the post-test for the control group was taken 16 days later, they had a 16-day advantage and adjusted the results to account for this difference. After adjustments the impact of the programme on academic outcomes was still negligible. Effect size for total score after adjustment was 0.08, effect size for vocabulary and comprehension was 0.04. Although the authors acknowledged that the children were from a wide range of age groups (grades 1 to 6), the analysis did not take into account the impact for the different age groups of children, but speculated that the effect might have been bigger if segregated analysis was carried out. Why this was not done as a routine analysis was not explained. Despite the small effect sizes, they claimed that "strong impacts were found on reading test scores" (p .37). They further added that the "impact of the program on test scores appears similar to that of a similar amount of school and is precise enough that it is unlikely that it was caused by chance. Despite finding no impact on most measures and negative effect on some and small effects on academic outcomes, Chaplin and Cappizano concluded with the statement: "*Our results suggest that the BELL program has positive and substantively important impacts*" (p.38).

This is an example where the conclusions are not warranted by the findings and reports of the findings are misleading. If improving attainment is the aim then the results of this study alone is not enough to justify introducing it across all schools.

To test the effect of the BELL summer school programme on pupils' academic outcomes in the UK, the EEF funded a pilot study in 2012 (Gorard et al. 2015). The EEF evaluation, involving 435 children aged 9 to 11 found positive effects on standardised test in English (effect size = +0.17) with stronger effects for free school meal children, but no effects on maths (Gorard et al. 2015). This was an efficacy trial, so a bigger trial will be needed to confirm these effects. Overall therefore, despite some claims to the contrary, there is no strong evidence that the BELL approach would work in England with disadvantaged pupils preparing for secondary school.

Effective Feedback

Almost all studies on effective pedegogy identify feedback as a characteristic of effective teaching and learning (Harris & Ratcliffe. 2005; Creemers 1994; Scheerens 1992 & 1999; Siraj-Blatchford & Taggart 2014; Coe et al. 2014; DfE 2000). In John Hattie's metananalysis feedback strategy was cited as having an effect size of + 1.13 (http://www.teacherstoolbox.co.uk/T_effect_sizes.html_Hattie). This is a huge effect for education, equivalent to a jump from a Grade C to a Grade A at GCSE. The Education Endowment Foundation Teaching and Learning Toolkit recorded feedback strategy as having an impact equivalent to +9months progress and Assessment for Learning with an impact of +3 months progress (Education Endowment Foundation 2015).

The evidence for the feedback strategy appears to be compelling. In 1988, the National Curriculum Task Group on Assessment and Testing (TGAT) for England and Wales recommended that assessment should be an "integral part of the education process, continually providing both 'feedback' and 'feedforward' and ought therefore to be 'systematically incorporated into teaching strategies and practices at all levels" (DES 1988, paragraph 4). And enhanced feedback strategy was one of the innovations recommended and implemented in schools across England under the National Strategy. These policies have been introduced with no robust evidence yet about the effectiveness of feedback or how it works. Much of the evidence so far has been based on correlational studies or passive non-experimental designs. It has not been tested or trialled on a large scale.

The most influential evidence on feedback strategy is the work by John Hattie in his article on the power of feedback (Hattie and Timperley 2007). He based his evidence on 74 meta-analyses of 4,175 studies involving 5,755 effect sizes. Hattie and Timperley reported wide variations in effect sizes depending on the types of feedback used, but there was no comment on the quality of these studies and the reliability of the evidence. The study that most informed their model was the one by Kluger and DeNisi (1996) because this study, according to Hattie and Timperley was 'the most systematic', and 'included studies that had at least a control group, measured performance, and included at least 10 participants'. This suggests that most of the other studies in the 73 other meta-analyses were not systematic, had no control group, had fewer than 10 participants and did not even measure performance. How many of such studies were in these meta-analyses was not known. Therefore, the number of studies whose evidence we can rely on is unknown. Hattie and Timperley added that many of these studies were not classroom based. Presumably these studies were undertaken in a laboratory condition. How would the results compare in real life classroom situations which normally has about 30 students and other atmospheric distractions, such as noise from outside classroom, interruptions from other students. Classes with SEN children can also affect classroom delivery because of the attention needed. These were not commented on, but are no less important because of threats to external validity. They then went on to say that their evidence based on 131 studies (conducted largely in control condition) included 470 effect sizes and involving 12,652 participants showed an average effect size of 0.38 (SE=0.09). Of these 32% showed negative effects or decreased performance. This cannot be explained by scale (sampling error) or even by theories of feedback use. But an explanation might involve the kinds of feedback such as praise, reward and punishment, the nature of the pupils or students involved, or the quality of implementation which is often

not very helpful. This is an example of unclear reporting, which could unintentionally confound readers. And assuming equal sample size in each of the 131 studies, there is an average of fewer than 100 participants in each study. Hattie and Timperley reported that the 'average sample size per effect was 39 participants'. This is a considerably small sample. Small samples pose a threat to external validity and reliability.

Also Hattie's study is based on a meta-analysis of meta-analyses. The meta-analyses themselves span across 20 years, some research reports cited went back as far as 1960. These meta-analyses used different calculations of effect sizes, for different measures of the same parameters (e.g. different types of reinforcement and a range of feedbacks) for different groups of children of different phases of schooling. Some studies were specifically for SEN children, children with behavioural, emotional and disruptive behaviour. How Hattie arrived at the effect sizes that he did in his paper was not explained.

Going back to the studies that Hattie cited in his paper, we could not locate the effect sizes listed in the summary table (Table 1, p. 83). For example, the review states that the effect size of 54 studies in Lysakowski and Walberg (1982) was +1.13 whereas the original paper reports it at 0.97 (study-weighted). The figure 1.13 appears nowhere in their paper. In Hattie (1992) and repeated subsequently, it is said that "Skiba, Casey & 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 the later 2007 paper reports this review as having 35 effect sizes not studies, and an effect size of +1.24. While none of this undoes the work that has been done or eliminates the evidence for the impact of enhanced feedback it ought to lead to caution. Overall, the evidence is not as clear as some commentators have suggested. We are also not sure what the outcome measures were for these studies.

Over half of these studies have no randomisation and yet quote p values. The use of p values in significant testing is based on the strict assumption that there is complete randomisation with no dropouts. And even if this condition was met significant tests would still not be appropriate because what significant test tells us is the probability of observing the results we get assuming that there is no difference between the groups. But the answer that we really want is whether there is a difference between the groups given the results that we have. Unfortunately significant test does not give us the answer to the latter question. This is a misinterpretation of what significant test does. This means that significant tests are invalid in a large number of cases where they are used (Colquoun, D. 2014; 2016; Gorard 2016). So if Hattie used meta-analyses based on p values but did not eliminate the majority where it was used incorrectly then the results will be completely wrong. For example, one study used MANOVA based on those who agree to participate and those who refused, presented as evidence of the intervention. This is not uncommon.

There have also been other studies conducted to evaluate the impact of a feedback strategy known as Assessment for Learning (AfL). AfL is a formative feedback strategy popularised by Paul Black and Dylan Wiliam (Black et al. 2003). It is simply a strategy where teachers use information about the learners to inform teaching and learning. Black and Wiliam (1998) conducted a review of 580 articles from 160 journals. Of these they selected 20 and included another 23 from an earlier study by Fuchs and Fuchs (1986). All the studies indicated a substantial impact of formative assessment on the learning of pupils

ranging from age 5 to undergraduates and across a range of subjects and nationalities. The authors found that the average effect sizes of the impact of formative assessment (FA) on pupils' attainment ranged between 0.4 and 0.7. It has to be mentioned that Fuchs and Fuchs' study was based on children with mild handicaps (their term). Black and Wiliam also found that many of the studies showed that FA was more effective for low achievers than other students. As in Hattie and Timperley's paper, the effect sizes were averaged across a wide age range (from pre-school to undergraduates). It may be the case that the approach is more effective for older children but less effective for younger ones. These differences were not teased out in the synthesis.

A newer review by Hopfenbeck and Stobart (2015) found 1,387 reports of research on AfL. Most of those concerned with impact on pupil attainment were small case studies (with perhaps one or two schools), and few were well-designed evaluations. This raises the question of the kind of studies that were synthesised by Hattie and Timperley (2007) and Black and Wiliam (1998).

Given the problems with all the studies so far, the evidence for feedback remains unclear. And yet policies are made based on such evidence, and schools are using it. Caution must therefore be taken when introducing such strategies en masse. Even if feedback was found to be effective, translating the research findings into classroom practice is not easy. Black and Wiliam acknowledged the complexities involved in implementing them in real classroom conditions (Black and Wiliam 2001, p. 2). Others have also suggested that this approach will not always be effective, perhaps especially if rolled out without due care (Smith and Gorard 2005).

The EEF evaluation of a feedback strategy based on Hattie & Timperley's model of enhanced feedback (See et al. 2016) found that getting teachers to apply the research findings in the classroom is complex and challenging. Teachers were not able to engage with academic paper. This is because academic papers are not written for practitioners. They do not give details about the intervention or how it is to be implemented in the classroom. The meta-synthesis used as evidence in John Hattie's as well and Black & Wiliam's work does not tell us what happens in the classroom. No account was taken of factors like teacher motivation, experience and competence of teachers. There was also no mention of the kind of tests used (e.g. were they standardised tests, intervention-specific, teacher-assessed, teacher ratings or pupil self-reports). If the tests were assessed by teachers then they cannot be blinded since they had knowledge of which students had been exposed to the intervention. This could influence their judgement.

For practitioners to use such evidence therefore requires more than simply reading the papers. There needs to be a clear and unbiased link connecting primary evidence to proposed classroom practice (Nelson and O'Beirne 2014). An impact evaluation of National Strategy support for Assessment for Learning (Ofsted 2008) in 27 primary and 16 secondary schools found no impact on English and maths (the two subjects evaluated) in over two-thirds of the schools visited. Impact was also variable across subjects. The evaluation also suggested that ineffective schools or ineffective lessons had been where there was a lack of understanding by teachers on what the approach entailed. The report stated:

> Where assessment for learning had had less impact, the teachers had not understood how the approaches were supposed to improve pupils' achievement. In particular, they used key aspects of assessment for learning, such as identifying and explaining objectives, questioning, reviewing pupils' progress and providing feedback without enough precision and skill. As a result, pupils did not understand enough about what they needed to do to improve and how they would achieve their targets. Teachers did not review learning effectively during lessons; opportunities for pupils to assess their own work or that of their peers were infrequent and not always effective.
>
> (Ofsted 2008, p.5)

In summary, there are several problems with relying on such evidence taken from meta-analyses of meta-analyses for policy and practice. First, much of it is not particularly robust (small-scale, involving non-randomisation of participants, based on summaries of effects across a wide range of subjects and age groups).

Second, no consideration was taken of the quality of research in the synthesis of existing evidence. For example, there are studies which involved only one participant, some had no comparator groups and some involved children with specific learning difficulties or had huge attrition as large as 70%. These may form the majority of studies reporting huge positive effects. On the other hand, the few good quality studies may report small effects. Averaging effect sizes from across studies of different quality giving equal weights to all can lead to misleading conclusions.

Third, there is the issue of translating research into practice.

Parental involvement interventions

In 1997 the White Paper England: Excellence in Schools set out the policy to enhance parental involvement (DfEE 1997). In 2003 the Green Paper: Every Child Matters (HMSO 2003) highlighted the role of parental involvement (PI) and suggested PI as an important contributory factor in raising attainment. And in 2009 Ofsted (a national school inspection organisation emphasised that schools should involve parents: get schools to engage with parents, improve communication with parents and develop strategies to help parents support learning at home. Some of these strategies have been evaluated and discussed in this section.

Shared reading/parent reading/dialogic reading

One popular approach in parental involvement is shared reading/parent reading/dialogic reading. It is often assumed that children, particularly those from disadvantaged backgrounds, are unprepared for school partly because the home environment is quite different to that of the school's. Some people believe that getting parents to read to their children at home prior to entering school may prepare them for school by bringing the home environment closer to that of the school's. Numerous programmes have been developed to encourage parents to read to their children at home before they start school.

Many studies have been conducted evaluating the effects of paired reading and shared reading. Results have been mixed. Of the eight studies included in the See and Gorard's (2015) review, six reported no effects on children's literacy. Baker (2011) also reported no association between intervention and outcomes. This was a correlational study so the evidence is not strong. A UK study found dialogic reading had no impact on children's language skills (Morgan 2008). A Canadian study (Sénéchal et al. 2008) also showed no relationship between children's reading ability and parental shared reading. In fact shared reading was negatively correlated with narrative ability. Parent's level of literacy had a greater effect on children's comprehension. Two US studies (Stevens 1996 and Terry 2011) also found no additional effects on measures of children's literacy (concepts of print awareness). Although Stevens reported that treatment children made significant improvements between pre- and post-test, no similar analysis was carried out for the control group. It is therefore not possible to attribute improvements to treatment. Another US study found positive effect on only one measure of vocabulary. This study was small scale (n=67) and did not randomise allocation of children.

Interestingly the study by the developer of the programme (Lonigan & Whitehurst 1998) actually found that dialogic reading had a negative effect on children's receptive vocabulary suggesting that training parents in dialogic reading may be more harmful than good. The study also suggested that to be effective, paired reading also needed to happen in the school. The biggest study involving a randomised controlled trial of 552 children in Australia found training parents in shared reading activities had a negative impact on vocabulary and home literacy (Goldfeld et al. 2011).

Despite these studies showing no effects and in some cases even negative impact, shared reading or dialogic reading is still being encouraged among parents.

Sure Start (SS)

SS is a home support programme in UK to engage mothers in scaffolding educational activities at home. A national evaluation (Melhuish et al. 2010) tracking children from age 3 to 5 and was the largest of its kind involved over 7000 families. They found no differences between Sure Start children and the Millenium Cohort children on 7 measures of cognitive and social development at age 5 teacher assessment and on 4 measures of socio-emotional development based on mothers' ratings. The data collected for children at age 3 and at age 5 were by different research teams. Attrition was high - over 50% of the original children were lost over the two years. This together with diffusion caused by introduction of pre-school education in non-Sure Start areas means that the results are no longer valid.

A smaller evaluation over a one-year period (Ford and McDougall (2009) involving 30 pre-school children reported positive effect on children's receptive vocabulary and academic knowledge. The report did not describe how the children were selected (i.e. random allocation or voluntary participation), hence children could be different at the outset. It was also unclear if comparisons were based on pre-post-test gain scores or differences in test scores at the end of the intervention.

A more recent evaluation of Sure Start Children's Centres (by Claire Crawford from Warwick University) compared children living in close proximity to SSCC with those at a distance from such centres (and

therefore less likely to use SSCC). Using intention-to-treat (i.e. those likely to be served by SSCC in high poverty areas) and difference-to-difference analyses, their preliminary findings suggest positive significant effects of access to SSCC on children's educational attainment at age 5 (measured using EYFS profile), but the effects dissipate towards the end of primary school. The data was not based on children who actually attend SSCC. The assumption is that children living nearby are more likely to attend SS schools. Data for children who were enrolled on SSCC were available and would have provided a more accurate assessment – but these were not used.

Raising Early Achievement in Literacy (REAL)

REAL is a family literacy programme which combines home instruction with parent support and training. An evaluation by *Nutbrown & Hannon (2011)* reported positive impact on literacy but it is not clear what aspect of literacy was assessed. This was a well-known study and was used as an impact case study in the REF (Research Excellence Framework) submission. Only positive results for some measures of literacy using developer-designed assessments (Sheffield Early Literacy Development Profile) were reported. Although they also assessed children using the standardised British Picture Vocabulary Scales, they did not report results on this assessment (presumably either no effect or negative effect). Results using standardised assessments at age 7 did not show any beneficial effects. Attrition was under-reported and there was inconsistency in the number of cases across tables. Also the study was small scale (only 88 children in each arm) involving families in one locale (Sheffield) in the UK who were mostly white and mono-lingual. There is thus a question of generalizability.

The author concluded that when compared to control children REAL children did not appear to have made greater progress, but on the developers' website (http://www.niace.org.uk/sites/default/files/documents/projects/Family/External_research/THE-UNIVERSITY-OF-SHEFFIELD-Impact-of-the-REAL-Project.pdf) they reported positive gains by programme children with a small to medium effect size (ES=0.4) immediately after the intervention. Results on the standardised BPVS test were not reported.

It has to be mentioned that the researchers in this case were also the developer of the programme. There is thus a conflict of interest. Yet this study was awarded the 2013 ESRC Outstanding Impact in Society prize for the wide dissemination and use of the outputs from this research.

Arts education

The House of Lords argued for arts to be part of the core curriculum to encourage the development of creativity, critical thinking, motivation and self-confidence - skills necessary for innovation (House of Lords 2014). Such skills are also believed to help children learn academically and improve children's cognition. However, the EEF-funded systematic review, (See and Kokotsaki 2016) suggests the evidence for the benefits of arts education on children's academic outcomes and cognitive development is not clear. The main reason is the lack of robust causal studies. The review looked at 200 international studies covering young people from pre-school through to age 16. Many of the studies had serious flaws in design, such as having no comparators and non-random allocation of participants. Using this review as an example, we

highlight the common flaws found in most social science research, such as lack of comparators, attrition, biased reporting and use of intervention-related tests.

For example, one famous myth is that listening to classical music (so–called Mozart effect) can improve cognitive skills of young children. Our review found that this was not the case. If policies were implemented based on such myths, it would either be a waste of money and time, or it could even be potentially harmful. We found six studies showing no or negative effects of listening to classical music. One study showed that listening to classical music had a negative effect on memory (Bressler 2003). In this study 24 five-year old children were assigned to either listening or silent condition. Treatment children listened to Mozart (K.448) being played in the background while engaged in colouring activity. Control children performed the colouring activity in silence. Children were then given a memory test (recall of visual and verbal information). Those who listened to Mozart did worse than those who worked in silence.

The second study (Albright 2012) examined the effects of listening to music on 102 children in 3rd (age 8/9) and 5th grade (age 10/11). Children were randomly assigned to either treatment or control. Treatment children listened to barogue and classical music for 50 minutes five days a week during maths lessons. After 16 weeks children took a state standardised test to measure progress in maths. The results showed that control group outperformed experimental group in the maths test.

One experiment (Hsieh 2011) comparing three groups of participants (Bach, Mozart and Silence) on the effects of music on spatial ability, found no differences between groups. This was a small study of only 90 participants. Similarly Crncec et al. (2006) also found no effects of listening to Mozart on the spatial temporal performance of 136 ten to eleven year old children who were randomly assigned to three conditions (Mozart, popular music and silence). Schellenberg and Hallam's (2005) randomised controlled trial also could not provide evidence for the beneficial effect of listening to Mozart. A matched comparison study (Lints and Gadbois (2003) showed that other conditions and not just listening to Mozart could explain enhancement in spatial reasoning.

On the other hand, there were studies reporting positive effects of listening to Mozart. Thompson (2005), for example, suggested that listening to classical music can enhance children's psycho-motor skills. The study was conducted on 91 four-year old children. Children listened to Mozart Sonata for Two Pianos in D Major for 30 days just prior to nap time, and the control group received no musical intervention. Pre- and post-test comparisons showed that experimental children did better than the control on only some measures of psychomotor skills, but not on others. This study involves a very small sample (under 100), with no random assignment of children to treatment conditions. The children were taken from one child-care centre from two classes taught by different teachers. This means that the children could be different at the outset in terms of ability, motivation or prior musical experiences. This may explain differences in performance.

One experimental study (Jausovec et al. 2006) suggested positive effects on spatial-temporal ability. This study reported two experiments both involving very small numbers (one had 14 participants in each arm of the intervention and the second experiment had only 12), and it was unclear how the participants were

allocated to the treatment conditions. A meta-analysis (Voracek 2008) revealed publication bias, lab differences and nonspecificity in the Mozart effect. All these suggest that the evidence for the so-called Mozart effect is unsubstantiated.

There is also no conclusive evidence that playing a musical instrument has a positive effect on young people's cognitive development. An experiment conducted on 10 sets of monozygotic twins (Nering 2002) aged 3 to 7 showed that the twin who received private piano instruction improve in IQ and arithmetic scores while the other twin who received no training did not show improvement. However, this study involved only 10 cases and there was no replication of this experiment to show consistent results. Also the researcher was the teacher that provided the music instruction. This may have affected the results due to experimenter expectancy effect and conflict of interest.

Three studies using brain scans showed changes in children's brains when trained to use a musical instrument. This is promising but one of the studies (Olson 2010; Schlaug et al. 2005) found that there was no difference between music and control children in terms of academic performance. Hudziak et al.'s (2014) study showed a positive correlation between music training and visuo-spatial and motor co-ordination as evidenced by the thickening of the brain cortex. However, the correlational design of the study (with no comparators) makes it difficult to say how much of the change was just due to musical training. The cortical thickening could be the result of maturation (natural development of the brain) or other practices that were not observed. Another study (Strait et al. 2013) reported that early music training could enhance neural encoding of speech-in-noise, but because the two groups of children were not randomly assigned the effects could not be solely attributed to music training.

A meta-analysis of 19 experimental studies (Hetland 2000) also showed that learning a musical instrument could improve students' spatial-temporal reasoning, but there was no evidence that this led to improvement in academic achievement.

The evidence for arts education is also weakened due to flaws in many of the studies. These include attrition, having non-equivalent comparison groups, using intervention-related or researcher-developed instruments, biased reporting and conflict of interest.

Biased reporting is very common. As the examples below illustrate, often the conclusions do not follow from the findings. For example, there were also studies that ignored negative results and reported only the positive ones. James (2011) concluded that infusing arts was effective in enhancing maths engagement and achievement because of the big gains made by experimental children between pre- and post-test. Although this was true, the results showed that control children made even bigger gains but this was not mentioned in the discussion. The author then went on to cite positive effects on engagement using teacher reports. For example: "'*I think the kids really enjoyed the lessons*' and '*One of students who rarely turned in homework started to turn in his maths homework*'.

Bettencourt (2009), despite finding no effects, concluded that the overall data showed that the intervention had positive benefits on students and educators, and that self-reflected learning was

beneficial to students. The researcher then made recommendations for introducing writing activities into maths lessons stating that traditional methods of teaching maths had been shown by previous research to be ineffective.

Similarly, Ayers (1993) reported no significant differences between groups, but concluded that pupils in the treatment group showed greater retention of information. This was not substantiated by the data presented. In fact the data suggest that control children made bigger gains than those in the experimental group. In terms of retention, the data showed that both groups registered a loss in retention of information, but analysis of data 6 weeks after the intervention showed that control group retained more information than experimental group. Even after controlling for reading and maths pre-test, no significant results were found.

Harland et al.'s (2000) study showed no relationship between arts participation and performance at national exam when prior attainment and social background were accounted for. Yet they claimed that arts participation boosted general academic performance and also resulted in greater personal development. It has to be noted that positive effects on creativity, critical thinking, self-confidence and other personal and social development skills were assessed using interviews.

There were studies that compared arts-focused schools with non-arts schools (e.g. Hetland and Winner; Vaugh and Winner 2000; Clark 2007; Yorke-Vinney 2007; Thomas and Arnold 2011). In the US, high performing pupils are encouraged to take up arts classes. High performing schools are also encouraged to provide for arts classes and so are able to retain their arts programmes which the lower performing schools are unable to do. So comparing outcomes of high arts performing schools with low performing non-arts schools is not a fair comparison. Arts-focused schools may also appeal to different types of pupils. It is also possible that integrating arts in the curriculum makes learning fun and enjoyable for pupils and teachers.

Peppler et al. (2014) conducted a longitudinal study looking at the impact of integrating arts in the curriculum. They reported that treatment pupils were more likely to pass the standardised test of English than control pupils. This was a large study consisting of over 6,000 children from six schools. They compared schools using the integrated arts curriculum with those using standalone arts programme. Allocation to treatment was not randomised, suggesting that there could be differences between schools. Perhaps schools that were on the integrated arts programme were progressive schools which may have other enrichment or interventions going on. These were not recorded. More importantly, comparisons were made with different cohorts of pupils over the three years. For example, grade 2 pupils in the first year were compared with grade 2 pupils in the second year. As pupil intakes varied from year to year, such comparisons do not show gains in pupils' performance but variation in pupil intake.

There were also studies that compared children who volunteered participation with those who did not. For example, Danner (2003) compared affective outcomes of 47 seventh grade children exposed to a creative drama enrichment programme with seven other children (representing 28% of all who were invited) in a private performing arts school not exposed to drama instruction. Although the study

described the design as a matched comparison, the two schools were clearly different: one was a charter school and the other a selective private performing arts school. A range of outcomes were assessed before and after the intervention. These included intention to use drugs, intention to engage in sexual activity and self esteem. There were no changes observed with the comparison group, but treatment pupils showed declining behaviour in several areas. The author concluded that the programme may even be harmful to the participants. This was a small study with only 54 students with attrition at 15%. Groups were not equal, and experimental students were volunteers. It may be that treatment pupils were more open about their behaviour after the treatment, hence the negative results. So the negative results do not actually mean that the intervention had failed.

Most of the studies that reported huge impact of arts on children's learning outcomes were based on teacher/researcher-developed tests or teacher surveys. A number of studies also reported significant improvements based on teachers' perceptions of impact, but no effects when tested using standardized tests. Brugar (2012), for example, examined the use of visual arts on 10 year old children's history learning. The results showed that both experimental children made bigger improvements compared to the control children (ES = +1.6). In this study the test questions were aligned with the activities used with the intervention group. The two questions for which experimental group did not show improvements were those that were not associated with the classroom teaching. The results are therefore not valid as the test disadvantaged the control children who were not taught the content.

Joseph (2014) conducted an experiment with 83 grade 4 (9-10 year old) children to test the effects of integrating creative dramatics on vocabulary. Vocabulary achievement was assessed using a teacher-researcher developed criterion-referenced test on vocabulary covered in the curriculum. This invalidates the test as the researcher may inadvertently pick vocabulary words that were taught in the creative dramatics lessons.

Another common problem in many of the studies is attrition. This is rarely considered when assessing the security of evidence. In our review of arts education, several studies had over 30% attrition. For example, Rose et al.'s study (2000) had attrition of up to 38% for some measures. Smith's (2011) study of the Georgia Wolftrap program (a drama-based programme) to develop children's social cognitive skills had 42% of missing data from the intervention group and 33% from the control. A study that explored the effects of creative dance on young and old people (von Rossberg-Gempton 1998) lost 50% of its participants. Smithrim and Upitis's (2005) longitudinal study involving 6,000 students had an attrition of 32%. Another study (Rossini 2005) started with 88 children, but only 65 parents gave consent and only 47 completed both pre and post tests (38% attrition). Costa-Giomi (1999) started with 119 children and ended up with only 78 representing an attrition of over 30%. Any missing values can create bias (Dong and Lipsey 2011). And where such attrition is not random (as is most often the case) it can bias the estimate of the treatment effect, and the bias can still be large even when advanced statistical methods like multiple imputations are used (Foster & Fang 2004; Puma et al. 2009). Such bias can distort the results of statistical significant tests and threaten the validity of any conclusion reached (Shadish, Cook & Campbell 2001; Campbell & Stanley 1963; Little & Rubin 1987).

There were also studies that did not report attrition giving only vague description of the sample. Walker (2008), for example, reportedly recruited 56 pupils from a total of 188, and reported that control group included only those who took the pre- and post-test, suggesting that those who did not take the test were dropped from the analysis, so it is not possible to estimate the attrition rate. Albright (2012) gave no data on attrition, while Runfola et al. (2012) reported that some teachers dropped out from the control group due to lack of motivation for being in the control. This means that those who stayed on are likely to be different to those who remained.

Studies conducted or commissioned by developers themselves also tended to report big effects because they have a vested interest in the programme. They are often more concerned that there is an impact rather than what the impact is. These are things that classroom teachers or policy makers may not be aware of when reading a research piece, and thus cannot judge the validity of the research. In one example, Fountain (2007) developed and evaluated a differentiated art instruction on students in her own school. She reported positive impact on creativity despite having no comparison group. Robinson's (2013) review of different arts integration reported that arts integration (apart from music integration) showed potential of positive effects on academic and other cognitive outcomes as well as social skills. However, it is worth noting that Robinson also writes and develops programme in arts integration.

Some of the examples described in this paper are not among the worst. In fact they are some of the better ones. Because of their clear reporting we are able to assess their quality. There are studies which do not even report on the sample or sampling strategy making it even more difficult to judge their quality.

What can be done?

The evidence presented above for some of the popular approaches in education shows that there is much to be done to ensure that the evidence that informs policy and practice is reliable and secure. There are a number of ways by which this may be achieved.

First, funders of research like the Education Endowment Foundation, the Economic and Social Research Council , Institute for Education Sciences What Works Clearing House, Nuffield Foundation and the British Academy, just to name a few, should insist on quality research using the simple criteria checklist: large sample, random allocation, minimal attrition, independent evaluation, independent standardised assessments to measure outcomes.

Second, research bodies like the British Educational Research Association can help raise awareness of bad practices in research, such as bias reporting, using causal terms like 'impact' in correlational studies and unclear reporting of attrition and sampling strategy. This can be done through conferences, workshops and on their websites. There needs to be more discussions on such issues.

Third, journal editors and reviewers should not accept and publish papers that do not report clearly their methods allowing readers to make their own judgements.

Fourth, introduce a system where people write an outline about their research: aims and objectives, policy and practical relevance, research design and journal editors accept or reject the research solely based on the soundness of the research rather than on the outcomes. This can help prevent dredging for data, biased reporting or publishing only positive results.

Fifth, it is the responsibility of individual researcher to do the best research with the resources they have. If the resources do not permit a sample beyond the classroom, then the claims made about the effectiveness of the intervention can only be limited. For example, pilot studies and studies by research students are likely to be small-scale. This is acceptable but the claims made about the impact would be restricted to only the sample. As far as possible research should involve as large a sample as possible so that the results can be generalised to a larger population. It is not good enough to test a programme on a class of children and claim that it will work with other children.

Finally, all research ethics should also include the moral responsibility of the researcher to be truthful and transparent to ensure that the research is of the best quality given the available resources. Researchers should be reminded that their research has implications on people's lives as schools, teachers, parents and policy makers make decisions based on their research. It is unethical to use tax-payers' money irresponsibly. Books on ethics are keen to remind us of the need to protect individuals, their identity and the on this for data protection and data security. These are very important issues indeed. But often what is forgotten is the need to protect those individuals who are using the outcome of the research.

Conclusion

There is a lot of research going on in education, perhaps too much. And the results are often not consistent. This makes it difficult for non-academic users of research to know which evidence to rely on. There are two issues here. One is the requirement of consumers of research to have the skills to read and judge the quality or trustworthiness of research. Another is the question of time. With so many studies being conducted on any one topic, how do policy makers and practitioners have the time to sieve through them all? Perhaps what is needed is a an avenue for research to be translated into easily digested summaries with tiered evidence, similar to the padlock ratings used by UK's Education Endowment Foundation (EEF) (https://v1.educationendowmentfoundation.org.uk/uploads/pdf/Classifying_the_security_of_EEF_findings_FINAL.pdf).

In a way, the EEF has initiated this step, but it could do more as its current Teaching & Learning toolkit is based on meta-analysis which takes no account of the quality of individual studies. One suggestion would be to re-analyse the evidence base used to inform the Teaching and Learning Toolkit using their own padlock ratings to judge the quality of individual studies. This could go a long way to making research evidence clear and accessible to policy makers and other non-academic consumers of research.

References

Albright, R.E. (2012) *The Impact of Music on Student Achievement in the Third and Fifth Grade Math Curriculum.* 3492175 Ed.D.thesis. Prescott Valley, AZ: Northcentral University.

Ayers, W.E. (1993) *A study of the effectiveness of expressive writing as a learning enhancement in middle school science.* Unpublished dissertation. Philadelphia, PA:Temple University.

Baker, C.N. (2011) *Relationships between contextual characteristics, parent implementation, and child outcome within an academic preventive intervention for preschoolers*, unpublished PhD thesis, University of Massachusetts, Amherst.

Bettencourt, C.L. (2009). *Promoting social change through writing: A quantitative study of research-based best practices in eighth-grade mathematics*. EdD thesis. Minneapolis: Walden University.

Black, P., Harrison, C., Lee, C., Marshall, B, and Wiliam, D. (2003) Assessment for Learning – putting it into practice. Maidenhead, UK: Open University Press.

Black, P. and Wiliam, D. (1998) *Inside the black box: raising standards through classroom assessment*, London: GL Assessment

Bressler, R.A. (2003) *Music and cognitive abilities: A look at the Mozart Effect.* Unpublished PhD thesis. Chicago, US: The Chicago School of Professional Psychology.

Brugar, K. A. (2012) *What difference does curricular integration make? An inquiry of fifth graders' learning of history through the use of literacy and visual arts skills*. Unpublished PhD thesis. Michigan: Michigan State University.

Bullock, H., Mountford, J. and Stanley, R. (2001) *Better policy-making London: Cabinet Office, Centre for Management and Policy Studies*, 83pp http://www.cmps.gov.uk/better_policy_making.pdf

Cabinet Office (1999) *Professional policy making for the twenty first century*. Report by Strategic Policy CentreMaking Team. London: Cabinet Office.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171–246). Chicago, IL: Rand McNally.

Chaplin, C. and Capizzano, J. (2006) *Impacts of a Summer Learning Program: A Random Assignment Study of Building Educated Leaders for Life (BELL).* Washington DC: The Urban Institute. http://www.urban.org/UploadedPDF/411350_bell_impacts.pdf

Clark, S.J. (2007) *The relationship between fine arts participation and the emotional intelligence of fifth-grade elementary students*. PhD thesis. Athens, GA: University of Georgia.

Coe, R., Aloisi, C., Higgins, S. and Major, L.E. (2014) *What makes great teaching? Review of the underpinning research*. London: Sutton Trust

Colquoun, D. (2014) An investigation of the false discovery rate and the misinterpretation of p-values, Royal Society Open Science, http://rsos.royalsocietypublishing.org/content/1/3/140216

Colquoun, D. (2016) The problem with p-values, Aeon, https://aeon.co/essays/it-s-time-for-science-to-abandon-the-term-statistically-significant

Costa-Giomi, E. (1999) The effects of three years of piano instruction on children's cognitive development. *Journal of Research in Music Education, 47*(3): 198-212.

Creemers, B. (1994) *The effective classroom*. London: Cassell.

Crncec, R; Wilson, SJ; Prior, M. (2006) No evidence for the Mozart effect in children. *Music Perception,* 23(4): 305-317.

Danner, S.A. (2003) *Prevention through performance: A creative drama enrichment program for arts school students.* Unpublished PhD thesis. Bowling Green, Ohio: Bowling Green State

Davies, H.T.O., Nutley, S.M. and Smith, P.C. (eds) (2000) *What works? Evidence-based policy and practice in public services.* Bristol: Policy Press, 380pp

DES (1988) *National curriculum task group on assessment and testing. A report to the Department of Education and Science and the Welsh Office.* London: DES. http://www.educationengland.org.uk/documents/pdfs/1988-TGAT-report.pdf

DfEE (1997) *White paper: Excellence in schools*. London: Her Majesty's Stationery Office.

Dong, N. and Lipsey, M. (2011) *Biases in estimating treatment effects due to attrition in randomised controlled trials: A Simulation study*. SREE Conference, 2011.

Ford, R. M. and McDougall, S.J.P (2009) Parent-delivered compensatory education for children at risk of educational failure: Improving the academic and self-regulatory skills of a Sure Start preschool sample. *British Journal of Psychology,* 100(4): 773-797.

Foster, M. E. and Fang, G. Y. (2004). Alternatives to Handling Attrition: An Illustration Using Data from the Fast Track Evaluation. *Evaluation Review, 28*:434-464.

Fountain, H.L.R. (2007) *Using art to differentiate instruction: An analysis of its effect on creativity and the learning environment*. Unpublished PhD thesis. West Lafeyette: Purdue University.

Fuchs, L. S., and D. Fuchs. 1986 Effects of Systematic Formative Evaluation: A Meta-Analysis. *Exceptional Children,* 53: 199–208.

Goldacre, B. (2013) *Building evidence into education.* London: DfE

Goldfeld, S., N. Napiza, et al. (2011) Outcomes of a universal shared reading intervention by 2 years of age: The Let's Read trial. *Pediatrics*, 127(3): 445-453.

Gorard, S. (2014) A proposal for judging the trustworthiness of research findings, *Radical Statistics*, 110, 47-60).

Gorard, S., Siddiqui, N. and See, B.H. (2015) How effective is a summer school for catch-up attainment in English and maths? *International Journal of Educational Research*, 73: 1-11.

Gorard, S. (2016) Damaging real lives through obstinacy: Reemphasising why significant testing is wrong. *Sociological Research Online*, 29 February 2016. http://journals.sagepub.com/doi/abs/10.5153/sro.3857

Gorard, S., Siddiqui, N. and See, B.H. (2015) How effective is a summer school for catch-up attainment in English and maths? *International Journal of Educational Research*, 73: 1-11.

Gorard, S., See, B.H. and Siddiqui, N. (2017) The trials of evidence-based education: *The promises, opportunities and problems of trials in education*. London: Routledge

Harland, J., Kinder, K., Lord, P., Alison, S., Schagen, I., Haynes, J. with Cusorth, L. White, R. and Paola, R. (2000) *Arts education in secondary schools: Effects and effectiveness*. Slough: NFER

Harris, R, and Ratcliffe, M. (2005) Socio-scientific issues and the quality of exploratory talk –what can be learned from schools involved in a 'collapsed day' project? *The Curriculum Journal*, 16(4): 439-453.

Hattie, J. (2008) *Visible Learning*, London: Routledge

Hattie, J. and Timperley, H. (2007) The power of feedback. *Review of Educational Research*, 77(1): 81-112.

Hetland, L. and Winner, E. (2001) The arts and academic achievement: what the evidence shows. *Reviewing Education and the Arts Project; executive summary*, 102(5): 3-6.

Hetland, L. (2000) *The relationship between music and spatial processes: A meta-analysis.* Unpublished EdD thesis. Cambridge, Massachusetts: Harvard University.

Higgins, S., Kokotsaki, D. and Coe, R. (2012) *The Teaching and Learning Toolkit*. London: Education Endowment Foundation

Hirsch, E. (1987) *Cultural literacy: what every American needs to know*, New York: Vintage

Hopfenbeck, T. and Stobart, G. (2015) Large scale implementation of assessment for learning, *Assessment in Education*, 22 (1): 1-2

House of Lords (2014) The case for arts education in schools. *House of Lords library note*, LLN-2014-037-2. London: House of Lords.

HMSO (2003) *Every Child Matters. Report presented by the Chief Secretary to the Treasury by Command of Her Majesty.* Cm 5860. Norwich: The Stationery Office

Hsieh, H.Y. (2011) The effect of music on spatial ability. Conference paper published in *Internationalisation, Design and Global Development*, 6775:185-19

Hudziak, J.J., Albaugh, M.D., Ducharme, S., Karama, S., Spottswood, M., Crehan, E. and Botteron, K.N. (2014) Cortical thickness maturation and duration of music training: Health-promoting activities shape brain development. *Journal of the American Academy of Child & Adolescent Psychiatry, 53*(11): 1153-1161.

Jacob, B.A., and Lefgren, L. (2002) *Remedial Education and Student Achievement: A Regression-Discontinuity Analysis*. National Bureau of Economic Research Working Paper 8918, May.

James, C.Y. (2011) *Does arts infused instruction make a difference? An exploratory study of the effects of an arts infused instructional approach on engagement and achievement of third, fourth, and fifth grade students in mathematics*. PhD thesis. Washington: American University.

Jausovec, N., Jausovec, K. and Gerlic, I. (2006) The influence of Mozart's music on brain activity in the process of learning. *Clinical Neurophysiology*, 117 (12) 2703-2714.

Joseph, A. (2014) *The Effects of Creative Dramatics on Vocabulary Achievement of Fourth Grade Students in a Language Arts Classroom: An Empirical Study*. EdD thesis. Washington: Seattle Pacific University.

Kaufman Best Practices Project. (2004). *Kaufman Best Practices Project Final Report: Closing the Quality Chasm in Child Abuse Treatment; Identifying and Disseminating Best* Practices. Kasas City, Missouri: The Ewing Marion KauffmanFoundation. http://www.chadwickcenter.org/Documents/Kaufman%20Report/ChildHosp-NCTAbrochure.pdf

Kluger, A.N. & DeNisi, A. (1996) The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2): 254-284.

Lints, A. and Gadbois, S. (2003) Is listenint to Mozart the only way to enhance spatial reasoning? *Perceptual and Motor Skills*, 97(3): 1163-1174.

Little, R. J. A., and Rubin, D. B. (1987) *Statistical analysis with missing data.* New York: Wiley

Lonigan, C. and Whitehurst, G. (1998) Relative efficacy of parent and teacher involvement in a sharedreading intervention for preschool children from low-income backgrounds. *Early Childhood Research Quarterly*, 13(2): 263-90.

Lysakowski, R., and Walberg, H. (1982) Instructional effects of cues, participation, and corrective feedback: A quantitative synthesis. *American Educational Research Journal*, *19*: 559–578

Martin, K., Sharp, C. and Mehta, P. (2013) *The impact of the summer schools programme on pupils*: Reseach Report. Slough: NFER

Melhuish, E., Belsky, J. and Leyland, A. (2010) *The impact of Sure Start local programmes on five year olds and their families*, DfE Research Report No. RR067. London: DfE

Morgan, A. (2005) Shared reading interactions between mothers and pre-school children: case studies

of three dyads from a disadvantaged community, *Journal of Early Childhood Literacy*, 5(3): 279-304.

Nelson, J. and O'Bierne, C. (2014) *Using evidence in the classroom: What works and why?* Slough: NFER

Nering, M.E. (2002) *The effect of piano and music instruction on intelligence of monozygotic twins.* Unpublished PhD thesis. Honolulu: University of Hawaii.

Nutbrown, C. and Hannon, P. (2011), Effects of the REAL Project: Raising Early Achievement in Literacy, available at: www.sheffield.ac.uk/research/impact/stories/fcs/22 (accessed 21 July 2015).

Nutley, S., Davies,H. and Walter, I. (2002) Evidence based policy and practice: Cross sector lessons from the UK. *ESRC UK Centre for Evidence Based Policy and Practice, Working Paper 9*. Available: https://www.kcl.ac.uk/sspp/departments/politicaleconomy/research/cep/pubs/papers/assets/wp9b.pdf

Ofsted (2008) *Assessment for Learning: The impact of National Strategy Support.* Ofsted Report reference no. 070244. London: Ofsted.

Olson, C.A. (2010) Music Training Causes Changes in the Brain. *Teaching Music, 17*(6): 22-22.

Peppler, K.A., Powell, C.W., Thompson, N. and Catterall, J. (2014) Positive impact of arts integration on student academic achievement in English language arts. *Educational Forum*, 78(4): 364-377.

Robinson, A.H. (2013) Arts integration and the success of disadvantaged students: A research evaluation. *Arts Education Policy Review*, 114(4): 191-204.

Runfola, M., Etopio, E., Hamlen, K., and Rozendal, M. (2012) Effect of Music Instruction on Preschoolers' Music Achievement and Emergent Literacy Achievement. [Article]. *Bulletin of the Council for Research in Music Education, No.*192.

Rychetnik, L., Frommer, M., Hawe, P. and Shiell, A. (2002) Criteria for evaluating evidence on public health interventions. *Journal of Epidemiology & Community Health*, Vol. 56, No. 2, pp. 119-127.

Scheerens, J. (1992) *Effective schooling: Research, theory and practice*. London: Cassell.

Schellenberg, .G. and Hallam, S. (2005) Music listening and cognitive abilities in 10 and 11 year olds: The blur effect. Conference paper presented in *Neurosciences and Music II* , Lipzig, Germany.

See B.H and Kokotsaki, D. (2016) The impact of arts education on children's learning and wider outcomes. *Review of Education,* 4(3): 234-262

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

See, B.H. and Gorard, S. (2015) Does intervening to enhance parental involvement in education lead to better academic results for children? An extended review, *Journal of Children's Services*, 10(3): 252-264.

Senéchal, M., Pagan, S., Level, R. and Ouellette, G.P. (2008) Relations among the frequency of shared reading and 4-year old children's vocabulary, morphological and syntax comprehension and narrative skills. *Early Education and Development*, Vol. 19 No. 1, pp. 27-44.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001) *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston: Houghton Mifflin.

Simpson, A. (2017) The misdirection of public policy: comparing and combining standardised effect sizes. *Journal of Education Policy*. 32(4): 450-466

Skiba, R., Casey, A., and Center, B. (1985–1986) Nonaversive procedures in the treatment of classroom behavior problems, *Journal of Special Education*, *19*: 459–481

Slavin, B. (2016) *Evidence and the ESSA*. Huffpost. 8 December 2016 http://www.huffingtonpost.com/robert-e-slavin/evidence-and-the-essa_b_8750480.html

Smith, H. (2011) *The effects of a drama-based language intervention on the development of theory of mind and executive function in urban kindergarten children*. PhD thesis. Atlanta, GA: Georgia State University.

Smith, E., Gorard, S. (2012), 'Teachers are kind to those who have good marks': a study of Japanese young people's views of fairness and equity in school, *Compare*, 42(1): 27-46.

Smithrim, K. and Upitis, R. (2005). Learning through the Arts: Lessons of engagement. *Canadian Journal of Education*, 28*(1/2): 109-127.*

Siraj-Blatchford, I. and Taggart, B. (2014) *Exploring Effective Pedagogy in Primary Schools: Evidence from Research*. London: Pearson.

Stevens, B.J. (1996) *Parental influences in getting children "ready to learn".* PhD dissertation, Texas Woman's University. Dissertation Abstracts International Section A: Humanities and Social Sciences, 1996. 57(5-A), 121pp

Strait, D. L., Parbery-Clark, A., Hittner, E. and Kraus, N. (2012) Musical training during early childhood enhances the neural encoding of speech in noise. *Brain and Language, 123*(3): 191-201.

Sunmonu, K., Larson, J., Van Horn, Y., Cooper-Martin, E. and Nielsen, J. (2002) *Evaluation of the Extended Learning Opportunities Summer Program.* Rockville, MD: Office of Shared Accountability, Montgomery County Public Schools.

Terry, M. (2011) *Exploring the additive benefit of parental nurturance training on parent and child shared reading outcome*. Unpublished PhD thesis, Texas A&M University, Texas.

Thomas, R. and Arnold, A. (2011) The A+ Schools: A New Look at Curriculum Integration. *Visual Arts Research*, 37(72): 96-104.

Thompson, D.J. (2005) *The impact of classical music on the developmental skills of preschool children.* Unpublished PhD thesis. Mississipi: Mississippi State University.

Vaughn, K. and Winner, E. (2000) SAT scores of students who study the arts: What we can and cannot conclude about the association*. The Journal of Aesthetic Education*, 34(3-4): 77-89

von Rossberg-Gempton, I.E. (1998) *Creative dance: Potentiality for enhancing psychomotor, cognitive, and social-affective functioning in seniors and young children.* PhD thesis. British Columbia, Vancouver: Simon Fraser University (Canada).

Voracek, M. (2008) Wishful thinking: Meta-analysis reveals publication bias, lab differences and nonspecificity in the Mozart effect. *International Journal of Psychology*, 43(3-4: 690-690

Walker, P.A. (2008) *The impact of a series of poetry workshops on the cognitive development of middle school students.* Unpublished dissertation. Los Angeles: University of California.

White, B. and Frederiksen, J. (1998) Inquiry, modelling and meta-cognition: making science accessible to all students. *Cognition and Instruction*, 16(1): 3-118

Yorke-Viney, S.A. (2007) *An examination of the effectiveness of arts integration in education on student achievement, creativity, and self perception*. PhD thesis. Scranton, PA: Marywood University.