First Thing Music

Behavioural Insights Team and UCL Institute of Education
Nikki Shure, Jake Anders, Dominic Wyse, Kimberly Bohling, Pantelis Solomon, Matt Barnard

PROTOCOL UPDATES 25 APRIL 2019:

p. 2 – update evaluation summary table

p. 8 – collection of Social Skills Improvement System data modified

p. 13-15 – modification to Implementation and Process Evaluation (IPE) data collection and analytical approach

p. 17 – updated evaluation team

p. 26 – addition of Appendix 2 – Additional detail about the IPE

PROTOCOL UPDATES 23 JANUARY 2019:

p. 7 – additional information added on Social Skills subscales, reference added as footnote

p. 8 – clarification on raw scores added, aggregate score updated, creative self-efficacy as secondary outcome clarified, EYFSP aggregate scores clarified,

p. 9 – details on creative self-efficacy added

p. 16 – ISRCTN number added

Evaluation Summary	
Age range	Year 1 (ages 5-6)
Number of pupils	3,058 (at randomisation)
Number of schools	65 schools (123 classes) (at randomisation)
Design	Cluster Randomised Controlled Trial (Classroom level randomisation)
Primary Outcome	Reading attainment
Date	24 April 2019
Version	3

Intervention

The First Thing Music programme aims to improve children's reading and social skills by providing them with daily music sessions. The sessions are part of a structured music education programme based on the Kodály approach. Students will learn the basics of music through daily singing and musical games with teachers who will be trained by music practitioners. Music education has been linked to improvements in academic attainment in areas such as literacy and language.[1] First Thing Music is part of a broader programme of work entitled 'Learning About Culture', which aims to improve the evidence base around arts-based education programmes. This is coordinated by the Education Endowment Foundation and the Royal Society for the Encouragement of Arts, Manufactures and Commerce.[2] It consists of five programmes: two in Key Stage 1 and three in Key Stage 2.

The model that will be tested in this programme is comprised of daily 15-minute music sessions for Year 1 pupils (5-6 year olds) over the course of three terms. The children will take part in singing games and movement activities, focusing mainly on steady beat, rhythm and pitch. These concepts will be introduced subconsciously at first, and later made conscious for the children as they are introduced to the basics of music theory and notation. Ideally, the music session will take place at the beginning of the school day as a "carpet" session or in a hall space. The intervention will be delivered by class teachers who will receive training and mentoring from a team of music practitioners recruited by specialists from the British Kodály Academy. Teachers will be asked to attend a day training session in September. In addition six half day sessions (one per half term) will be provided for teachers and music practitioners aimed at consolidating and expanding learning, and to give opportunities for feedback and sharing of experiences.

The intervention itself will commence with one week of sessions every day led by the music

[1] Henley, D. (2011). *Music Education in England*. Department for Education. Retrieved from: http://webarchive.nationalarchives.gov.uk/20110305040317/http://education.gov.uk/publications/eOrderingDownload/Music%20Education%20in%20England%20-%20A-Review.pdf

[2] https://www.thersa.org/globalassets/pdfs/reports/rsa-learning-about-culture-report.pdf

practitioners who will provide training for the teachers and an opportunity for them to participate in the intervention. The teacher will lead the sessions independently for the second week. Specialists will return in week three and from then on will make weekly visits to support the teachers. Two after-school training sessions will be provided per term. Headteachers are encouraged to attend and participate in at least one in-class session per term. The teacher will also be supported by a Resource Booklet, including an introduction to the Kodály-based approach, and songs/rhymes and games, both in visual and recorded format.

The daily music sessions will follow a Kodály-based approach, introducing the children to musical concepts through experiences such as listening, singing and movement. These concepts will be constantly reviewed, reinforced and extended through games, songs and exercises. More advanced skills such as written notation, subdivisions of beats, patterns of longer/shorter sounds and awareness of melodic shape by pitch discrimination, will be introduced gradually according to the progress of the children.

The First Thing Music programme will be delivered by the First Thing Music team led by Lindsay Ibbotson with the delivery partners Tees Valley Music Service and tutors from the British Kodály Academy, who will provide ongoing support and supervision to each participating school.

Further information on the intervention theory of change and logic model is provided in the Logic Model section.

Significance

A growing body of research supports the argument that difficulty in processing rhythm can have a negative impact on children's reading behaviour. Variations in attainment in spelling and reading have been found to be related to performance on tests of rhythmic discrimination.[3] Some studies evaluating the Kodály approach with young children found that it led to significant gains in mathematics and reading attainment[4] while others found that improvements could only be seen among boys.[5] The EEF's Arts Education literature review identified the Kodály approach as promising, based on earlier studies. More broadly, the review found encouraging results for musical approaches with young children.[6]

This evaluation is part of a round of funding between the Education Endowment Foundation (EEF) and the Royal Society of Arts to test the impact of different cultural learning strategies in English schools. The programmes will be supported by Arts Council England.

---

[3] David, D., Wade-Woolley, L., Kirby, J. R., & Smithrim, K. (2007). Rhythm and reading development in school-age children: a longitudinal study. *Journal of Research in Reading*, *30*(2), 169-183.
[4] Gardiner, M. F., Fox, A., Knowles, F., & Jeffrey, D. (1996). Learning improved by arts training. *Nature*.
[5] Hurwitz, I., Wolff, P. H., Bortnick, B. D., & Kokas, K. (1975). Nonmusical effects of the Kodály music curriculum in primary grade children. Journal of learning Disabilities, 8(3), 167-174.
[6] Huat See, B., & Kokotsaki, D. (2015). *Impact of arts education on the cognitive and non-cognitive outcomes of school-aged children.* London, United Kingdom: Education Endowment Foundation. Retrieved from: https://v1.educationendowmentfoundation.org.uk/uploads/pdf/Arts_Education_Review.pdf.

Methods

Research questions

The primary objective of this evaluation is to test whether First Thing Music improves reading attainment over the course of one school year for Year 1 pupils.

The evaluation will also address whether the programme impacts pupils' social skills.

Design

This will be a cluster randomised controlled trial, randomised at the class level. The school will serve as the primary stratifying variable. This means that within each school at least one class will be assigned to receive the intervention (treatment) and one to the control condition.[7] Classes in the control group will be expected to continue with 'business as usual', and will be offered the opportunity to take part in the programme following the completion of the study.

Randomisation

Randomisation will be conducted following the recruitment of schools, including the signing of Memoranda of Understanding (MoUs) and the completion of parental withdrawal forms and data collection processes. Random allocation will occur at the class level. The randomisation will proceed in the following steps:

1. Classes will be stratified by school.
2. A random number will be generated for each class within each school.
3. In the case of a two-form school, the class with the highest random number in the school will be assigned to the treatment group, and the other class to the control group.
4. In the case of schools with more than two forms, BIT will do as per step 3 for schools with an even number of forms (e.g. four forms). For schools with an odd number of forms (e.g. three forms), the randomisation will be done as follows. Half the schools will randomly be chosen to have one treatment and two control classes and the other half will have two treatment and one control class.
5. If one-form entry schools are recruited, they will be grouped into a single stratum for the purposes of randomisation.

Randomisation will be conducted by BIT staff using data analysis and statistical software Stata. The code used to carry out this randomisation will be recorded and reported in the

[7] In the event the project delivery team are not able to recruit the required number of classrooms using two- or three-form entry schools, one-form entry schools will be accepted into the trial. These will be pooled into a single strata for the purposes of randomisation.

final report.

Participants

Schools will be recruited into the study on the basis that they:

- Are located in the North East region of England. If there is significant interest from localised areas not in the North East (e.g. Sheffield), music practitioners may be recruited to these areas as well. This will typically require at least 10 schools in a localised area to be recruited in the study.
- Have discussed participation with First Thing Music and signed an MoU detailing the conditions of participation (opt-out process, pupil data provision and endline assessment, participation in IPE activities, etc.).

Strong preference will also be given to two-form entry schools. If the project team are unable to recruit the required number of two-form schools, other schools (e.g. single and three-form entry) will be considered after discussions with the evaluation team and the EEF.

Schools with an average or above average share (14.1%[8]) of Free School Meal (FSM) children will receive priority in recruitment.

As this trial will be delivered to the whole class, there are no pupil-level eligibility requirements.

Sample size calculations

Sample size calculations are based on the following assumptions and with reference to the primary outcome measure (reading attainment):

- **Randomisation will be at the classroom level, stratified by school.** This will be performed as specified in the Randomisation section.
- **There will be two trial arms (treatment and control).**
- **The intra-cluster correlation (ICC) is estimated to be 0.19.** Estimating ICC values for class-level randomisations is difficult as there is less guidance available relative to school-level randomisation. Other EEF trials that used class-level randomisation have found the estimated ICC values when performing sample size calculations were overly optimistic.[9,10] While a school-level ICC value for a reading outcome measure in

[8] www.gov.uk/government/uploads/system/uploads/attachment_data/file/650547/SFR28_2017_Main_Text.pdf

[9] Foreign Language Learning in Primary Schools, a trial testing an intervention on English literacy involving Year 3 and 4 children, estimated an ICC of 0.05 when performing power calculations, but found it to be 0.13 when post-hoc analysis was performed (see pg 23). https://educationendowmentfoundation.org.uk/public/files/EEF_Project_Report_FLL.pdf

[10] Grammar For Writing, a trial testing an intervention on writing involving Year 6 children, estimated an ICC of 0.19 when performing power calculations, but found it to be 0.32 when post-hoc analysis was performed (see pg 26). https://educationendowmentfoundation.org.uk/public/files/Support/Campaigns/Evaluation_Reports/EEF_Project_Report_GrammarForWriting.pdf

KS1 would be estimated at 0.11 for schools in the North East[11], we adjust this upwards to 0.19 to provide a margin of error commensurate with the experiences of prior EEF studies.[12]

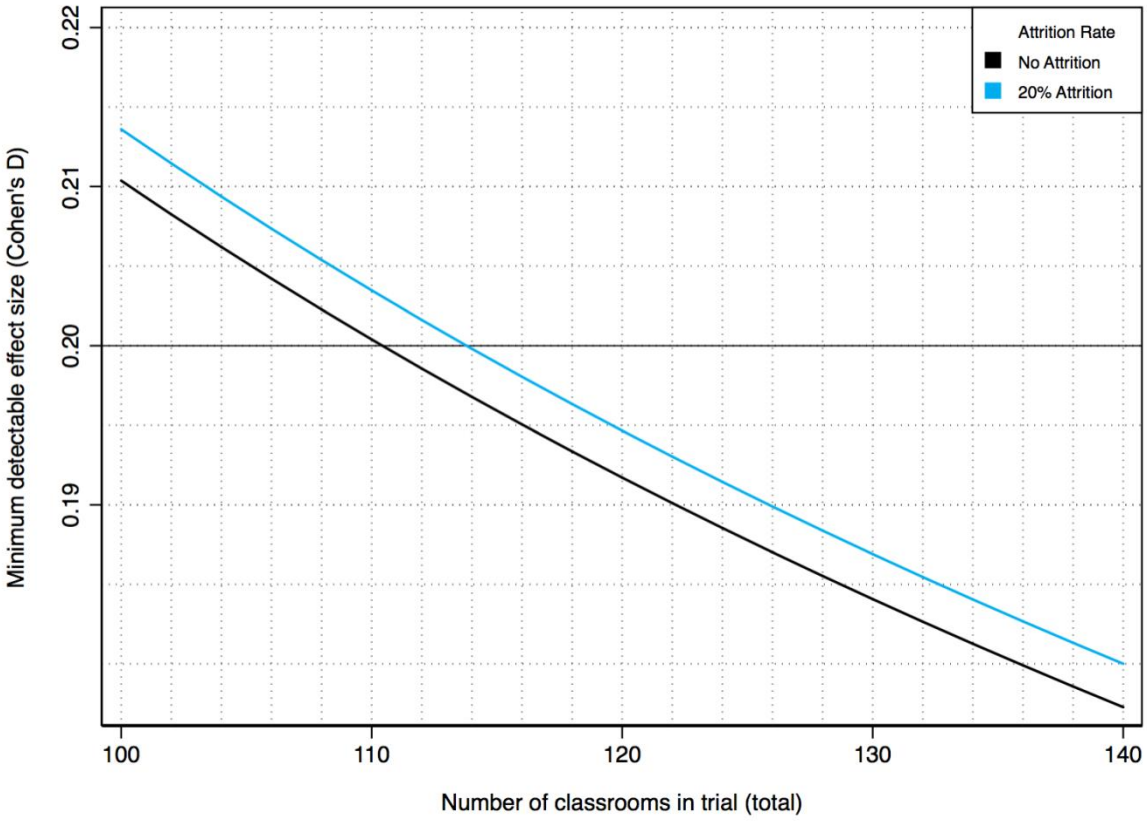
- **There are an average of 27.4 pupils per class using ONS statistics from 2016.[13]**
- **20 per cent of children in each school will withdraw their data or be unable to participate in the collection of an endline outcome measure (attrition due to changing school, inability to complete assessment etc.).** This estimate is based on the 15% standard post-randomisation attrition rate in EEF studies, plus an additional allowance for children being opted-out of the study (5%).
- Hypothesis
    - **Null hypothesis:** There is no difference in standardised PIRA scores between children who participate in the First Thing Music intervention and those who do not.
    - **Alternative hypothesis:** There is a difference in standardised PIRA scores between children who participate in the First Thing Music intervention and those who do not.
- **The required minimum detectable effect size (MDES) is 0.20 standard deviations (Cohen's d).** This specifies the minimum effect size our trial is powered to detect, in terms of a given standardised difference between two means (of a continuous outcome measure).
- **Test-retest correlation of 0.61.** As we will use Early Years Foundation Stage Profile (EYFSP) scores as a baseline when analysing our primary outcome measure, the predictive power of this baseline will also factor into our sample size calculations. We estimate this value using unpublished Fisher Family Trust (FFT) analysis of the test-retest correlation coefficient of EYFSP score and PIRA assessments collected at the end of year 1 for a prior EEF trial (ABRA: Online Reading Support).[14]
- **Power:** 80%; Significance level: 5%. These are standard assumptions in social policy trials.

[11] https://educationendowmentfoundation.org.uk/public/files/Evaluation/Writing_a_Protocol/ICC_2015.pdf

[12] This was estimated by inflating the school level ICC value expected by 70%, as per change between expected and observed ICC values in the Grammar For Writing trial.

[13] www.gov.uk/government/uploads/system/uploads/attachment_data/file/552342/SFR20_2016_Main_Text.pdf

[14] https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/EEF_Project_Report_ABRA.pdf



Given the above assumptions, including the 20 percent attrition assumption, we would require 112 classrooms (for a total of 3,069 children) to detect an effect of 0.202 (Cohen's d).

Assuming the FSM subgroup is 25 per cent of the total sample (based on data from DfE statistics[15] for Middlesbrough), and maintaining all other assumptions (which is likely to be a conservative approach, given lower levels of within-group variation in this subgroup), there is an estimated minimum detectable effect size for this group of approximately 0.23 standard deviations.

Outcome Measures

The primary outcome measure is reading attainment, with social skills as a secondary measure.

Reading

To measure the primary outcome, we will use the Progress in Reading Assessment (PIRA) by Rising Stars[16]. PIRA is a standardised assessment of pupils' reading attainment and profile of reading skills. It measures reading ability in the following areas: phonics, literal comprehension, and reading for meaning. This is a standardised and well-known test, which

[15] https://www.gov.uk/government/statistics/schools-pupils-and-their-characteristics-january-2017
[16] https://www.risingstars-uk.com/Series/Rising-Stars-Pira-Tests

has been used in a number of prior EEF evaluations.[17,18]

Endline PIRA assessments will be conducted during May - June 2019 by trained research assistants (RAs) who will be blind to trial arm assignment. Rising Stars, the publisher of PIRA, will mark the assessments. Analysis will use raw PIRA total scores (0-50).

Social Skills

Our secondary outcome measure will be social skills, which will be assessed at endline using the Social Skills scale of the Social Skills Improvement System (SSiS)[19]. The SSiS contains three scales: the aforementioned Social Skills scale, a Problem Behaviours scale and an Academic Competence scale. As the intervention logic model most supports detecting a change in social skill, we will not administer the other two scales. The SSiS Social Skills scale assesses pupils' skills across the following subscales: communication, cooperation, assertion, responsibility, empathy, engagement and self-control.

SSiS is the most commonly used social skills assessment for young children, and it is standardised and has been used in prior EEF evaluations.[20] We chose to use SSiS, over an equally popular instrument, the Strengths and Difficulties Questionnaire (SDQ) because it is more thorough and in-depth than SDQ. We also felt that the SSiS was a better tool to assess the types of behaviours we would expect to change as a result of participating in the programme (e.g., communication, engagement), whereas the SDQ was more oriented to identify problematic behaviours (e.g., conduct problems, peer relationship problems). The questionnaires will be delivered to teachers electronically. As with all measures of social skills at this age, this must be completed by the child's teacher and thus cannot be blind to trial arm assignment. The Social Skills subscale consists of 46 items, each scored 0-3. The analysis will use raw SSiS total scores (0-138).

We estimate that completing a 46-item survey for 30 students may take teachers 2.5 hours. We have concerns that teachers may not complete the survey for all students in their class and will instead complete it for a non-random sub-sample (e.g., the students with the best behaviour), which would bias our results. Instead, we will randomly select a sub-sample of 10 students per classroom for which each teacher will be asked to complete the survey. We believe asking teachers to complete the survey for a relatively small sub-sample of their students raises the chance that they will complete all requested surveys and – given that we select students randomly – provides better assurance that the data will not be biased.

---

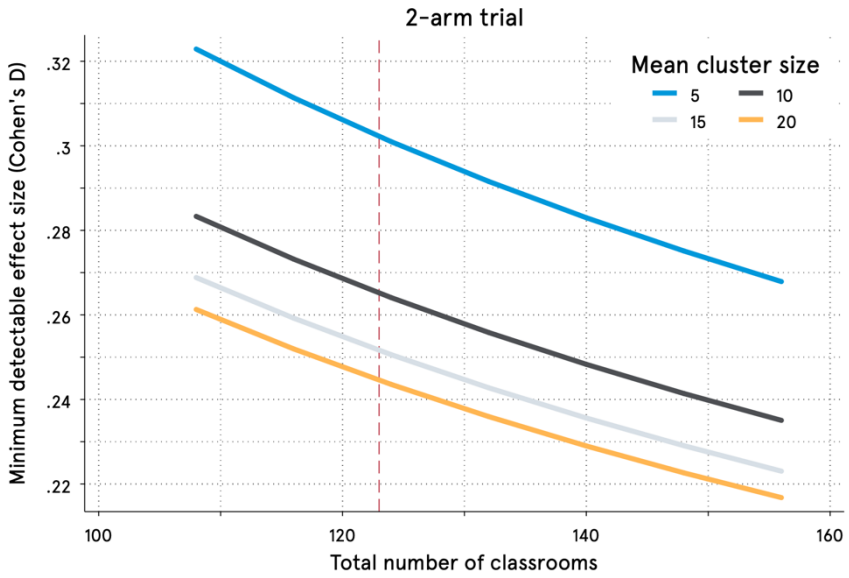
[17] McNally, S. (2016). *Evaluation Protocol: An Evaluation of Teaching Assistant-Based Small Group Support for Literacy*. London, United Kingdom: Education Endowment Foundation. Retrieved from https://v1.educationendowmentfoundation.org.uk/uploads/pdf/Digital_-_Small_Group_Support_for_Literacy.pdf.

[18] McNally, S., Ruiz-Valenzuela, J., & Rolfe, H. (2016). *ABRA: Online Reading Support.* London, United Kingdom: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/EEF_Project_Report_ABRA.pdf

[19] https://www.pearsonclinical.com/education/products/100000322/social-skills-improvement-system-ssis-rating-scales.html

[20] Centre for Effective Education, Queen's University Belfast. (2016). *Evaluation Protocol: Zippy's Friends.* London, United Kingdom: Education Endowment Foundation. Retrieved from: https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Protocols/EEF_Project_Protocol_Character_Zippys_Friends_protocol.pdf.

The proposed sub-sample size of 10 is based on power calculations. The figure below shows the minimum detectable effect sizes (MDESs) for sub-sample sizes of 5, 10, 15, and 20. We selected a sub-sample size of 10, as we think it strikes the best balance between power and a reasonable ask of teachers that is more likely to lead to full compliance of survey completion.



Creative self-efficacy

As highlighted in the logic model, the impact of the intervention on writing outcomes may have an effect through pupils' engagement with and motivation for reading. For this reason, we consider pupils' self-perception of ability to generate and use ideas in their school work as a secondary outcome measure. To measure this, we will use an adapted version of the ideation sub-measure of the writing self-efficacy measure proposed by Bruning et al. (2013), with significant simplification of language to make it appropriate for this age group (the original measure was designed for secondary school pupils).[21] This approach has been taken to provide some scope for comparisons with other trials being conducted at the same time (evaluation of Young Journalist Academy, Power of Pictures and Craft of Writing) in which we will also examine this subscale as part of the wider measure of writing self-efficacy. This measure will be captured using three, three-category likert scale items asked by RAs after completion of the PIRA assessment. The analysis will use raw total scores (3-9).

Analysis plan

Primary analysis

We will estimate the effect of the trial using a linear model on pupil-level data with class-level

[21] Bruning, R., Dempsey, M., Kauffman, D., McKim, C. & Zumbrunn, S. (2013) Examining Dimensions of Self-Efficacy for Writing. Journal of Educational Psychology, 105(1), 25-38

clustered standard errors including a class-level treatment indicator, a school-level fixed effect, and baseline covariate. PIRA raw scores will be used in all analyses.

Our baseline covariate will be the child's Early Years Foundation Stage Profile (EYFSP) aggregate score for four learning goals:

1) understanding;
2) speaking;
3) reading; and
4) writing.

These goals were selected as they are most closely linked to reading, our primary outcome measure. Past research found that neither the total EYFSP score nor the score for personal, social and emotional development correlated well with later attainment.[22] This aggregate score will range from 4 to 12.

The estimated impacts will be "intention to treat" (ITT) effects and will be reported with 95% confidence intervals. We will calculate Hedge's g effect size by dividing this coefficient by an estimate of the pooled total variance of the outcome variable and applying the appropriate correction factor.

Secondary analysis

The secondary analysis will measure the impact of the intervention on the pupils' social skills and creative self-efficacy.

Social skills will be measured using scores on the social skills subset of SSiS produced by teachers, with a baseline covariate consisting of EYFSP scores aggregated across the following learning goals:

1) self-confidence and awareness;
2) managing feelings and behaviour; and
3) making relationships.

The EYFSP aggregate score will range from 3 to 9.

Creative self-efficacy will be measured using the ideation sub-measure of the writing self-efficacy measure with a baseline covariate consisting of EYFSP scores aggregated across the following learning goals:

1) exploring and using media and materials
2) being imaginative

The EYFSP aggregate score will range from 2-6.

[22] Snowling, M. J., Hulme, C., Bailey, A. M., Stothard, S. E., & Lindsay, G. (2011). Better communication research project: language and literacy attainment of pupils during early years and through KS2: does teacher assessment at five provide a valid measure of children's current and future educational attainments?. London: Department for Education.

Subgroup analysis

We will carry out a subgroup analysis to measure the impact of the intervention on **everFSM** pupils. Following EEF guidance, we will first test for an interaction of the treatment and **FSMever status**. If a significant interaction is found, we will estimate a separate model on the restricted sample of only **FSMever pupils**. This procedure will be carried out for both our primary and our secondary outcomes.

Other

We will report the distribution of missing observations by treatment arm and explore whether baseline characteristics are balanced across trial arms.

An estimate of the intra-cluster correlation of the primary outcome measure will be extracted by estimating a variance components model for this purpose.

Definition of fidelity/on-treatment minimum

We outline below the fidelity measure and on-treatment minimum for First Thing Music. This measure assesses the minimum standards required in order for the delivery team to be satisfied that it is on-treatment – it is not an assessment of quality of engagement. The delivery team believe no less than 80% of all daily sessions must be conducted throughout the year. The teacher may only miss a maximum of two training sessions; if the teacher misses a session, the teacher must be followed up with individually. The delivery team will ask schools to conduct a register in order to document the number of sessions.

This measure provides useful contextual information for the process evaluation and may help us decide which schools to sample for the case studies. We will also estimate the treatment effect across all three outcome measures for compliers using a Complier Average Causal Effect (CACE) analysis using the minimum compliance definition above.

Implementation and process evaluation methods

Introduction

A robust and in-depth implementation and process evaluation (IPE) is vital to ensure we understand the extent to which First Thing Music achieves positive outcomes for children. In the first section, we outline the overarching implementation questions that will be explored across all projects, including First Thing Music. These cross-project similarities in delivery and in what they are aiming to achieve are outlined in the appendix[23]. We highlight, for each question, the dimension or factor affecting implementation it relates to, as specified in the guidance set out by the EEF.[24]

The second section outlines the IPE questions that are unique to First Thing Music.

A flexible research approach will be employed to capture the unifying and distinct elements of the five programmes. We will use similar methods to capture both the overarching IPE questions, as well as the project specific questions.

[23] For an overarching flow diagram of the programme similarities, please see appendix 1.

[24] Humphrey, N., Lendrum, A., Ashworth, E., Frearson, K., Buck, R., & Kerr, K. (2016). Implementation and process evaluation (IPE) for interventions in education settings: An introductory handbook. *Education Endowment Foundation (Ed.)*.

Cultural Learning IPE Questions

1. In what ways was the programme implemented? What are the barriers and facilitators of delivery (Fidelity)? In particular:
    1. Senior Leadership Team buy-in;
    2. Delivery of training and supporting materials – a) the extent to which is it consistent across sites; and, b) whether it appears to be effective in ensuring that teachers understand the aims and main features of the intervention;
    3. Delivery of the intervention – a) consistent across sites; b) whether it appears to facilitate children's engagement
2. To what extent did the schools engage with the intervention in line with the intervention aims? (Responsiveness).
3. How was the quality of the intervention perceived by teachers, senior leaders and teaching assistants? (Quality)
4. To what extent is the knowledge of arts practitioners delivering the intervention integrated with the pedagogic knowledge of teachers involved? (Implementer support system)

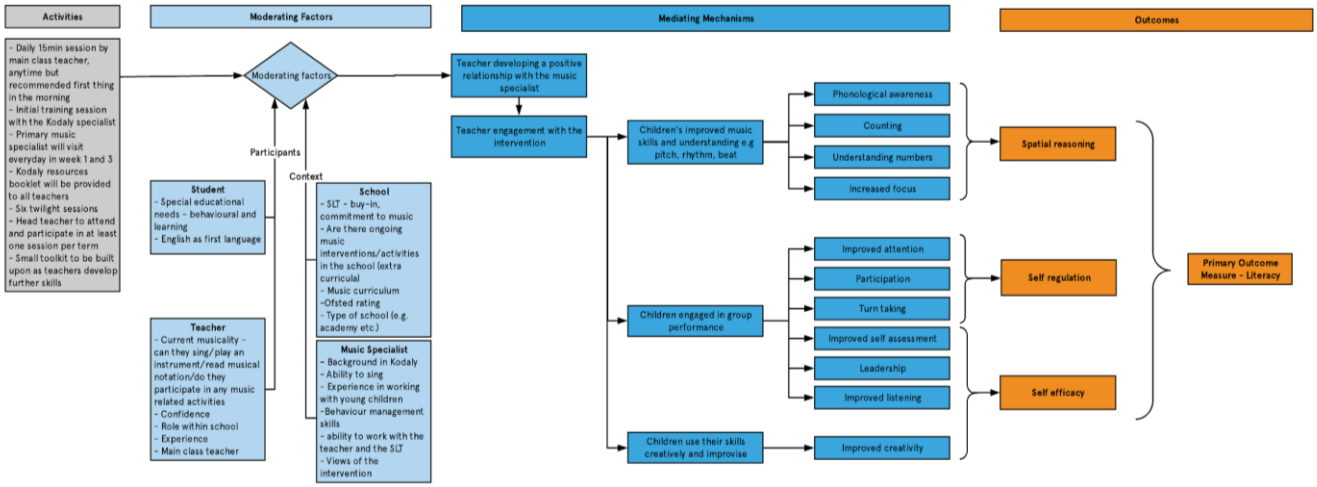
First Thing Music Questions

Beyond the overarching questions which will be asked, additional areas which will be important to explore are as follows:

1. What are the mechanisms that are taking place in the intervention and to what extent are they bringing about change? (Mechanisms)
2. To what extent does initial teacher confidence to deliver music lessons affect implementation, and how is the training adapted to support their needs ? (Quality and responsiveness)
3. What influences teachers' willingness to engage in music, and what music expertise do the teachers possess prior to engaging with the intervention? (Implementer characteristics)
4. To what extent can the time be created for the intervention every day? (Fidelity)
5. To what extent does the intervention differ from the music experience of those in the control group in class and around school? (Programme Differentiation)
6. To what extent does the intervention vary across schools, and does this affect implementation? (Adaption)

Logic Model

An IDEA workshop was held, utilising the TIDieR framework, to develop a logic model in collaboration with First Thing Music. The Logic Model will be instrumental in directing the IPE. Throughout the IPE, we will attempt to monitor the proposed mediating mechanisms as well as understand the role played by potential moderators. A summary of the similarities across all the logic models for the Cultural Learning interventions can be found in the appendix.



Methods

A suite of methods will be used to answer the research questions outlined above. These will be analysed in conjunction with the other sources of data to provide an in-depth yet generalisable understanding of the intervention. These methods will be the same across all projects to ensure consistency, but will vary according to the project delivery timetables. We will work closely with the First Thing Music team to ensure we conduct the data collection when appropriate. In addition to main project team input Professor Andrew Burn, specialist in English, Media and Drama; Professor Gemma Moss, literacy specialist; and Emerita Professor Sue Hallam, specialist in music education, will be invited to give feedback on the methods.

Observation of training. The IPE team will attend and observe the initial and midpoint training session delivered by the training provider, as well as review the materials used in the courses. Members of our team with expertise and knowledge of arts in education will lead the observations and fieldwork. The project team is also carrying out an informal evaluation of the training for their own purposes and will share anonymised data from the evaluation with us, which will be used to triangulate insights. This will be particularly valuable around measuring engagement in programmes and consistency of training.

Administrative data. Working closely with the delivery partners, we will devise measures of engagement in the intervention and triangulate these metrics with the sampling to ensure our case studies (see below) target a variety of intervention settings. These measures may include online metrics, attendance or other relevant engagement related data. This will help us ascertain the feasibility and scalability of projects.

Case studies of schools. These will consist of interviews and classroom observations with a subset of approximately 6 schools. These case studies will consist of

- Teacher interview (where feasible, both before and after the observation)
- Observation of a Kodály session in each treated classroom
- Informal interviews with children

- Interview with SLT

The schools will be sampled based on a range of characteristics such as geography, Ofsted rating and engagement (see defining fidelity above). We will use documentary analysis of the resources at the heart of an intervention. Case study is a powerful research strategy to use within sequential explanatory mixed method designs and adds completeness to the exploration of complex issues in situ (Yin, 2013).

Online surveys. To gather data from all participating schools, we propose carrying out an online survey of control and treatment schools. The purpose of this survey would be to collect information on "business as usual" schools and classrooms, differences between "business as usual" and intervention classrooms, additional cost data, and a wider view of implementation and/or impact as measured qualitatively. To encourage participation and minimise the burden on respondents, it is expected that the survey would take teachers no more than 20 minutes to complete.

Triangulation

Multiple sources of data will be brought together to best answer the IPE questions. How these methods will be triangulated is outlined in the table below.

Cultural Learning IPE Questions	**Methods**
In what ways was the programme implemented? What are the barriers and facilitators of delivery (Fidelity)? In particular: 1. Senior Leadership Team buy-in; 2. Delivery of training and resources – a) the extent to which is it consistent across sites; and, b) whether it appears to be effective in ensuring that teachers understand the aims and main features of the intervention; 3. Delivery of the intervention – a) consistent across sites; b) whether it appears to facilitate children's engagement	Survey; Administrative Data; Case studies; Observation
To what extent did the schools engage with the intervention, in line with the intervention aims? (Responsiveness)	Survey; Administrative Data; Case studies
How was the quality of intervention perceived by teachers, senior leaders and teaching assistants? (Quality)	Case studies; Survey
To what extent is the knowledge of arts practitioners and other practitioners integrated with the pedagogic knowledge of teachers involved? (Implementer support system)	Case studies
First Thing Music Questions	
What are the mechanisms that are taking place in the intervention and to what extent are they bringing about change? (Mechanisms)	Case studies
To what extent does teacher confidence to deliver music lessons affect implementation, and how is the training adapted to support their needs? (Quality and responsiveness)	Survey; Case studies; Observation
What influences teachers' willingness to engage in music, and what	Survey; Case studies

music expertise, do the teachers possess? (Implementer characteristics)	
To what extent can the time be created for the intervention every day? (Fidelity)	Survey; Administrative data; Case studies
To what extent does the intervention differ from the music experience in the control group classes and school? (Programme Differentiation)	Survey
To what extent does the intervention vary across schools and does this affect implementation? (Adaption)	Survey; Case studies

IPE Analysis

Structurally, this will draw upon the analytical strategy of multi-case studies – whereby a programme is first coded individually and then a large cross-sectional analysis is conducted, which encompasses all programmes (Stake, 2013). The original plan was that the deductive analysis would be conducted on NVivo by the lead researchers, who will co-code 3 transcripts to ensure coder similarity and robustness of coding framework. The current research team uses the Framework approach to data management and analysis (Lewis et al, 2013), a robust and widely-used methodology within educational research and particularly appropriate for exploring implementation and identifying barriers and facilitators to effectiveness. This approach uses matrices within which the data is organised using columns (codes) and rows (cases). Codes will be pre-specified in a coding framework which reflect the research questions, but additional codes will be created as new themes emerge. The Framework approach does not use co-coding, as the approach to reporting makes the analysis transparent to the reader, so transcripts will not be co-coded as originally envisaged. A number of pieces of software can be used to support the approach, including NVivo, which was the software originally specified in this protocol. However, we propose using Excel as the team does not currently have access to NVivo and the planned volume of data means that a specialist software is not needed. More details about IPE analysis can be found in Appendix 2.

The analysis will be conducted in stages, first on the school, or case level, then across the cases involved in the trial. Finally, a cross-project analysis of the Cultural Learning aspects of the data will be conducted to ensure we identify significant patterns relevant to all interventions. This will take the form of a flexible, yet robust, thematic framework, which will include elements that are unique to each, but also relevant to all projects. It will be important to understand how the same theme may be manifested in a different way for different programmes (Bazeley, 2013).

IPE Data Collection Timeline

We understand that each project will follow a similar delivery schedule, with variation in the numbers and timing of training sessions across the year. This similarity allows us to map our data collection activities on to one timeline. We have arranged the timeline by term as the First Thing Music team are yet to specify exact timings for their programme delivery. We can

therefore consider this an indicative schedule of events across the academic year of 2018-19.

Date	Item
Autumn Term 2018	Observation of training
	Collection of baseline survey to measure school buy-in and teacher attitude towards intervention
	Collection of school characteristics
Spring Term 2019	Observation of mid-point training
	Conduct in-school case studies
	Collection of fidelity data to inform case study sampling
	Finalise sampling strategy
Summer Term 2019	Conduct in-school case studies
	Administer end of intervention survey
	Conduct analysis

Costs

An estimate of the per-pupil cost of the intervention will be calculated by the evaluation team. This estimate will focus on cost from the perspective of a participating school and will be based on the direct, marginal financial costs of implementing the intervention. This includes anything which the school needed to pay for beyond business as usual.

The cost estimate will make use of information from the project team (particularly with regard to the actual cost of delivering the intervention, e.g. the cost of providing the training), as well as that collected directly by the evaluation team from schools about the costs of preparing and implementing the intervention. Information on costs, especially any hidden costs or resource implications, will be explored through the process evaluation as part of the interviews with teachers and school visits. The purpose of collecting such data in the process evaluation will be to identify the main areas of expenditure required by the project. This process will also help to establish whether it may be appropriate to include any questions on costs/resource use in the survey. This will need to strike a balance between collecting sufficient cost information and not damaging response rates; it will also need to take account of whether a teacher is well placed to provide accurate information on particular types of costs.

Time spent by schools, such as the amount of time for which schools need to arrange supply cover for teachers to attend training, but also to prepare for delivery, will be reported separately from the financial costs. Any costs in terms of prerequisites will also be considered, such as musical instruments, books or other resources. Control group schools will also be asked about the time they invested in CPD, to ascertain how much time above and beyond business is usual is needed. We may also triangulate national data on this if available.

An estimate of cost per pupil per year will also be calculated based on the trial period, as once trained, teachers would also be able to deliver the programme in subsequent years. Any costs associated purely with the trial will be excluded.

Ethics and registration

Ethical approval has been sought following UCL Institute of Education staff ethics approval procedure. It was approved on the 20th of March 2018

Personal data for this trial will be processed under the legitimate interests provision of the GDPR (Article 6(1)(f)). Nevertheless, parents will be provided with information explaining their right to withdraw their child's data from the trial if they object to this processing of their child's data. This use of data has been allocated the following UCL Data Protection Registration Number: Z6364106/2018/02/09.

This trial protocol has been pre-registered at www.controlled-trials.com and assigned an International Standard Randomised Controlled Trial Number (ISRCTN) of 14035536.

Personnel

Delivery team:

- Lindsay Ibbotson
- Susan Robertson
- Lucinda Geoghegan
- Zoe Greenhalgh

Evaluation team:

BIT

- Pantelis Solomon (Principal Investigator)
- Kimberly Bohling
- Florentyna Farghly
- Fabian Gunzinger
- Johanna Frerichs
- Juliane Wiese
- Millie Devereux
- Bridie Murphy
- Eleanor Collerton
- Matt Barnard

UCL Institute of Education

- Jake Anders (Principal Investigator)
- Dominic Wyse
- Gemma Moss

- Andrew Burn
- Nikki Shure
- John Jerrim
- Susan Hallam

Responsibilities

Outcome measures administration and collection - BIT

Design of the trial

- sample size calculations - BIT
- refinement of randomisation approach - BIT

Delivery of the intervention

- recruitment of schools - First Thing Music
- session delivery - First Thing Music

Data collection

- Collection of pupil data - BIT
- Outcome measure collection (research assistant recruitment and coordination) - BIT
- Linking of UPN to NPD - IoE
- Data for process evaluation - BIT

Impact analysis - BIT (lead) and UCL

Qualitative analysis - BIT (lead) and UCL

Risks

The data security policies of UCL and BIT and the Data Sharing Agreement between BIT, UCL, and Tees Valley Music Service will be added to this protocol once drafted and approved.

Issue/risk	Risk level	Action to address issue/reduce risk
Dropout / non-compliance of settings	Medium	We want to avoid attrition of schools from the project as much as possible. We plan to minimise attrition by ensuring that schools that sign up are committed (by asking them to sign a Memorandum of Understanding). Keeping schools informed of progress and providing reminders of next steps will be important for engagement. The project team will also be asked to monitor changes in key personnel to ensure ongoing commitment. Minimising the data collection burden on schools will also be important for retention. We will also randomise only after schools have followed opt-out collection procedures and provided the necessary student data.
Difficulty in collecting data needed prior to randomisation (i.e. pupil data)	Medium	We will work closely with delivery teams and maintain regular contact. A school recruitment timetable, which includes a communication schedule, will be shared and agreed with the delivery partners. As part of this, delivery partners will be asked to send a weekly email, or to update a shared spreadsheet with contact details of recruited schools. Pupil data will be submitted directly to BIT, who will screen each data set as it comes in to check for any incomplete or incorrectly entered data, to ensure a school is eligible for randomisation. The school recruitment timetable builds in sufficient time to follow up with schools who have either not returned their data on time or have returned incomplete data to ensure that randomisation is not delayed.

Difficulty recruiting schools	Medium to high	We are confident that the project team will convey the importance of the evaluation to settings and the value to them of taking part. As classroom-level randomisation will mean all schools have at least one treated class and the programme will be offered to other classes in the subsequent year, this should mitigate any unease about children or the school 'missing out' (for 2+ form schools at least). Schools with 2+ forms will be prioritised during recruitment for this reason.
Objection to NPD linking	Medium	We plan to collect the necessary data to allow this long-term follow-up. We believe this processing of personal data is justified under the legitimate interests basis of GDPR (Article 6(1)(f)). Nevertheless, we will inform parents of the right to withdraw their child from processing of their data. We do not anticipate high or non-randomly varying levels of opt-out.
Missing Outcome Data	Medium	For directly collected assessments, attrition is a potential risk. BIT will ensure schools and research assistants understand the need to collect endline measures for as many students as possible in order to maximise external validity. Schools will also be contacted sufficiently far ahead of the endline primary data collection window to ensure we arrive at a convenient time for RAs to visit and run the PIRA tests (in order to avoid weeks or days in which large numbers of pupils are likely to be absent). Upfront notice will also give school teachers ample time to complete SSiS surveys. RAs will report to the BIT project coordinator the number of children not able to sit the PIRA after each visit. If the rate is high (>5% of sample) the project coordinator will contact the school to determine whether an additional testing day is appropriate and/or feasible. BIT will provide regular updates to the EEF and project team about testing completion and attrition. To ensure PIRA response papers are not lost in the postal system, they will be couriered to the test publisher for marking. Once marking is complete the test publisher will then send BIT an electronic record of marks (over a secure service) and courier the hard copy papers themselves.

Parent and teacher concern about 'over-testing'	Low	Communications to schools (during recruitment) and parents (when informing them of the trial and their right to withdraw their child's data from processing) should emphasise that only one assessment will be taken by children due to this study (the endline PIRA).
Contamination	Medium	Communications from both the project and evaluation team will stress that the class (or classes) assigned to the control condition cannot be given access to First Thing Music materials or sessions. The school MoU will also be explicit on this point. Music specialists delivering the intervention and teacher training will also be instructed to report any instances of children or teachers in control groups attempting to access the intervention. To address any unease about this in schools, First Thing Music will offer all control group classes access to the programme at a reduced cost once endline outcome measure collection has been completed. Process evaluation will also explore whether any contamination occurred.
Treatment compliance	Low	We view this not so much as a risk but as the reality of implementing such an intervention. The impact estimates (Intention to Treat) therefore relate more to the type of treatment likely to prevail in practice rather than the type of impact that could be seen were it possible to achieve laboratory-type conditions. Nevertheless, understanding treatment variation is important and will be explored through CACE analysis of the on-treatment sample as well as being a key focus of the implementation and process evaluation.
Unexpected absence or loss of team members	Low	The team will substitute for each other during any short-term absence. In the event of longer periods of unplanned absence or departure, we will recruit replacements. As BIT and UCL are joint evaluators, there is a relatively large pool of staff with experience in education evaluation who could substitute for members of the team, should this be necessary.

Timeline

Date	Activity
October 17 - March 18	School recruitment (First Thing Music)
March- April 18	MoU signing (First Thing Music)
April - May 18	Distribution of withdrawal forms to parents (BIT)
May 18	Final date of return of withdrawal forms before schools send pupil data to evaluators (BIT)
May - June 18	Randomisation (BIT)
September 18 - July 19	Intervention delivery (First Thing Music)
September 18	Observe training (BIT)
October 18	NPD application (UCL) and IPE baseline survey (BIT)
November 18	Baseline survey (BIT)
February 19	Observe second training (BIT)
March-April 19	Conduct sampling for case studies (BIT)
March-May 19	Case studies for IPE conducted (BIT)
May - July 19	Endline (PIRA & SSiS) administered by RAs (BIT)
July 19	Endline IPE survey (BIT)
July - August 19	Marking of PIRA endline assessments (Hodder, contracted by BIT) Data entry of writing self-efficacy measure
September - December 19	Analysis and report writing (BIT and UCL)

Appendix 1 Implementation and Process Evaluation Overarching Similarities

Similarities across projects

The logic models from the five cultural evaluations were compared to understand their similarities and differences. From this, an amalgamated flow chart was designed to show the general route that all the programmes can take (Figure 1).

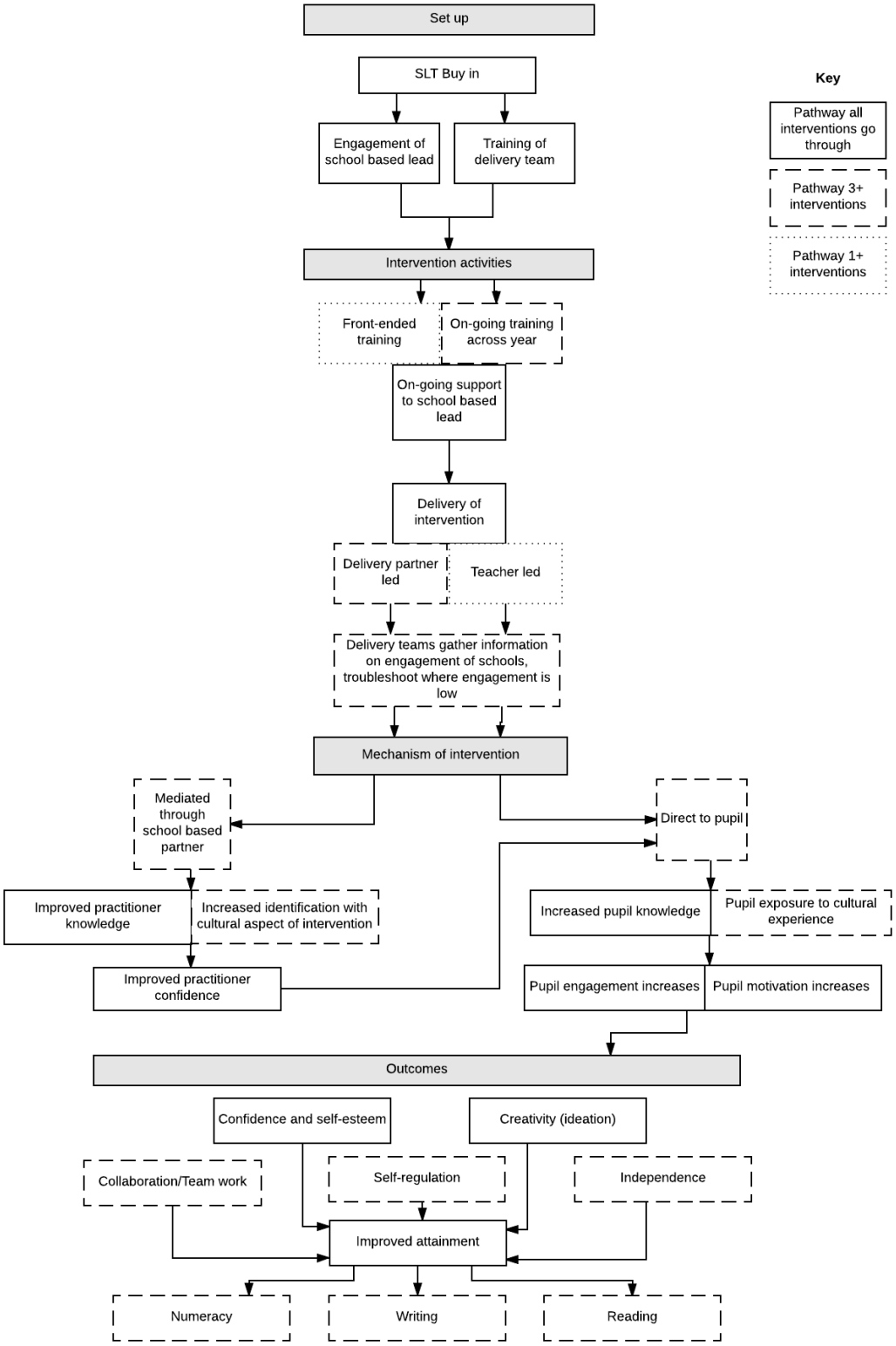


Figure 1 Amalgamated logic model of the five interventions

From Figure 1, we can see that the following are standard across all five interventions:

Implementation Similarities

1. Senior leadership buy-in
2. On-going (yet varied) support from delivery team staff - relationship with school, and teachers or teaching assistants
3. Training days for teachers or teaching assistants
4. Delivery teams gather information which helps them understand how the schools are engaging in the intervention

When considering the differences in implementation, there are two possibilities which all of the five interventions take: 1. The intervention is mediated through school-based partners, or 2. The intervention is delivered directly to pupils. These two possibilities should be measured in a standardised fashion as they may have implications for how arts-based programmes are designed in the future. These 'options' are outlined below:

1) Training model - front-end loaded and/or on-going across the year
2) Direct delivery of intervention - via a member of school staff or via a delivery partner
3) Mechanisms of change - mediated through a member of school staff or delivered directly to pupils
4) For writing orientated interventions, the extent the practices reflect robust evidence of what works

Moderating factors

Across the five interventions, several common moderators emerged from the logic model IDEA workshops. We will aim to capture these systematically when drawing up the MOU with the schools. Of all the 29 different moderators outlined, we will systematically capture those referenced by four or more of the projects. These are as follows:

1) School Ofsted rating
2) Current activities relevant to the intervention
3) Pupil SEND/EAL
4) Teacher/TA experience (years)
5) Teacher/TA background knowledge in arts-related programmes

Mediating factors

There was generally much less overlap between projects overall in relation to mediating factors and the 43 mediating mechanisms listed (although many between-project similarities). The only ones which were relevant for four or more of the projects were broad, and the first is being captured in some of the projects already. The second, creativity, will also be captured as part of the overarching Ideation measure.

1) Improved pupil self-efficacy
2) Improved creativity

Appendix 2 – Additional detail about the IPE

A) Research questions: What data will be used to answer each question (from how many data sources)?

The implementation and process evaluation will use both qualitative and quantitative data sources to answer the key research questions.

1) Qualitative interviews and observations: these data sources will be used to map the range of ways the programme has been implemented, identify the barriers and facilitators to delivering the programme and ensuring fidelity, and to building on the programme's theory of change to develop explanations for understanding what helps and hinders the programme achieving its key outcomes
2) Surveys and administrative data: these data sources will be used to measure the level of engagement of schools and teachers, quantify levels of perceived fidelity and effectiveness and get feedback on the perceived quality of the intervention

B) Case Studies

1) Selection of case study units

   Six schools will be sampled for the case studies. We seek to select schools that have different characteristics, so that we can capture variation in the experiences that teachers will have of the programme and implementing what they have learnt in their schools. The primary sampling criteria will be:

   a) FSM: a high or low percentage of pupils receiving free school meals, which is defined as above or below the median (17.5%)

   b) Engagement: based on the average number of First Thing Music sessions completed each week as reported by the class teacher with high engagement delivering an average of 4-5 sessions per week and low engagement delivering fewer than 4 sessions per week..

   The secondary sampling criteria will be:

   a) Location as defined by local authority, so as to capture the range of practitioners working with the schools

   b) Ofsted rating: Good, Outstanding, or Requires Improvement

   To conduct sampling, we will begin with primary sampling criteria by selecting three schools with high FSM and three schools with low FSM, of which 3 are High Engagement and 3 Low Engagement. While selecting schools, we will ensure that we are also achieving variety in terms of Ofsted and location.

2) Research methods/data sources informing each 'case'

   Case study schools will be asked to nominate a date that is convenient for them for a BIT researcher to conduct a one-day visit. On the day, the researcher will observe a First Thing Music session and talk to the children about what they think of the programme. Generally the discussions with children take place as a group, as this tends to make the children more comfortable and allows them to more naturally

stimulate conversation. The researcher will also conduct audio-recorded interview with the class teacher and a member of the senior leadership team. The purpose of interviews with teachers is to understand their experience of the programme, the barriers and facilitators to engaging with the programme, how the training influences their practice and any subsequent effect this has on their pupils. The purpose of interviews with the senior leadership team is to understand their perception of the quality of the programme, barriers and facilitators to delivering it, any support they have provided for the teacher to engage with the programme, and any potential impacts of the programme that they have observed. Dom Wyse will attend one of these case study visits alongside a BIT researcher. The aim of this is to ensure that Dom has seen the programme in action, which will enhance his ability to feed his expert knowledge into the analysis and reporting for the IPE.

3) Participant numbers and selection criteria for each method

At each case study school, we will interview a teacher in an intervention classroom (N=1) and a member of the senior leadership team (N=1), who has overseen the teacher's engagement with the First Thing Music programme. We will also observe the class of the interviewed teacher as they take part in a First Thing Music session. We will have informal conversations with students about the programme; the number of students is dependent on how many parents return the consent form.

4) Analysis

The analysis of the qualitative data will be undertaken using tools and processes associated with the Framework approach[25], which emphasises the importance of being comprehensive, systematic and transparent. In practical terms, the first step in analysis is to manage the data using Framework, a case and theme-based approach to data management. Key themes that emerge from the interviews will be identified through familiarisation with the transcripts. An analytical framework is then drawn up and a series of matrices are set-up in Excel or specialist software such as NVivo, each relating to a different issue. The columns in each matrix represent the key sub-themes or topics and the rows represent individual participants. Data from each interview is noted in the appropriate cell, so the data are ordered systematically and grounded in participants' accounts. This means that each part of a transcript that is relevant to a particular theme will be noted, ordered and accessible and can easily be shared within the team.

The second part of the analytical process is to work through the charted data, drawing out the range of experiences and views and identifying similarities and differences. This involves undertaking thematic analysis of specific issues (crudely looking down columns in Framework) as well as case-based analysis (crudely comparing and contrasting rows in Framework), through which typologies will be identified where they exist and explanatory accounts developed. During the analytical process a balance will be maintained between deduction (using existing knowledge and concepts relevant to the issue) and induction (allowing new concepts and ways of interpreting experience to emerge from the data). As qualitative data cannot be generalised in terms of prevalence, the analytical outputs will focus on the range and

[25] Ritchie, J., Lewis, J., McNoughton-Nichols, C., Ormston, R. (2013) Qualitative Research Practice (2nd edition).

diversity of experiences and key concepts, avoiding numerical summaries or quasi-numerical language such as ‘most’ and ‘majority’[26].

5) Data collection timeline

March 2019

- Sampling schools
- School recruitment

March – May 2019

- Case study data collection

[26] Ibid.