LEARNING ABOUT CULTURE

Overarching Evaluators' Report

September 2021

Jake Anders, Nikki Shure, Dominic Wyse, UCL Institute of Education

Kimberly Bohling, Alex Sutherland, Matthew Barnard, Johanna Frerichs,

Behavioural Insights Team

The Education Endowment Foundation (EEF) is an independent grant-making charity dedicated to breaking the link between family income and educational achievement, ensuring that children from all backgrounds can fulfil their potential and make the most of their talents.

The EEF aims to raise the attainment of children facing disadvantage by:

- identifying promising educational innovations that address the needs of disadvantaged children in primary and secondary schools in England;
- evaluating these innovations to extend and secure the evidence on what works and can be made to work at scale; and
- encouraging schools, government, charities, and others to apply evidence and adopt innovations found to be effective.

The EEF was established in 2011 by the Sutton Trust as lead charity in partnership with Impetus (formerly Impetus Trust) and received a founding £125m grant from the Department for Education.

Together, the EEF and Sutton Trust are the government-designated What Works Centre for improving education outcomes for school-aged children.

For more information about the EEF or this report please contact:

Jonathan Kay
Education Endowment Foundation
5th Floor, Millbank Tower
21–24 Millbank
SW1P 4QP

0207 802 1653

jonathan.kay@eefoundation.org.uk

www.educationendowmentfoundation.org.uk

Contents

Summary and key findings

This report sets out findings and lessons learned from an ambitious and innovative programme of work funded by the Education Endowment Foundation (EEF), the Royal Society of Arts (RSA), Arts Council England and the Paul Hamlyn Foundation (PHF), delivered over four years, including 8500 children in 400 state schools across England. This body of work represents a major step forward in terms of evaluations of arts-based interventions on pupil attainment, making a significant contribution to what we know, and what we don't, about the impact of arts-based interventions.

The impact evaluations of the five trials have not delivered evidence of statistically significant improvements in the measures of pupils' academic attainment used in the project, compared to pupils in the business-as-usual control groups. However:

- making substantial improvements in pupil attainment compared to business-as-usual is an extremely challenging barrier to clear. These trials are not unusual among interventions evaluated by EEF-funded trials in this respect;
- the reported outcomes of the trials equally provide little evidence of detrimental effects from implementing such programmes. We emphasise this point because risks of neglecting the delivery of 'core' curriculum subjects are sometimes raised as a concern when considering use of arts-based education interventions in schools;
- the findings do *not* mean that schools should not implement these programmes or discontinue programmes they already use. Improving pupil academic attainment is not the best or only reason for schools to implement arts-based interventions in schools;
- however, and in line with previous systematic reviews, implementing arts-based programmes (specifically the ones that we have trialled) primarily in order to improve pupil attainment in literacy continues not to have a strong basis.

We highlight important considerations for organisations engaged with arts-based education on how to refine their programmes to help ensure implementation within schools is as smooth as possible. This includes:

- ensuring adaptability to differing contexts within different schools;
- engagement with school staff at multiple levels (senior leadership and classroom teachers, for example, not one or the other);
- planning to be flexible within school structures such as timetabling (rather than asking schools to work around the programmes in this respect);
- an increased focus on planning to scale-up programmes successfully: it cannot just be assumed that an intervention that is highly effective when delivered by its developers in a small number of schools will continue to work on a larger scale.

The findings raise important considerations both for us as evaluators (and others in the evaluation community), and also for EEF and others as funders who are interested in continuing to develop the evidence base about arts-based interventions in education. Some of the most important are:

- building evidence requires an accumulation of evidence first, about how interventions affect their most direct intermediate outcomes, before moving to evaluate outcomes that are more distant from the intervention, such as pupil attainment. This project shows the limitations of attempting to skip any of the steps that are implied in this necessary accumulation of evidence.

We were limited by our methods for collecting different outcomes, particularly at scale. The reliance on traditional, more formalised survey methodology may prevent us from capturing important outcomes that are on pathways to attainment, or to outcomes that are important for pupils but unrelated to attainment. New thinking, development and investment in these areas are vital to helping this kind of project: measuring what we value, rather than risk valuing only what we can (currently) measure.

Introduction

In 2017 the Royal Society for the Encouragement of Arts, Manufactures and Commerce (RSA) and the Education Endowment Foundation (EEF), with support for programme delivery from Arts Council England and the Paul Hamlyn Foundation (PHF), announced a partnership to deliver the UK's largest ever study of arts-based education. The resulting Learning about Culture (LAC) programme[1] is a major study into the role that arts-based education plays in improving educational outcomes for children. Its intention is to develop more evidence of what works, and support schools and other organisations who engage with the arts to use that evidence, as well as evidence from their own work and elsewhere, to improve their practice and their evaluations of practice.

Concern that a focus on increasing attainment in literacy and numeracy has led to marginalisation of art, music and cultural studies in English schools (Neelands et al., 2015) has increased the motivation for understanding links between arts-based learning and academic attainment. The UK Government's Culture and Sport Evidence review (Newman et al., 2010), showed pupil participation in cultural learning programmes (from piano training to theatre-based drama projects) to be correlated with higher levels of achievement in mathematics and literacy / English in both primary and secondary school. The review also linked participation in cultural learning programmes to faster language development in the early years and improved cognitive ability. Additionally, large cohort observational studies in the US have suggested that the mathematics and literacy gains from participation in arts activities (including, for example, out-of-school dance lessons) are particularly large for pupils from low-income groups (Catterall, 2009; Catterall et al., 2012).

However, these **associations** between arts-based activities and higher levels of achievement are not the same as evidence of a causal link. An international systematic review commissioned by the Organisation for Economic Co-operation and Development (OECD) highlighted the lack of robust causal evidence linking arts-based education interventions and academic attainment (Kautz et al., 2014). Similarly, EEF's previous work on arts education, including its toolkit (EEF, 2018) and specific review of the existing state of the evidence in this area, 'found no convincing evidence that demonstrated a causal relationship between arts education and young people's academic ... outcomes' (See & Kokotsaki, 2015). Nonetheless, in launching that review, former EEF director Sir Kevan Collins highlighted that 'All children, including those from disadvantaged backgrounds, deserve a well-rounded, culturally rich, education. However, many have gone further than this, arguing that arts education itself directly improves pupil attainment' (Collins, 2015). This project, as outlined further in the launch prospectus (Londesborough et al., 2017), was undertaken in the spirit of both parts of this argument: that the best argument for arts education is for its own sake, and that we should rigorously interrogate claims of impact on pupil attainment, given they are both regularly deployed and not as well understood as we would like.

At the core of the LAC programme are five school-based randomised controlled trials of arts-based education interventions carried out involving around 8,500 children in 400 state schools across England, with a considerable proportion of pupils eligible for pupil premium. These trials represent the biggest study of its kind ever undertaken in the UK and provide much-needed insight into both what works and how it works. Two of these randomised controlled trials were focused on Key Stage 1 (KS1: children aged 5 to 7; Year 1: children aged 6 to 7, and Year 2: children aged 7 to 8) and three are focused on Key Stage 2 (KS2: children aged 7 to 11; Year 5: children aged 10 to 11). Despite the unique aspects of each of the interventions, there are important similarities in how they are delivered and what they hope to achieve. The five programmes are as follows:

1	Speech Bubbles	(SB)	KS1
2	First Thing Music	(FTM)	KS1
3	The Craft of Writing	(CoW)	KS2

[1] See the RSA website for further details (https://www.thersa.org/globalassets/pdfs/reports/rsa-learning-about-culture-report.pdf).

4 Power of Pictures (PoP) KS2
5 Young Journalist Academy (YJA) KS2

The projects have been evaluated by a collaboration between UCL Institute of Education and the Behavioural Insights Team (BIT), who worked with the developer teams and with the EEF and the RSA, in order to design trials that would maximise what we can learn from the individual studies, but also reveal some overarching concepts derived from learning from all of the trials. Particularly notable with regard to the rigorous approach to both individual trials and learning from all five trials was:

- the use of common outcome measures where possible;
- harmonised aspects of trial design;
- a common team across all five projects designing and refining the projects' logic models.

With regard to outcome measures, the teams were able to agree on:

a a common primary outcome measure (Progress in Reading Attainment, PIRA), aiming to capture pupils' reading skills, across both KS1 studies;
b a common primary outcome measure (the writing assessment measure, WAM), aiming to capture pupils' writing skills, across the three KS2 studies;
c a common secondary outcome measure (the writing self-efficacy measure, WSEM), aiming to capture writing self-efficacy across the same three KS2 studies; and the WSEM3, an adapted version of the WSEM, to capture writing self-efficacy for the two KS1 studies with younger pupils;
d a further similar secondary outcome measure (a measure of creative self-efficacy drawn built on the Ideation sub-domain of the WSEM, albeit with some adaptation due to age range variation) across all five studies.

Individual reports are available for each of the interventions, reporting the findings of each intervention's impact evaluation and implementation and process evaluation. As such, this overarching report from the teams at UCL and BIT focuses on what we can learn from the evidence the five studies have generated as a whole. In particular, as well as summarising and synthesising the study findings, we draw out the lessons from the process of carrying out the research. There have been particular challenges in carrying out the research, due to the nature of the interventions, their expected outcomes and how we have tried to design the trials to allow for this overarching synthesis as its culmination.

We aim to speak to three key audiences, with particular focus on these different readers in particular parts of the report:

1 researchers looking to conduct evaluations of interventions of this type in future, as well as EEF and other potential funders planning to commission evaluations of interventions of this type (see Section 3);
2 organisations that currently deliver interventions of this type and are looking to build the evidence base to develop their programmes (see Section 4);
3 schools considering use of this type of programme as part of their curriculum (see Section 5).

While the impact evaluations of the five trials have not delivered evidence of statistically significant improvements in pupils' academic attainment relative to the control group, we emphasise that this is an extremely challenging barrier to clear. The outcomes of these trials are not unusual among EEF-funded trials in this respect and, while it is easy to focus on the lack of a positive impact in the outcome measures compared to business-as-usual, it is equally the case that the trials provide little evidence of detrimental effects from introducing such programmes. We emphasise this point, because the risk of neglecting the 'core' curriculum subjects is sometimes raised as a concern when considering the use of innovative arts-based approaches in schools, and such concerns may be behind reported declines in arts teaching in primary schools (Cooper, 2018). We find little evidence of such harms, and a great deal of enthusiasm for the interventions among the teachers and pupils who were part of them.

There are also important findings that developers of arts-based programmes should reflect on. Our implementation and process evaluations, in particular, highlight findings that appear common across interventions and which we judge to be relevant to many organisations who work with schools on programmes of this type. We also advocate an increased focus on planning, to scale-up programmes successfully: it cannot just be assumed that an intervention that is highly effective when delivered by its developers in a small number of schools will continue to work on a larger scale. Hence, it is vital to plan carefully to stand the best chance of retaining effectiveness at scale.

The report proceeds as follows. We begin, in Section 2, with a summary of impact and implementation and process evaluation findings across the five trials, including reporting of a pooled analysis of primary and secondary outcome measures across the three KS2 trials. Next, in Section 3, we move on to lessons for evaluations and for research funders based on the process of carrying out this research, highlighting particular issues around choice and availability of suitable outcome measures, and practicalities of trial design with interventions of this type. We then discuss lessons for arts organisations in Section 4, focusing particularly on considerations for scaling, self-evaluation and practical considerations for implementation in schools that may not otherwise be obvious, before concluding, in Section 5.

Summary of evaluation findings across the five trials

The primary goal of these five evaluations was to measure the impact of arts-based education activities on reading and/or writing attainment. The secondary goal was to measure their impact on writing self-efficacy, creativity and the ability to generate ideas. We were also interested in understanding which factors affected implementation. The overall impact evaluation results showed no statistically significant effect of participation on the attainment measures used (explained below in more detail). Similarly, there was no statistically significant impact on the creativity and self-efficacy secondary outcome measures; however, the Power of Pictures intervention showed promise on the writing self-efficacy measure. Equally, the interventions were not found to be harmful to the measures of pupil attainment used, nor was there much evidence that they crowded out other beneficial activities; most teachers viewed the programmes very positively. The evidence from these trials does not indicate that schools should not continue to pursue arts-based education. The implementation and process evaluation highlighted individual-level and structural-level factors that affected implementation, with special attention drawn, for example, to the role of the school's senior leadership team (SLT).

Outcome measures

Depending on whether the trial was targeted at the younger primary school pupils (KS1 trials: First Thing Music and Speech Bubbles) or at the older primary school pupils (KS2 trials: Craft ofWriting, Power of Pictures and Young Journalist Academy), different instruments were used to capture the literacy outcomes. These measures were selected by the evaluators in discussion with project teams, the EEF and the RSA. They all have their limitations, but were felt to be appropriate (in the light of prior evidence on arts-based education highlighting improved literacy as a potential outcome) and pragmatic choices.[2] Writing was selected as the outcome for the KS2 trials, since all three interventions have an explicit focus on writing and/or producing written content. This is a difficult outcome to capture in an age-appropriate way at scale, so we adapted an existing measure, the writing attainment measure (WAM) for the KS2 trials (more detail provided below). Due to similar challenges for younger pupils, reading was selected for the KS1 trials as the measure of literacy skills, as well as oral communication due to the nature of Speech Bubbles. Table 1 provides an overview of these measures.

Table 1: Summary of instruments used to measure outcomes

Trial	Primary/secondary outcome	Outcome	Measure
First Thing Music (FTM), Speech Bubbles (SB)	Primary	Reading attainment	Progress in Reading Assessment (PIRA)
SB	Primary	Oral communication (narrative recall skills)	Renfrew Bus Story (RBS)
FTM, SB	Secondary	Social skills	Social skills sub-scale of the Social Skills Improvement System (SSIS)

[2] Further detail on each of the measures, including reasons for selecting them, administration and limitations, is provided in each of the individual programme reports.

FTM, SB	Secondary	Creative self-efficacy	Adapted version of the Ideation sub-measure of the writing self-efficacy measure (WSEM), referred to as WSEM3.
Craft of writing (CoW), Power of Pictures (PoP), Young Journalist Academy (YJA)	Primary	Writing attainment	Writing attainment measure (WAM) score (with the Ideas scale double weighted)
CoW, PoP, YJA	Secondary	Writing self-efficacy	Writing self-efficacy measure (WSEM) proposed by Bruning et al. (2013), adapted for primary school use.
CoW, PoP, YJA	Secondary	Ideation	First five questions of WSEM (Ideation sub-scale).

For the KS1 trials, the primary outcome measure was the Progress in Reading Assessment (PIRA), and in addition, for Speech Bubbles, the Renfrew Bus Story (RBS). The PIRA is a standardised assessment of pupils' reading attainment and profile of their reading skills. It measures reading ability in the following areas: phonics, literal comprehension and reading for meaning. The RBS is a short, standardised test that assesses narrative aspects of oral language. The administrator tells the pupil a story using a short picture book with no text as an aid. The pupil is then asked to retell the story using the picture book.

The secondary outcome measure for the KS1 trials was the social skills sub-scale of the Social Skills Improvement System (SSIS). This commonly used standardised measure assesses pupils' skills across the following sub-scales: Communication, Cooperation, Assertion, Responsibility, Empathy, Engagement and Self-control. Teachers were asked to complete the assessments of their pupils.

We note that robust assessment of writing is challenging, particularly during primary schooling. There are few measures available, and none have been used in a similar context. In the face of these constraints, the measure we chose for the three KS2 trials, the Writing Assessment Measure (WAM), was a pragmatic choice, which comes with some limitations (e.g., in terms of how relatively new it is); however, it was designed for the context of the English educational system and existing evidence suggests that it is a valid, consistent and reliable measure (Murphy et al., 2013; Dunsmuir et al., 2015). The evaluation team for these trials also conducted a small-scale pilot of the WAM, which affirmed this. The WAM is designed to assess narrative writing in response to a written prompt, for which pupils are given 15 minutes to write.

The secondary outcome measures for the KS2 trials were the Writing Self-Efficacy Measure (WSEM) and the Ideation sub-scale of the WSEM. A simplified and shortened version of that WSEM Ideation sub-scale was created for the KS1 trials (which we referred to as the WSEM3); further details below. To measure writing self-efficacy, we used a version of the measure proposed by Bruning et al. (2013), which has been adapted for primary school pupils with some simplification of language. The WSEM involves 16 statements capturing pupils' perceptions of their writing capabilities, including 'I can think of many ideas for my writing' and 'I can avoid distractions while I write', with pupils giving marks out of 100 for their self-assessment in each of these. We used slightly simplified versions of some of the statements to better suit the primary school context. In addition, we requested responses on a five-point Likert scale. To measure ideation for KS2 pupils, we used the first five questions of the writing self-efficacy measure, a sub-scale which focuses on creativity. A further shortened and simplified version of this Creativity sub-scale (including switching to a three-point Likert scale) was used as an analogous measure of ideation for KS1 pupils.

Impact evaluation results

An overarching finding based on all five trials was 'no statistically significant impact' of participation on any of the attainment measures. By not statistically significant, we mean that the statistical uncertainty (which is inherent to quantitative impact evaluations) around our estimated impacts is such that there is a substantial probability that such a finding could have arisen by chance. There were particular challenges in this study which exacerbated this uncertainty.

Table 2 presents the impact evaluation results for each trial and primary outcome measure along with pooled analysis for the KS2 trials which had the same design and same outcome measures (see Appendix A for an overview of the aims and method for this pooled analysis). The goal of the additional pooled analysis (which was not part of a pre-specified analysis plan) is to trade intervention specificity for an increased sample size, and thereby statistical power, to see if participation in a non-specific KS2 arts-based education intervention has the potential to lead to pupil progress on the primary or secondary outcome measures, based on averaging across the three trials.

The effect sizes presented in Table 2 are small in magnitude, ranging from minus two to positive one months' progress, but due to uncertainty in the estimates, they are not statistically significantly different from zero. Even when results are pooled across the three KS2 trials, in the case of the WAM the overall picture does not change. Pupils who participated in these interventions did not, on average have statistically significantly higher attainment at the end of the school year as compared to their peers in the business-as-usual control group.

Table 2: Summary of impact on primary outcomes

Trial	Outcome	Effect size (95% CI)	Estimated months' progress (95% CI)	EEF security rating	No. of pupils	*p*-value	EEF cost rating
First Thing Music (FTM)	PIRA[a] score	0.07 (–0.02,0.19)	1 (0, 3)	2	2150	0.13	£ £ £ £ £
Speech Bubbles (SB)	PIRA[a] score	–0.05 (–0.2, 0.1)	–1 (–3, 2)	4	821	0.93	£ £ £ £ £
SB	Renfrew Bus Story score	–0.04 (–0.19, 0.11)	0 (–3, 2)	4	811	> 0.99	£ £ £ £ £
Craft of Writing (CoW)	WAM[b] score	–0.03 (–0.19, 0.12)	0 (–3, 2)	2	1697	0.68	£ £ £ £ £
Power of Pictures (PoP)	WAM[b] score	0.09 (–0.07; 0.24)	1 (–1, 3)	3	1945	0.27	£ £ £ £ £
Young Journalist Academy (YJA)	WAM[b] score	–0.13 (–0.32, 0.05)	–2 (–4, 1)	3	1613	0.16	£ £ £ £ £

Pooled Key Stage 2	WAM[b] score	–0.04 (–0.14, 0.07)	0 (–2, 1)	N/A	5255	0.49	N/A

Notes. [a] Progress in Reading Assessment; [b] Writing assessment measure (Ideas scale double weighted).

The results for the secondary outcome measures are more positive. For each trial, there is again little evidence that participation in these programmes led to changes in pupils' creativity or how they viewed themselves as writers. Most of the effect sizes presented in Table 3 are positive, small and not statistically significant. One exception is the PoP, which achieved progress (effect size of 0.11) on the writing self-efficacy measure for participating pupils, and was statistically significant at the 10 percent significance level. This provides suggestive evidence that participation in the PoP programme can improve pupils' perception of themselves as writers and their confidence, even with the uncertainty around these estimates. The pooled results for the three KS2 trials show a small and positive effect for the writing self-efficacy measure, which is small in magnitude and statistically significant at the 10 percent level. They also show a small, positive effect on the ideation measure, which is not statistically significant. Taken together this provides suggestive evidence for the potential for arts-based education interventions in KS2 to achieve positive (if small) effects on pupils' confidence and creativity as writers but highlights the challenges of achieving statistical power to do so.

Free school meal (FSM)-eligible pupils were examined as a separate sub-group of interest, but no differential effects were found (more details are available in each of the separate evaluation reports). Consistent with the outcomes for all pupils, they also did not experience statistically significant improvements in attainment or self-efficacy/creativity as a result of participating in these trials.

Table 3: Summary of impact on secondary outcomes

Trial	Outcome	Effect size (95% confidence interval)	No. of pupils	*p*-value
First Thing Music (FTM)	WSEM3 score	–0.03 (–0.14, 0.07)	1931	0.51
Speech Bubbles (SB)	WSEM3 score	0.05 (–0.09, 0.19)	770	0.47
FTM	SSIS score	–0.09 (–0.24, 0.08)	852	0.27
SB	SSIS score	0.03 (–0.12, 0.17)	746	0.71
Craft of Writing (CoW)	WSEM score	0.04 (–0.08, 0.15)	1674	0.55

Power of Pictures (PoP)	WSEM score	0.11 (–0.00, 0.22)	1907	0.06
Young Journalist Academy (YJA)	WSEM score	0.03 (–0.09, 0.16)	1571	0.61
Pooled Key Stage 2 (KS2)	WSEM score	0.07 (–0.00, 0.14)	5152	0.06
CoW	Ideation	0.02 (–0.09, 0.14)	1674	0.69
PoP	Ideation	0.09 (–0.03, 0.21)	1907	0.14
YJA	Ideation	0.03 (–0.10, 0.16)	1571	0.68
Pooled KS2	Ideation	0.06 (–0.01, 0.13)	5152	0.12

Implementation and process evaluation (IPE) results

The IPEs for these five interventions were designed with several common overarching questions to allow for the identification of themes across the findings. The goals of the overarching questions were to understand:

- how the intervention was delivered and what contributed to delivery with fidelity;
- how responsive schools were to the intervention;
- how schools perceived the quality of the intervention;
- how the knowledge of the arts-based practitioners was integrated with the knowledge of the teaching staff involved.

For each individual trial, there were also additional IPE research questions that were specific to the programme.

The IPEs for these five trials used a range of methods to answer these questions. This included case studies of six purposively sampled intervention-arm schools per trial. The case studies included:

- interviews with teaching staff and SLT members;
- informal discussions with pupils participating in the intervention;
- observations of intervention lessons.

The sampling criteria for all interventions included the proportion of pupils receiving FSM and an indicator of engagement (either level of delivery of the intervention or attendance at teacher training). Teachers in all participating schools were asked to complete a survey (response rates were 80%, on average across the trials). The research team also conducted observations of programme training and collected administrative data from the delivery teams. The design of the five trials was compared and an amalgamated logic model was developed to highlight similarities. Figure 1 shows this logic model, which was developed during the design phase of the evaluation.

From the IPE findings across the five interventions, we observed three main areas that appeared to affect implementation: intervention characteristics; individual-level factors; and structural factors. These themes are drawn from the 30 case studies which were purposively sampled for each intervention, so as to observe the range and diversity of implementation experiences. As such, these themes may not be representative of all schools who implemented the interventions; however, given that these findings emerged across five very distinct interventions, they are worth noting. The factors affecting implementation are described briefly here and presented in more detail in section 4.2.

a **Intervention characteristics:** The interventions can be seen as being on a spectrum of the degree of specification of implementation. Highly specified interventions generally had higher levels of fidelity, but may have been more challenging to scale. Less specified interventions were generally more adaptable to the local context, but also required more effort from the teacher to plan elements that drew on the specific classroom and school context.
b **Individual-level factors:** Factors at the teacher and pupil levels were seen to influence implementation.
c **Structural factors:** SLT support, resources and timetabling were identified as the key structural factors affecting implementation.

In all five trials, in surveys and interviews (see Table 4 for survey over-arching survey findings), teaching staff reported perceived benefits of their programme on pupils. The majority of respondents from the KS2 evaluations perceived a positive impact of their intervention on writing skills, the primary outcome measure, and ideation, a secondary outcome. The majority of respondents from the KS1 evaluations were not sure about the impact of their intervention on reading, a primary outcome; however, a majority did perceive a positive impact on speech and language and communication skills. Across all five interventions there was a perceived benefit in terms of pupil creativity. In the light of the importance of

consensual judgement to assess creativity, noted in the research literature, the perceptions of the respondents are important.

Improved pupil engagement was a theme that emerged in both surveys and interviews across the five interventions. A majority of survey respondents indicated their respective intervention had a positive impact on pupil engagement. In interviews with some teaching staff, a theme emerged about the interventions being effective in engaging some pupils who might not otherwise be as engaged with the standard literacy lesson or activities.

Table 4: Summary of teaching staff survey results

	First Thing Music (FTM)	Speech Bubbles (SB)	Craft of Writing (CoW)	Power of Pictures (PoP)	Young Journalist Academy (YJA)
Percentage who perceived positive impact					
Reading	41% + (56% not sure or neither positive nor negative)	38% + (62% not sure or neither positive nor negative)	–	86%	57% (43% limited impact)
Speech and language	77%	81% +	–	–	–
Writing skills	–	–	92% +	84% +	69% +
Communication	83%	86%	–	81%	74%
Social skills	93% ++	95% ++	–	59% (41% no impact or don't know)	74%
Ideation	–	–	97% ++	95% ++	63% ++
Creativity	85%	86%	97% (specifically writing)	97%	69%
Engagement	88%	86%	98% (specifically writing)	98%	74% (specifically, with culture and the wider world)
Confidence	–	91%	97% (specifically writing)	95%	77%

Notes. + indicates primary outcome measure in impact evaluation; ++ indicates secondary outcome measure; – indicates not asked.

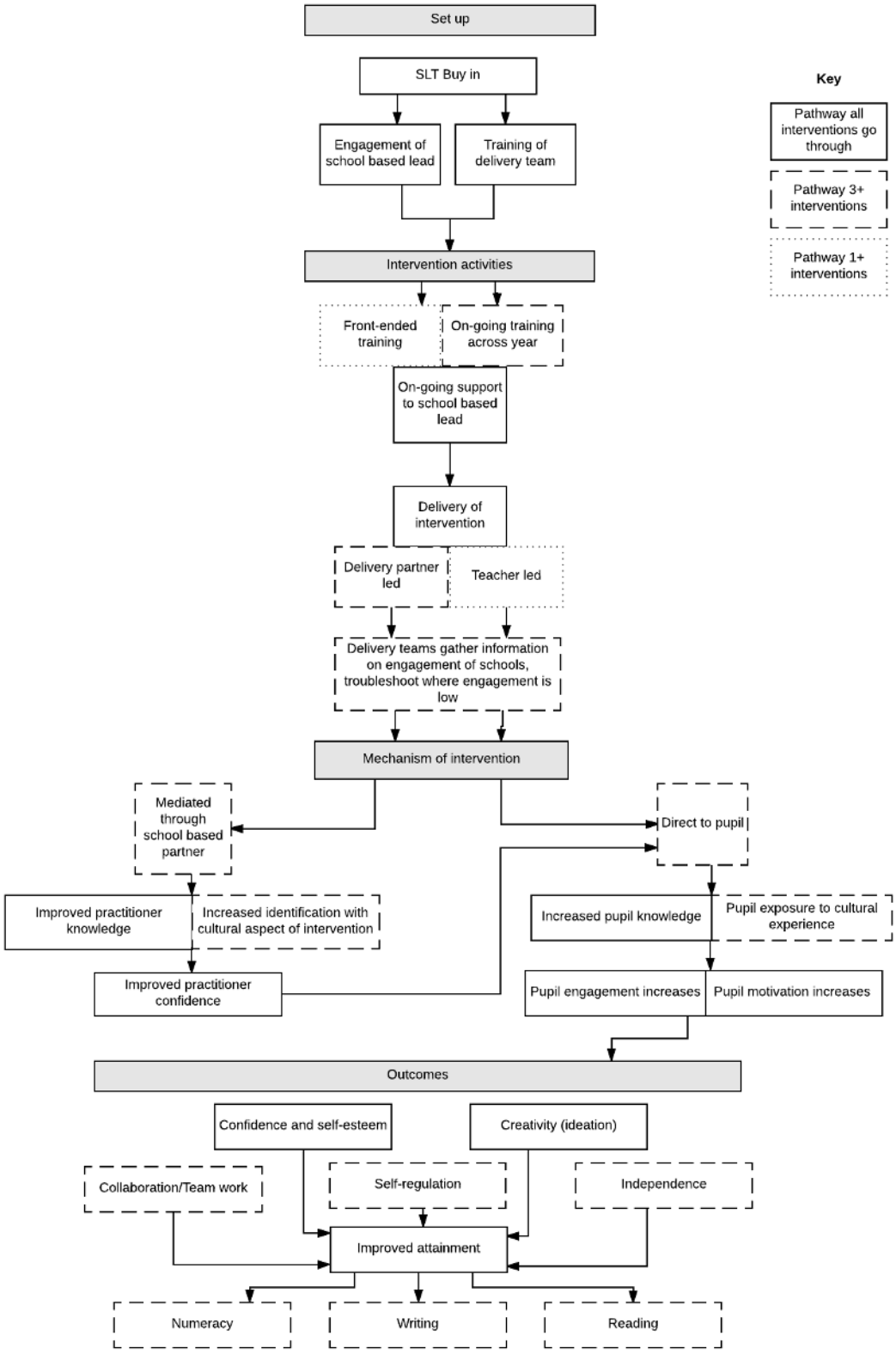
Taken together, these findings highlight what is shown in the amalgamated logic model (Figure 1). None of the interventions were straightforward to implement with high fidelity and achieve impact. In some schools, pupils and teachers found interventions more straightforward. This may have been due to prior experience with these types of programme, different pupil cohorts between schools, and varying levels of support and buy-in from SLTs. Even for interventions with a long track record of implementation, shortages of supplies and space proved challenging for schools. This indicates that schools may need more preparation (e.g., more staff meetings) for these types of intervention. Ultimately, SLT support and buy-in was identified as an overarching factor that could positively affect implementation.

How does this compare with existing evidence?

These interventions built on a small existing evidence base on the role of cultural learning interventions in improving attainment. Prior evidence, based mostly on observational and qualitative studies, showed pupil participation in cultural learning programmes to be correlated with higher levels of attainment in both primary and secondary school (Newman et al., 2010). Large cohort observational studies in the US suggested arts participation (either in or out of school; e.g., taking out-of-school dance lessons) may bring particularly large attainment gains for pupils from low-income groups (Catterall, 2009; Catterall et al., 2012). Gains in both cases were found for mathematics and literacy. There was no significant evidence base from trials on the impact of cultural learning activities on creativity or self-efficacy (see also the evidence review, which underpinned the initial EEF call, in See & Kokotsaki, 2015).

None of the five interventions evaluated here was found to have a statistically significant impact on the literacy outcome measures. There was also no evidence of a differential effect for low-income groups, as measured by FSM eligibility. There was, however, a high degree of enthusiasm reported by participating teachers, who also highlighted some of the programmes' ability to engage pupils from disadvantaged backgrounds, or those who tended not to enjoy many literacy activities.

Figure 1: Amalgamated logic model for all five LAC trials



What can evaluators and funders learn from Learning about Culture (LAC)?

In this section, we set out what we believe are the main points for researchers and funders to consider, and they largely point to stepping back and thinking again about assumptions underpinning interventions, but also about what we *can* measure versus what we *should* measure.

It is important to acknowledge upfront the context for these evaluations. Given the limited evidence base for arts-based education, the funders aimed to generate new causal evidence on five distinct programmes that also had some key similarities in approach. The projects were selected as a result of open competition as those that were most suitable for being evaluated using a randomised control trial (RCT) and that had a theory of change that led to improvement in attainment outcomes (not just arts-related outcomes). The evaluation designs needed to carefully balance accurately evaluating each individual intervention, while also standardising some research questions, methods and outcome measures to allow for some comparability across the projects and synchronisation of reporting.

The learnings and recommendations presented in this section are made in the spirit of supporting further research in this space and building upon what was learned through these evaluations, as there is still much more to learn and understand about how arts-based education impacts pupil development.

These results are not the 'last word' in whether these interventions are effective or not.

What we *can* measure in experiments (and research) is not always what we should or need to measure. This is not a reason to ignore the results of these trials – the results are robust and transparent for scrutiny – but as we set out below, we think there are ways to improve how evaluations of arts-based education interventions are undertaken, particularly if the focus is on linking to attainment. The outcomes from trials tell us that one area of renewed focus should be on basic research that focuses on the relationship between arts-based education and intended outcomes.

The need for more basic research is perhaps most obvious with how we try to measure creativity in the context of educational interventions. From the beginning of the projects, creativity has been difficult to accommodate, including as part of the data and analyses, but it is arguably fundamental to researching arts-based education interventions to be able to capture this. That is a methodological challenge as well as a challenge for theory. Consider that:

- The creativity research field has moved from measures of creativity such as Torrance tests, critique of these, then towards sociocultural perspectives (see e.g., Eckhoff & Urbach, 2008; Kim, 2011; Glăveanu, 2015; but also Kaufman & Beghetto, 2009; Kaufman, 2019; Sarma & Thomas, 2020).[3]
- Creativity is a capacity demonstrated when something new and valued is created by a child (in this case), as judged by appropriate observers, but also enabled and recognised by teachers (see Wyse, 2017; Bereczki & Kárpáti, 2018).

Taken together, the two points relating to measurement illustrate that creativity is difficult to capture in a test measure. This does not mean that it should be ignored, but we must acknowledge that including creativity in evaluations is problematic, as is its omission, particularly for arts-based education. But building on these first two points:

a The methodology had an impact on the extent to which the design could assess creativity, even in the process evaluation (as noted).
b The ideation measure (Bruning et al., 2013) was a good solution within the inherent limitations outlined above.

[3] In particular, creativity as a single construct versus a construct with multiple components.

c The essence of these arts-based education interventions was actually music; speech and theatre; reading and writing, respectively, not creativity explicitly, first and foremost.

These points raise questions about selection of, and conception of, the interventions in relation to creativity per se. They are also linked to the phenomenon of creativity being the essence of some human activity, yet not necessarily being an explicit focus of thinking during the activity.

Interrogate the science behind claimed associations

It is worth again acknowledging that, overall, there was little evidence on arts-based education to begin with, and that a key aim of the project was to grow that area of research. However, there was promising research relevant to the different interventions:

- **First Thing Music (FTM)**: as noted by Sala and Gobet (2020) in their recent meta-analysis, there was no experimental evidence (from 54 RCTs) linking music training with children's cognitive skills measured later. One reading of this is that the causal arrow might flow *from* improved cognitive skills *to* an interest in music, rather than the other way around. If that is the case, it might mean that if we intervene to improve cognitive performance then it could lead to an increased interest/participation in music in any of its diverse forms, including listening, performing or composing (if a child has such opportunities that fit with their particular interests).[4]
- **Speech Bubbles (SB)** seemed to be a promising intervention in respect of attainment outcomes. Prior evaluation evidence supported improvements in spoken language and there were claims of effectiveness relating to attainment outcomes such as literacy (Barnes, 2014).
- **Craft of Writing (CoW)**: This teacher-based intervention relied on a 'trickle down' from improving teachers' writing skills to being able to teach pupils. As with FTM, a prior systematic review found the evidence base regarding the impact of teachers' writing on students' outcomes was small and does not show a clear impact (Cremin & Oliver, 2017).
- **Power of Pictures (PoP)**: unlike other interventions, PoP had very little rigorous supporting evidence, relying on self-report and a design without a comparison group. Noting that PoP now seems to be the most promising intervention in terms of attainment outcomes, it reminds us of the difficulty in predicting successful interventions, and the scope for exploring 'wisdom of the crowd' approaches for predicting success (Simoiu et al., 2019).
- **Young Journalist Academy (YJA)**: There is significant evidence relating to training children in writing, both on writing as an outcome, but also reading (e.g., Graham & Herbert, 2011; Graham & Sandmel, 2011). There is little evidence relating to framing the process of writing as a *journalistic* output, but there is evidence of the need to situate writing as a process involving feedback and requiring an array of skills (Deane et al., 2014), which YJA fulfilled.

Despite these being the strongest applications in the funding round, with the benefit of hindsight there might have been a stronger case for not starting all the projects on the same trajectory towards large-scale trialling.

Allow more time for planning/piloting outcome measures and linking interventions to outcomes

If empirical research does demonstrate that arts-based education interventions do causally affect 'closer' outcomes which could lead to changes in outcomes that are further away, it will be important to use that new information to revisit what

[4] Meta-analyses and the literature reviews underpinning them are also variable in quality, and we have not assessed the quality of those cited here. According to Ahn et al.'s (2012) review, there was a need to improve both data evaluation and analysis for meta-analyses in education. The quality of RCTs themselves is also of course variable, as noted by many authors, including in relation to education RCTs (e.g., Wyse & Torgerson, 2017).

those ultimate outcomes of interest are. The attainment measures available to the study team were, in hindsight, several steps removed from the activity. Put another way, we may have been focusing on attainment in reading and writing to improve comparability between projects, but the outcomes might have accrued elsewhere (if at all). And by 'elsewhere', this might be on *other* attainment measures, or on pathways upstream of attainment. The secondary outcomes we used were supposed to pick up changes upstream from attainment but did not do so in many cases. Again, that might be because they were on a pathway that was untouched by the intervention. Similarly, a drive for consistency across projects might also have meant additional trade-offs in outcome selection, compared to optimising outcomes for a given intervention. We also note that there may have been benefits to otherwise disengaged or 'non-traditional' learners (as noted by one of the peer reviewers), but we were not able to capture this.

We expand one example a little more. The idea to link music directly to academic attainment, such as reading and writing, means a narrowing of vision about what else music might be affecting. If, for example, music training can improve pupils' self-control (see e.g., Moffitt et al., 2011) then there are direct pathways to better life outcomes which need not run through education. Similarly, music might be a way to help otherwise disengaged learners attend school through enjoying and looking forward to school. This would not be the only domain in which music is used in a way that might be seen as surprising: instead of using beatboxing for laryngectomy patients (Moors et al., 2020), we're using it to get pupils to come to school and (for some) using that opportunity for add-on skills like maths. This isn't detracting from the notion of 'music for pleasure' or the wider literature on the correlation between music and other outcomes, but if we're trying to illustrate that music has some 'instrumental value' (no pun intended), such as getting pupils into school, then it's worth keeping that open.

Measure what is close to the intervention first

The EEF is focused on pupil attainment because that is the outcome they care most about moving and they were careful to select programmes that had a hypothesised theory of change that led to improvements in attainment. The evaluation team selected outcomes with this focus in mind, whilst also aiming to incorporate comparability across trials where possible.

Yet this series of arts-based education evaluations has highlighted that this focus and approach may not have been well suited to evaluating the impact of these interventions. Our recommendation would be to continue to use rigorous designs but focus on measures that are much closer to the individual intervention itself. The rationale for that focus on 'closeness' to intervention is also driven by the lack of effects noted with the outcomes used. Rather than use that information to judge whether an intervention was effective or not, we think there is a case for starting with measurements closer to the interventions (mechanisms) to see if they are affected by it (leaving aside issues with measurement for now). If not, then this would suggest a theory failure and a need to revisit more basic assumptions about the relationships and direction of causality proposed by given interventions. As concrete examples:

- **a** Phonological awareness is a hypothesised attribute that sits much closer to music training and also appears in the First Thing Music logic model as a hypothesised mechanism that leads to improved reading attainment. Other discrete skills that contribute to reading attainment and were assessed by the PIRA, such as making inferences, do not appear in the FTM logic model. As such, a more targeted assessment of an individual reading skill, rather than overall reading attainment, might be a better outcome to examine for this programme.
- **b** With Craft of Writing, as a teacher-focused intervention, the logic model includes mechanisms related to improved teacher knowledge and confidence as key changes leading to improved pupil outcomes. Measuring teacher outcomes would provide evidence about whether changes are being observed at the teacher level, which would be necessary to observe changes at the pupil level. Due to the aforementioned desire to allow for comparability across trials, this approach was not taken for this evaluation, as it would not be relevant to any of the other interventions.

Pilot interventions and outcomes prior to trialling

The need for a better understanding of outcome measures and the link(s) between interventions and outcomes means that smaller steps need to be taken before claims are made about intervention impact. Beyond the basic science needed to understand what music training, for example, might plausibly impact, meticulous piloting work is also crucial regarding the measures used and how they are collected. Unless there is a plausible causal pathway that's supported by basic research, then assume that the work is exploratory and should be treated more like a pilot (if a trial). If it's not a trial, then the focus should be on establishing and testing plausible pathways between the activity and the outcome.

Classroom-level randomisation should be avoided with education trials where possible

This is primarily because, in England at least, class allocations can happen as late as September of a new academic year. If programmes require training or preparation prior to the start of the school year, this means that randomisation must take place in the year prior to implementation. What this ultimately means is that class lists used for randomisation purposes need updating (increasing the risk of mis- allocation/contamination if interventions begin in summer terms). If classroom-level randomisation is unavoidable, then schools will need more support due to the additional burden on schools/administrators to provide additional data.

What can arts-based education organisations learn from Learning about Culture (LAC)?

- The first point to make is that **these trials are not the final word regarding programme impact.** As set out in section 3, there are significant points that the research and funding community needs to address in order to ensure that assessments of impact are even better targeted. There is also a need for claims regarding the effectiveness of arts-based education interventions to be tempered by available evidence. The results of these studies, in particular feedback on implementation, mean those organisations have a basis for reflecting on how to improve delivery.
- **Arts-based education learning organisations *engaging* in trials and other robust research designs means that claims of effectiveness can be rigorously tested, often for the first time.** Alongside that, lessons can be learnt, both about the intervention and the ways that interventions might be researched. (There is a more general point about how funding is spent – because without testing we cannot say whether an approach has a positive or negative impact, or no impact at all.)[5]
- **There is a growing 'ecosystem' around evidence-based philanthropy that supports 'better giving.**[6] As many arts organisations are charities, being able to evidence effectiveness may offer an advantage in an increasingly competitive 'market' for donations. Where results are null or negative, this might be viewed as undermining, but the response is that at least those organisations now know whether or not their work affects a given set of outcomes – many other organisations cannot say that (even if they might claim it). Beyond these points, we discuss some specifics from our evaluations below.
- **Scaling requires planning from the point of piloting onwards.** Successful scaling does not follow automatically from an intervention that has positive effects – it requires intention, planning and delivery. There can be many reasons that successful interventions fail to scale, ranging from what is scaled not being what was tested ('voltage drop': Al-Ubaydli et al., 2017; 2019), the absence of necessary and/or sufficient conditions required for success (Cartwright & Hardie, 2012), or poor fidelity overall (see Sarama et al., 2008). What is clear from literature on scaling is that **planning for scale** early on is needed, acknowledging that successful scaling necessitates (or forces) alterations to interventions: typically simplification, codification, local 'credible' implementers who can run with programmes at a distance, specific implementation research, or strong alignment with existing policy that means changes are small enough to be implemented easily (see Horner et al., 2013; Smith et al., 2015; Hallsworth & Kirkman, 2020; Shenderovich et al., 2020; Bird et al., 2021).[7]

The points about scaling-up suggest the need for thinking through what it would take someone to deliver an intervention with no input from developers – perhaps more easily posed as questions:

a What resources do they need? Teaching materials, but also physical resources.
b What training? Is it one-off? Repeated?

---

[5] As an example of a well-regarded intervention that seems to 'make sense' that it would be effective, consider teaching children chess in primary school. A 2016 EEF evaluation led by UCL found that most teachers and children liked the introduction of chess teaching, but that teaching children chess had **no impact on maths attainment** a year later (Jerrim et al., 2016). This result contrasts sharply with other studies in the area – including two other RCTs – and it may be that because the research team measured outcomes a year after intervention, any impact had diminished. But the evidence sends a clear message: it would be wrong to claim that chess improved maths attainment among primary school children similar to those included in the study. That does not mean children should stop playing chess, or that chess offers no other benefits, such as learning rules, turn-taking and understanding strategy.

[6] See, for example, https://giving-evidence.com/ and https://www.chewgroup.org.uk/ for examples of evidence-based philanthropy and an organisation supporting charities with evaluation.

[7] There is also the idea of starting from the position of running experiments at scale (e.g., at a population level) – typically thought of as 'too difficult' – but these can be, and have been, done (e.g., Muralidharan & Niehaus, 2017; Nagin & Sampson, 2019).

c What support? What **people**? How many people? What skills?
d Will a set of instructions make sense on a wet Tuesday morning in February in a school very different from the places it was originally tested out?

Successful scaling also depends on how complex an intervention is. The more complex (i.e., the more 'moving parts'), the more planning, stages and structure needed (Greenhalgh & Papoutsi, 2019; Zamboni et al., 2019 report 30 or so points that scaling frameworks set out). If there are lots of moving parts, the challenge for scaling-up is working out what can be left out and the approach still be effective – often that question is left unanswered, or eligibility criteria are changed to encourage participation.

How organisations think about evaluation and participation in studies

As evaluators, we appreciate the courage and commitment from the arts-based education organisations who put themselves forward to fully participate in a year-long evaluation. It's also important to reflect on what the burden of participating in research brings for organisations and schools, and the issues that randomised trials, in particular, raise. We think there are three main takeaways:

a **Fear of losing 'control' in an RCT:** organisations have to accept that randomisation means they do not get to choose who gets offered support/an intervention (be that a school or pupil). That might feel very far from how they normally operate, but accepting and sticking to that is crucial for the scientific study of interventions.
b **Constraints of participating in RCT:** Pupil attrition poses challenges to balancing trial needs versus programme needs. For example, Speech Bubbles pupils moved schools and then developers felt they no longer had enough pupils to run sessions with fidelity. Ordinarily developers might just bring new students into groups, which may still be possible during a trial, but the new pupils will typically be excluded from analysis because they were not randomly allocated.
c **Data collection:** Using administrative data is a way to reduce the burden on schools and organisations, but what can be measured doesn't necessarily match the outcomes of interest. If funders, delivery organisations or (perhaps less likely) schools, think that other outcomes need collecting, then there needs to be a recognition of the burden this will impose on schools. (We return to this point in section 5.)

Challenges and successes: key takeaway points from implementation

From the IPE findings across the five interventions, we observed three main areas that appeared to affect implementation: intervention characteristics; individual-level factors; and structural factors.

Intervention characteristics

There are trade-offs between *specificity* (prescriptiveness), *fidelity*, *adaptability* and *scalability*. Generally, there is unlikely to be a 'unicorn' in the sense of an intervention that (i) is easy to implement, (ii) has high fidelity, and (iii) is effective. The programmes we evaluated can be loosely mapped on a spectrum of how tightly interventions were specified in terms of implementation in the classroom. To be clear, specificity (see Figure 2) is *not* a measure or indicator of programme quality, but it did have implications for implementation.

There were four key dimensions identified to programme specificity:

- **a** training;
- **b** external specialist involvement;
- **c** practical hands-on teaching;
- **d** programme fit with existing school approach.

For example, Speech Bubbles was highly specified in that external practitioners were leading intervention delivery with a specified role and training for the teaching assistant (TA) in each school, whereas Craft of Writing had little direct programme delivery by external specialists and relied upon training teachers to implement it (see Figure 2). As teachers were given autonomy in implementation, it was possible for them to adapt the programme to better sit alongside the school's existing approach to writing instruction.

Highly specified (prescriptive) interventions generally had higher levels of fidelity, often because the intervention was delivered by an external team, but also because they tend to have greater levels of detail regarding what needs to be delivered. However, this means the interventions may be less adaptable to a particular school context or constraint. Highly specified interventions – particularly if they require specialist staff – may be less scalable due to this specificity of implementation (lack of adaptability) and cost.

Less specified interventions are more adaptable to local context, but may require more effort from teachers. Greater flexibility and scope for adaptability are more likely to lead to less fidelity or, at the very least, mean that what is delivered may differ in appreciable ways between School A and School B. That said, differentiating between what are 'core components' and what are 'modular' / optional components is important, and thus gives greater agency (which may also help with acceptability) (see Blase & Fixsen, 2013; Aarons et al., 2017).

Individual-level factors

For all the interventions we studied, there were teachers and pupils who were more and less comfortable with them. We split these factors into school staff level and pupil level.

School staff-level factors

- **Teachers/TAs:** previous experience/background of the teacher; belief in the likely efficacy of the intervention; comfort with teaching in a way that was different from their usual practice;
- **SLT:** There is a need for support from a SLT key facilitator ('champion') to support successful implementation. We also found that support was stronger among SLT who had a background in the intervention topic (e.g., writing), or had perceived a positive impact on pupils.

Pupil-level factors

- **English as an additional language (EAL)** was a key consideration for some interventions.
- **Lower attainment pupils / SEN**[8]: Perceptions that some pupils struggled to engage (meaning that programmes might need further adaptation).
- **'Less engaged pupils':** One theme across the IPEs was the potential for these interventions to engage some pupils who are typically less engaged in the standard literacy lessons and activities. It may be useful to consider whether these sorts of creative intervention might be better targeted at pupils who are harder to engage.

[8] Note. These terms are conflated because teachers used them interchangeably.

Structural factors

Resources and timetabling were identified as the main structural challenges affecting implementation. Interventions placed different demands on schools in terms of the resources needed to take part. However, even where the interventions were not time and resource intensive, challenges to implementation were still reported.

Resources
In some cases, it was challenging for schools to make the required classroom space and equipment available, though this could be overcome with support from SLT (a finding linked to the *Individual-level factors* point above). As part of school recruitment, it may need to be clearer and more obvious up-front what schools will need to provide, to take part.

Timetabling
In terms of timetabling, some teachers reported finding it difficult to fit the interventions into what was seen as an already crowded timetable, with pressure to teach the school or national curriculum (e.g., balancing demands to teach grammar with time for 'free writing' as part of the Craft of Writing) or concrete timetabling conflicts (e.g., FTM First Thing Music required teaching music first thing in the morning). Again, support from the SLT was important in facilitating prioritisation of the intervention in the timetable.

Timetabling was specifically identified as an implementation barrier, for different reasons, across all five projects:

- **First Thing Music:** given as a reason for not completing delivery in the morning – and it is notable that interventions involving timetabling changes (rather than changes of content) can be very difficult to implement.
- **Speech Bubbles:** timetabling difficulties were a barrier for releasing the TA required for delivery.
- **Craft of Writing** and **Power of Pictures:** the need to work through curriculum content meant that sometimes that was prioritised ahead of intervention activities. (At times there was conflict between intervention content and national curriculum requirements, e.g., free-flow writing versus curriculum grammar requirements.)
- **Young Journalist Academy:** timetabling made it harder to prioritise YJA activities outside days when the mentor visited school.

Senior leadership team (SLT)
As with almost any new approach in schools, successful implementation was strongly influenced by the level of commitment of the SLT. This indicates that processes for ensuring that this commitment is in place *need to be enhanced for successful scale up.* This is especially the case given that SLT in schools that are later adopters are likely to be less committed and less enthusiastic than early adopters. *One suggestion for the EEF / delivery partners is to assess levels of commitment at SLT using anonymous feedback forms following presentations* – for example, if there is little agreement, or most do not feel engaged in the idea, this suggests implementing may be more difficult. Similarly, asking early on directly whether SLT members would be willing to 'champion' is valuable.

Conclusions and lessons for the future

In this report, we have brought together and reflected on carrying out five school-based randomised controlled trials of arts-based education interventions. We have particularly focused on lessons from our findings and the process of conducting this research for fellow evaluators, research funders, arts organisations delivering such interventions and schools.

Conclusions

Given the primary status of the analysis of pupil attainment within these five evaluations, we begin our conclusions by stating that they have not produced statistically significant evidence of positive impacts of these interventions on our academic attainment outcomes. But these trials should not be seen as the last word on the impact of these interventions and others like them. We knew when starting these evaluations that we would not be capturing all the effects of the interventions. Not finding effects on these specific measures of pupil academic attainment shines a light on the need to think more deeply about how we can measure (and hence quantitatively evaluate) what we value, not just value what we can easily measure.

Schools and education policymakers should very much interpret the findings in this light. Our findings do not provide evidence that using these programmes within schools gets in the way of pupils developing their core academic skills. In that sense, **it should be seen as encouraging to schools that engaging in appropriate arts-based programmes appears engaging for staff and pupils, including those who might otherwise be a cause for concern. In this respect, is a complementary activity to other parts of the curriculum.**

Perhaps our biggest conclusion, and the one we reflect on most below, is that **these trials have important lessons for future attempts to evaluate interventions in this space**. Building evidence for any educational intervention requires an accumulation of evidence first about how interventions affect their most direct intermediate outcomes, before moving to evaluate more distant outcomes (such as pupil attainment). This project shows the limitations of attempting to skip any of the steps that are implied in such an accumulation, including new and innovative thinking as to how to capture such intermediate outcomes at fairly substantial scale (allowing for their use in impact, not just implementation and process evaluations) in the highly time-pressured context of a busy school.

There are pointers for arts-based education organisations on how to refine their programmes to help ensure implementation within schools is as smooth as possible. This includes ensuring adaptability to differing contexts within different schools, engagement with school staff at multiple levels (senior leadership and classroom teachers, for example, not one or the other), and planning to be flexible within school structures such as timetabling (rather than asking schools to work around the programmes in this respect).

Reflections and lessons for the future

As stated above, we cannot shy away from the fact that the trials as a whole have not produced statistically significant quantitative evidence implying improvements in pupils' academic attainment (in the forms that we measured) relative to business-as-usual control groups who did not receive these interventions. However, while we did not find evidence from the trials that these interventions shifted our outcomes of interest faster than they were moving in the control group, equally we note that **there is little evidence that schools using these programmes, on average, get in the way of pupils' academic development in these forms**. Moreover, we stress that the outcomes we measured are far from the be-all and end-all of pupils' educational experience. Indeed, the implementation and process evaluations revealed substantial enthusiasm for the programmes among teachers and pupils, as well as suggestions that these interventions may have particular benefits in encouraging engagement among children who are otherwise at risk of disengagement in

the classroom. Ultimately, we return to the point made at the outset: the best argument for arts-based education is for its own sake, not for its instrumental value.

Perhaps the largest issue that the lack of impacts on quantitative outcomes raises, despite the evident enthusiasm evident from IPE findings, is **whether the evidence base for these interventions was sufficiently well developed to carry out randomised controlled trials aiming to estimate effects on pupil *attainment* outcomes,** particularly given the relative lack of quantitative evidence on intermediate outcomes within the interventions' theories of change. If our five studies had focused first and foremost on estimating treatment effects on more intermediate outcomes implied by the interventions' logic models (e.g., the change at teacher level in the case of the Craft of Writing programme), we may ultimately have advanced our understanding of the potential impacts of such interventions further, placing us in a position of understanding whether these interventions show unambiguous effects on these intermediate outcomes or whether these are not being achieved, for example, due to some of the implementation challenges that we have identified. Instead, we are in a position of uncertainty as to whether the weakness in these interventions' actual model of changing pupil attainment lies between intervention delivery and intermediate outcomes, between those intermediate outcomes and academic attainment, or around theory / implementation failure.

A related challenge for evaluators is that we are limited by our methods for collecting different outcomes, particularly at scale. The reliance on traditional survey methodology that is more formalised may prevent us from capturing important, but non-attainment, outcomes that sit on pathways **to** attainment, or to outcomes that are important for pupils more generally (such as engagement in learning). There are stand-out examples of where technology is being used to collect reliable and valid measures of pupil outcomes, such as the Automated Test of Language Abilities (ATLAS),[9] which has been developed by a team at the University of Oxford. That app allows for at-scale data collection by teachers or TAs. COVID-19 presents a 'reset' opportunity to think about how we collect data, but app development isn't something that can/should be done in the midst of a trial (the development of ATLAS was funded by a grant from the Nuffield Foundation). This implies that if EEF (and/or other research funders, perhaps including Nesta and the Nuffield Foundation) intend to fund further work analysing the impact of interventions with this level of prior evidence, there is a real need for them, explicitly, to first fund work on the development of scalable data collection approaches, such as apps, that are useful to schools but also light in terms of burden. These could be specialist, such as ATLAS, or more general platforms that can be adapted easily. If forced to choose, we would recommend developing high-quality apps that are narrower in focus, so that the time can be spent making sure the information is useful to teachers and schools.

Ultimately, we stress that our findings provide plenty of reasons to remain optimistic about the importance of arts-based educational interventions as part of a balanced curriculum, helping pupils to develop skills and interests, whether or not they can ever be fully captured in measures of academic attainment. They do this while providing plenty of sources of challenge for the research and arts-based education communities, which we hope all sides will step up to move forward in the coming years.

[9] See the project webpage for more details (http://www.education.ox.ac.uk/research/standardisation-of-a-computerised-oral-language-test/).

References

Aarons, G.A., Sklar, M., Mustanski, B., Benbow, N., & Brown, C.H., 2017. 'Scaling-out' evidence-based interventions to new populations or new health care delivery systems. *Implementation Sci* 12, 111. https://doi.org/10.1186/s13012-017-0640-6

Ahn, S., Ames, A.J., & Myers, N.D., 2012. A review of meta-analyses in education: Methodological strengths and weaknesses. *Review of Educational Research* 82, 436–476. https://doi.org/10.3102/0034654312458162

Al-Ubaydli, O., List, J.A., LoRe, D., & Suskind, D., 2017. Scaling for economists: Lessons from the non-adherence problem in the medical literature. *Journal of Economic Perspectives* 31, 125–144.

Al-Ubaydli, O., List, J.A., & Suskind, D., 2019. *The science of using science: Towards an understanding of the threats to scaling experiments.* National Bureau for Economic Research Working Paper.

Barnes, J., 2014. Drama to promote social and personal well-being in six- and seven-year-olds with communication difficulties: the Speech Bubbles project. *Perspect Public Health* 134, 101–109. https://doi.org/10.1177/1757913912469486

Bereczki, E.O., & Kárpáti, A., 2018. Teachers' beliefs about creativity and its nurture: A systematic review of the recent research literature. *Educational Research Review* 23, 25–56. https://doi.org/10.1016/j.edurev.2017.10.003

Bird, K.A., Castleman, B.L., Denning, J.T., Goodman, J., Lamberton, C., & Rosinger, K.O., 2021. Nudging at scale: Experimental evidence from FAFSA completion campaigns. *Journal of Economic Behavior and Organization* 183, 105–128.

Blase, K., & Fixsen, D., 2013. *Core intervention components: Identifying and operationalizing what makes programs work*. U.S. Department of Health & Human Services.

Bruning, R., Dempsey, M., Kauffman, D.F., McKim, C., & Zumbrunn, S., 2013. Examining dimensions of self-efficacy for writing. *Journal of Educational Psychology* 105, 25.

Cartwright, N., & Hardie, J., 2012. *Evidence-based policy: A practical guide to doing it better*. Oxford University Press, Oxford/New York.

Catterall, J., 2009*. Doing well and doing good by doing art: The effects of education in the visual and performing arts on the achievements and values of young*. Los Angeles/London: Imagination Group/IGroup Books.

Catterall, J.S., Dumais, S.A., & Hampden-Thompson, G.,2012. *The arts and achievement in at-risk youth: Findings from four longitudinal studies*. Research Report# 55. National Endowment for the Arts. [online] https://www.researchgate.net/publication/263479321_The_Arts_and_Achievement_in_At-Risk_Youth_Findings_from_Four_Longitudinal_Studies.

Collins, K., 2015. *Why Arts Education Matters*, Education Endowment Foundation blog. https://educationendowmentfoundation.org.uk/news/why-arts-education-matters

Cooper, B. 2018. *Primary colours: The decline of arts education in primary schools and how it can be reversed*. Fabian Policy Report, Fabian Society. https://fabians.org.uk/wp-content/uploads/2019/01/FS-Primary-Colours-Report-WEB-FINAL.pdf

Deane, P., Odendahl, N., Quinlan, T., Fowles, M., Welsh, C., & Bivens-Tatum, J., 2008*. Cognitive models of writing: Writing proficiency as a complex integrated skill*. ETS Research Report Series 2008, i–36. https://doi.org/10.1002/j.2333-8504.2008.tb02141.x

Dunsmuir, S., Kyriacou, M., Batuwitage, S., Hinson, E., Ingram, V., & O'Sullivan, S., 2015. An evaluation of the Writing Assessment Measure (WAM) for children's narrative writing. *Assessing Writing* 23, 1–18.

Eckhoff, A., & Urbach, J., 2008. Understanding imaginative thinking during childhood: Sociocultural conceptions of creativity and imaginative thought. *Early Childhood Educ J* 36, 179–185. https://doi.org/10.1007/s10643-008-0261-4

EEF, 2018. *Arts participation, Education Endowment Foundation evidence toolkit*. https://educationendowmentfoundation.org.uk/evidence-summaries/teaching-learning-toolkit/arts-participation/

Glăveanu, V.P., 2015. Creativity as a sociocultural act. *J Creat Behav* 49, 165–180. https://doi.org/10.1002/jocb.94

Graham, S., & Hebert, M., 2011. Writing to Read: A meta-analysis of the impact of writing and writing instruction on reading. *Harvard Educational Review* 81, 710–744. https://doi.org/10.17763/haer.81.4.t2k0m13756113566

Graham, S., & Sandmel, K., 2011. The process writing approach: A meta-analysis. *The Journal of Educational Research* 104, 396–407. https://doi.org/10.1080/00220671.2010.488703

Greenhalgh, T., & Papoutsi, C., 2019. Spreading and scaling up innovation and improvement. *BMJ* l2068. https://doi.org/10.1136/bmj.l2068

Hallsworth, M., & Kirkman, E., 2020. *Behavioral insights*. MIT Press.

Horner, R.H., Kincaid, D., Sugai, G., Lewis, T., Eber, L., Barrett, S., Dickey, C.R., Richter, M., Sullivan, E., Boezio, C., Algozzine, B., Reynolds, H., & Johnson, N., 2014. Scaling up school-wide positive behavioral interventions and supports: Experiences of seven states with documented success. *Journal of Positive Behavior Interventions* 16, 197–208. https://doi.org/10.1177/1098300713503685

Jerrim, J., Macmillan, L., Micklewright, J., Sawtell, M., & Wiggins, M., 2016. *Chess in schools. Evaluation report and executive summary*. Education Endowment Foundation, London.

Kaufman, J.C., 2019. Self-assessments of creativity: Not ideal, but better than you think. *Psychology of Aesthetics, Creativity, and the Arts* 13, 187–192. https://doi.org/10.1037/aca0000217

Kaufman, J.C., & Beghetto, R.A., 2009. Beyond big and little: The four C model of creativity. *Review of General Psychology* 13, 1–12. https://doi.org/10.1037/a0013688

Kautz, T., et al. 2014. *Fostering and measuring skills: Improving cognitive and non-cognitive skills to promote lifetime success*, OECD. Available at: www.oecd.org/edu/ceri/Fostering-and-Measuring-Skills-Improving-Cognitive-and-Non-Cognitive-Skills-to-Promote-Lifetime-Success.pdf

Kim, K.H., 2011. The creativity crisis: The decrease in creative thinking scores on the Torrance tests of creative thinking. *Creativity Research Journal* 23, 285–295. https://doi.org/10.1080/10400419.2011.627805

Londesborough, M., Partridge, L., Bath, N., & Grinsted, S., 2017. *Learning about culture: Programme prospectus*. RSA. https://www.thersa.org/globalassets/pdfs/reports/rsa-learning-about-culture-report.pdf

Moffitt, T.E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R.J., Harrington, H., Houts, R., Poulton, R., Roberts, B.W., Ross, S., Sears, M.R., Thomson, W.M., & Caspi, A., 2011. A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences* 108, 2693–2698. https://doi.org/10.1073/pnas.1010076108

Moors, T., Silva, S., Maraschin, D., Young, D., Quinn, J., de Carpentier, J., Allouche, J. & Himonides, E. (2020). Using beatboxing for creative rehabilitation after laryngectomy: Experiences from a public engagement project. *Frontiers in Psychology* 10, 2854. https://doi.org/10.3389/fpsyg.2019.02854

Muralidharan, K., & Niehaus, P., 2017. Experimentation at scale. *Journal of Economic Perspectives* 31, 103–124. https://doi.org/10.1257/jep.31.4.103

Murphy, V.A., Kyriacou, M., & Menon, P., 2013. *Profiling writing challenges in children with English as an Additional Language (EAL)*. University of Oxford, Department of Education 14.

Nagin, D.S., & Sampson, R.J., 2019. The real gold standard: measuring counterfactual worlds that matter most to social science and policy. *Annual Review of Criminology* 2, 123–145.

Neelands, J., Belfiore, E., Firth, C., Hart, N., Perrin, L., Brock, S., Holdaway, D., & Woddis, J., 2015. *Enriching Britain culture, creativity and growth: 2015 report by the Warwick Commission on the future of cultural value*. University of Warwick, Coventry.

Newman, M., Bird, K., Tripney, J., Kalra, N., Kwan, I., Bangpan, M., & Vigurs, C., 2010. *Understanding the impact of engagement in culture and sport: A systematic review of the learning impacts for young people*. CASE: The culture and sport evidence programme.

Sala, G., & Gobet, F., 2020. Cognitive and academic benefits of music training with children: A multilevel meta-analysis. *Mem Cogn* 48, 1429–1441. https://doi.org/10.3758/s13421-020-01060-2

Sarama, J., Clements, D.H., Starkey, P., Klein, A., & Wakeley, A., 2008. Scaling up the implementation of a pre-kindergarten mathematics curriculum: Teaching for understanding with trajectories and technologies. *Journal of Research on Educational Effectiveness* 1, 89–119. https://doi.org/10.1080/19345740801941332

Sarma, U.A., & Thomas, M.T., 2020. Breaking the limits of executive functions: Towards a sociocultural perspective. *Culture and Psychology* 26, 358–368. https://doi.org/10.1177/1354067X19898673

See, B. H. & Kokotsaki, D., 2015. *Impact of arts education on the cognitive and non-cognitive outcomes of school-aged children: A review of evidence*. Education Endowment Foundation, London. https://educationendowmentfoundation.org.uk/public/files/Presentations/Publications/Arts_Education_Review.pdf

Shenderovich, Y., Ward, C.L., Lachman, J.M., Wessels, I., Sacolo-Gwebu, H., Okop, K., Oliver, D., Ngcobo, L.L., Tomlinson, M., Fang, Z., Janowski, R., Hutchings, J., Gardner, F., & Cluver, L., 2020. Evaluating the dissemination and scale-up of two evidence-based parenting interventions to reduce violence against children: study protocol. *Implement Sci Commun* 1, 109. https://doi.org/10.1186/s43058-020-00086-6

Simoiu, C., Sumanth, C., Mysore, A., & Goel, S., 2019. Studying the 'wisdom of crowds' at scale, in: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 171–179.

Smith, J.M., de Graft-Johnson, J., Zyaee, P., Ricca, J., & Fullerton, J., 2015. Scaling up high-impact interventions: How is it done? *International Journal of Gynecology and Obstetrics* 130, S4–S10. https://doi.org/10.1016/j.ijgo.2015.03.010

Wyse, D., 2017. *How writing works: From the invention of the alphabet to the rise of social media*. Cambridge University Press.

Wyse, D., & Torgerson, C., 2017. Experimental trials and 'what works?' in education: The case of grammar for writing. *Br Educ Res J* 43, 1019–1047. https://doi.org/10.1002/berj.3315

Zamboni, K., Schellenberg, J., Hanson, C., Betran, A.P., & Dumont, A., 2019. Assessing scalability of an intervention: why, how and who? *Health Policy and Planning* 34, 544–552. https://doi.org/10.1093/heapol/czz068

Appendix A: Pooled impact evaluation

As an additional analysis, we ran a simple pooled analysis of the impact estimates from across the three KS2 trials. This decision was taken only in this case because these trials have the same design (two-armed, clustered trial with randomisation at the school level), the same broad target population (Year 5 pupils in English state schools), and the same outcome measures: the WAM, the WSEM and the ideation measure.

The goal of this additional analysis is to estimate the impact of participation in any of the three cultural learning activities evaluated in later primary school (Key Stage 2 level: CoW, the Power of Pictures and the Young Journalist Academy) on our outcome measures of interest. We think the pooled analysis approach is a useful thing to do, despite being for a rather ill-defined treatment (an averaging of all three of these interventions – note that it also says nothing about using the interventions in combination) given the commonalities of the interventions (i.e., arts-based education methods with writing attainment as primary outcome and implemented in Year 5), in order to provide an additional source of evidence as to whether interventions such as these may be having impacts on our outcomes of interest that our individual trials may be missing due to lack of statistical power. Specifically, we are trading off treatment specificity with the additional sample size and, hence, power provided by pooling our samples.

We estimate our pooled models on the combined analysis samples from the three trials, with the regression models that we estimate taking the form:

$$Y_{ij} = \alpha + \beta_1 Treat_j + \beta_2 FSM_{ij} + \beta_3 EAL_{ij} + \beta_4 FSMProp_{ij} + \beta_5 EALProp_{ij} + \gamma' X_{jt} + \varepsilon_{ij}$$

where Y_{ij} is our outcome of interest (either the WAM, WSEM, or ideation measure) for individual i in school j, FSM_{ij} is whether individual i is eligible for free school meals (FSM) and EAL_{ij} is whether individual i is recorded as having English as an additional language (EAL), while $FSMProp_{ij}$is the FSM composition of treated class in school j and $EALProp_{ij}$is the same for its EAL composition. X_{jt}includes a series of trial t specific binary variables for each strata used in the randomisation. The inclusion of a full set of trial-specific stratum variables allows us to account for structural differences between the trials while still estimating our parameters of interest in pooled models.

This regression model is estimated separately for our primary outcome (the WAM) and then the two secondary outcome measures (WSEM and ideation measures).

Pooled intention-to-treat estimates are recovered from the estimate of β_1 when it is estimated on the pooled samples as they were at randomisation. As with the individual trials, our final analysis samples are affected by attrition, so are not truly intention to treat in this sense, but are the closest approximations with the data available.

This document is available for download at https://educationendowmentfoundation.org.uk

Education Endowment Foundation

The Education Endowment Foundation
5th Floor, Millbank Tower
21–24 Millbank
London
SW1P 4QP

https://educationendowmentfoundation.org.uk
@EducEndowFoundn
Facebook.com/EducEndowFoundn