Education Endowment Foundation

THE CRAFT OF WRITING

Evaluation Report

September 2021

Jake Anders, Nikki Shure, Dominic Wyse, Matthew Barnard, Johanna Frerichs, Kimberly Bohling

UCL Institute of Education

THE 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

About the evaluator

The project was independently evaluated by a team from University College London (Dr Jake Anders, Dr Nikki Shure, Professor Dominic Wyse, Professor John Jerrim, Professor Gemma Moss, Professor Andrew Burn) and the Behavioural Insights Team (Dr Matt Barnard, Kim Bohling, Johanna Frerichs, Faisa Abdi).

The lead evaluator was Dr Jake Anders.

Contact details:

UCL Centre for Education Policy and Equalising Opportunities
Institute of Education
20 Bedford Way
London WC1H 0AL
Email: cepeo@ucl.ac.uk
Telephone: +44(0)20 7612 6000

Executive summary

The project

The Craft of Writing (CoW) is an intervention aimed at improving the writing skills, writing self-efficacy and writing creativity (ideation) of primary school pupils by developing teachers as 'writers' and improving their own writing practice and their teaching of writing. This intervention was delivered to Year 5 pupils between June 2018 and July 2019 by the University of Exeter, the Open University and Arvon, a creative writing charity.

CoW consists of training for self-nominated teachers, which takes place over two residential weekends and three follow-up continuing professional development (CPD) days. The training, led by professional authors, focuses on learning to apply the CoW framework in the classroom. The CoW framework is made up of five overarching elements: language choices, text-level choices, the reader–writer relationship, being an author, and the writing process.

The study was a two-arm school-level clustered randomised efficacy trial with 94 schools independently evaluated by a team from University College London (UCL) and the Behavioural Insights Team (BIT). The trial looked at the impact of the intervention on pupils' writing attainment as primary outcome, and writing self-efficacy and ideation as secondary outcomes, although we note the challenges inherent in capturing pupils' writing attainment at scale. The study also included an implementation and process evaluation (IPE) which measured compliance and fidelity, and qualitative analysis of case studies from six schools that took part in the trial. This evaluation was jointly funded by the EEF and the Royal Society of Arts (RSA).

Key conclusions
1. Children in the Craft of Writing (CoW) treatment schools made no additional progress in writing compared to children in business-as-usual control schools. This is our best estimate of impact which has a low to moderate security rating. However, as with any study, there is uncertainty around the result: the possible impact of this programme ranges from three months' less progress to positive effects of two additional months of progress.
2. FSM-eligible pupils in the CoW treatment schools made no additional progress in their writing or self-efficacy test scores as compared to children in business-as-usual control schools. However, free school meals (FSM)-eligible pupils in the CoW treatment schools made progress in ideation, on average, compared to children in the business-as-usual control group. This provides some indication that the intervention may be beneficial for increasing the creativity of children from disadvantaged backgrounds.
3. 94% of surveyed teachers, who self-nominated to participate in the study, said that CoW improved their confidence as a writer and 67% said that the intervention had a 'very positive' impact. However, surveyed teachers, reported a tension between using the materials they had been given and following the national curriculum.
4. Teachers reported that the intervention challenged them to rethink how they teach writing and their own writing practice, which, if successful and useful, may only have an effect in a longer-run timeframe than measured in this trial.

EEF security rating

These findings have a low to moderate security rating. This was a well-designed efficacy trial, which tested whether the intervention worked under developer-led conditions. The trial lost three padlocks, due to missing data and the number of pupils and schools who started the trial but were not included in the final analysis. Over 30% of the pupils who started the trial were not included the final analysis, either because schools declined to participate in testing at the end of the intervention, or because pupils were absent at the point of testing. In addition, it is not entirely clear how many pupils were randomised to the intervention.

Additional findings

Pupils in CoW schools made no additional progress in their writing test scores compared to those in the control group equivalent. This is our best estimate of impact, which has a low to moderate security rating. As with any study, there is always some uncertainty around the result: the possible impact of this programme also includes small negative effects of three months of less progress and positive effects of up to two months of additional progress.

Pupils in CoW also made no additional progress in their self-efficacy and ideation scores (secondary outcomes) compared to those in the business-as-usual control group. While there was no evidence of a differential effect on the primary outcome measure for FSM-eligible pupils, the sub-group analysis seems to suggest three months' progress on FSM pupils' ideation or creativity as a result of participating in the CoW intervention. However, this result may be less secure than the primary outcome findings, due to the smaller sample within the FSM sub-group.

Whilst its development (Dunsmuir et al., 2015) and the observations from a pilot of the writing assessment measure (WAM) suggested that 15 minutes was sufficient for pupils to produce a writing sample that could be meaningfully assessed, the programme developers raised concerns on the limited amount of time. The evaluators took this concern seriously while aiming to maintain broad alignment with previous use of the WAM. As such, in a change from the WAM as previously used, five minutes of planning time in which pupils could make notes were provided at the start of the activity.

The IPE reported low levels of compliance in terms of attendance at the training sessions (with 45% of schools not complying to the expected training requirements); however, further analysis did not find evidence of differential treatment effects among schools with higher levels of compliance for the primary outcome.

The IPE and qualitative case studies analysis found evidence supporting the majority of elements of the CoW logic model (Figure 1) and theory of change, and identified buy-in from the senior leadership team (SLT) and embedding of the approach in teaching the national curriculum as key factors for successful implementation. However, interviewed teachers felt that, whilst this intervention challenged them to change their thinking about writing and their own writing practice, there was a tension between using the materials they had been given and following the national curriculum.

This study found that 'pupils' enthusiasm about writing' and 'pupils applying the CoW framework in their written work' were also important elements or mechanisms that could lead to improved pupils' outcomes, and so these were ultimately added to a revised version of the CoW logic model (Figure 8).

The existing evidence on whether teachers' writing practice impacts pupil outcomes showed mixed results. The results from this trial support this existing evidence, since the effect sizes obtained are small in magnitude and are unclear (with the primary outcome's effect size being slightly negative and the secondary outcome effect sizes being slightly positive).

Cost

The average cost of CoW for one school was around £669 for the year, or £9 per pupil per year when averaged over three years.

Impact

Table 1: Summary of impact on primary outcome

Outcome/ group	Effect size (95% confidence interval)	Estimated months of progress	EEF security rating	Number of pupils	*p*-value	EEF cost rating
WAM score (ideas scale double weighted) / overall	–0.03 (–0.19, 0.12)	0		1697	0.68	£ £ £ £
WAM score (ideas scale double weighted) / FSM-eligible	–0.04 (–0.28, 0.20)	0	N/A	495	0.73	£ £ £ £

Introduction

Background

This evaluation is part of a round of funding between the Education Endowment Foundation (EEF) and the Royal Society of Arts (RSA) to test the impact of different arts-based learning strategies in English schools entitled 'Learning about Culture'.[1] The aim is to improve the evidence base around arts-based education programmes. It consists of five programmes: two in Key Stage 1 (KS1) (Reception and Year 1) and three in Key Stage 2 (KS2) (Year 5). Despite the unique aspects of these intervention models, there are many similarities in how they are delivered and what they hope to achieve.[2] The programmes have been supported by Arts Council England.

The background for the study is that a focus on increasing attainment in literacy and numeracy has been criticised for leading to a marginalisation of art, music and cultural studies in English schools (Neelands et al., 2015). The UK Government's Culture and Sport Evidence review (Newman et al., 2010), which summarised much of the observational and qualitative research in this area, 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 to cultural participation are particularly large for pupils from low-income groups (Catterall, 2009; 2012).

Research into the relationship between providing teachers with an opportunity to develop their own identities and skill sets as writers and pupil outcomes is patchy. There is some evidence of impact on teachers' skills, knowledge and confidence in facilitating creative writing (Redmond, 2010), and on pupils' attitudes and engagement as writers (Wilson, 2010), but a systematic review by Cremin and Oliver (2017) found that the evidence base regarding the impact of teachers' writing on pupils' outcomes is small and does not show a clear impact. Specifically, they reviewed all studies from 1990–2015, finding only 22 that met the standards to be included in their systematic review[3]; these studies lead to an inconclusive picture of whether or not teachers who teach writing 'need to be writers', but highlight the role of pre-service and professional development programmes in helping teachers develop their assurance and identities as writers.

Building on this limited prior evidence base, this evaluation was designed to estimate the effect of participating in the Craft of Writing (CoW) over the course of one school year on pupils' writing skills. The CoW is an intervention to develop teachers as writers, thereby improving their own writing practice and their teaching of writing (more details below in the *Intervention* section). This trial was designed as a two-arm clustered randomised trial with randomisation occurring at the level of the school. This level of randomisation was selected since entire classes participate in the programme and thus the risk of contamination within schools is very high. The two arms were: (i) participation in the CoW (treatment); and (ii) business-as-usual (control). Blocking was used in the randomisation to improve cross-arm comparability of schools, to improve precision of estimates, and to allow schools that sign up early to receive their allocation sooner than they otherwise would (this is important because of the nature of the intervention requiring as much notice of allocation as possible to be given to teachers, given it requires activity outside of their normal working hours; see the *Intervention* section for more details). Initially, we had planned to look at the long-term effects of participating in the CoW after one further year, looking at results from the end of KS2 national curriculum tests in English grammar, punctuation and spelling.

1 See the RSA website for further details (https://www.thersa.org/globalassets/pdfs/reports/rsa-learning-about-culture-report.pdf).
2 For an overarching flow diagram of the programme similarities, please see Appendix 1 in the evaluation protocol (https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Protocols/Craft_of_Writing_evaluation_protocol.pdf).
3 The inclusion criteria for this review were: 'studies focused on teachers' identities and practices as writers; addressed primary, secondary or pre-service teachers in mainstream education; and were peer-reviewed reports of empirical investigations with some connection to the classroom. Studies were excluded if they were not published in English; were EFL focused; or were autobiographical accounts from writing teachers rather than investigations' (Cremin & Oliver, 2017, p. 271). Studies were evaluated by two independent reviewers on two measures: 'methodological detail (research questions, sample, methods of data collection and analysis); and contribution of findings to the review questions' (Cremin & Oliver, 2017, p. 271). They were rated as low, medium or high.

We note upfront that it has been necessary for the analysis of this trial to deviate substantially from our initial plans set out in the project protocol and statistical analysis plan (SAP). These stem from issues in accessing the pre-test data that we expected to be able to obtain from the Department for Education (DfE)'s National Pupil Database (NPD). During the implementation of the trial the DfE changed the way in which data from the NPD are made available to researchers, switching from providing extracts that can be used alongside project data within evaluator's own secure computing systems to requiring access within the Office for National Statistics (ONS) Secure Research Service (SRS). In turn, this means that it is now necessary for project data to be uploaded to the SRS. Given that this project data is considered personal data over which we as evaluators are data controllers, this requires the conclusion of an appropriate data sharing or processing agreement between the evaluator and the DfE and/or the ONS in order to provide legally required reassurance by the DfE/ONS about the treatment of personal data over which the evaluator is controller.

Due to extended negotiations and delays between the evaluators and DfE/ONS, which we understand to have been severely exacerbated by additional workload due to the COVID-19 pandemic, in the interests of completing these evaluations and after discussion with the EEF and project teams, the decision was made to proceed with analysis, with deviations from protocol flagged as we move through the *Methods* section. These deviations were reported to the EEF ahead of conducting the analysis. Beyond issues inherent in deviating from pre-registered protocol, the main implications for the analysis are reductions in the statistical power relative to expectations.

It is important to understand the implications of this change. The purpose of including baseline measures in the current evaluation is to increase its statistical precision (i.e., to reduce the uncertainty around intervention impact estimates, which makes them more likely to be statistically significant). Importantly, both the original and the substituted baseline measures are taken from prior to the randomisation and intervention. Therefore, due to the randomised nature of the evaluation, their inclusion does not bias any intervention impact estimates, but only affects the statistical uncertainty around these estimates (i.e., the extent to which they are detectable as statistically significant). Given EEF policy to report impact estimates whether or not they are statistically significant, there is an increased risk that headline positive or negative effects are just due to this uncertainty, rather than representing a true effect. As a result, we particularly stress the importance of statistical significance as a check on interpretation of the results in this report.

Intervention

1. **Brief name.** Craft of Writing (CoW)

2. **Why (rationale / theory).** The Craft of Writing (CoW) intervention provides a sustained 'Arvon experience', developing teachers as writers, and is combined with a more explicit focus on pedagogical implications for the classroom. The intervention is underpinned by a framework of craft knowledge co-created with professional writers.[4] The CoW Framework represents a model of subject knowledge for teachers to inform their teaching of writing, and is structured around five aspects of knowledge: the writing process; being an author; the reader–writer relationship; textual choices; and language choices. It is used to support teachers in integrating what they learn from the Arvon teachers' confidence as writers and as teachers of writing.

3. **What (materials).** All resources and materials for CoW are available online from **http://socialsciences.exeter.ac.uk/education/research/centres/writing/projects/craftofwriting/interventionschools/**. The resources used included (see *Activities* in Figure 1):

    - the laminated CoW Framework, a booklet summarising the framework (its features and feedback elements), available as Appendix M of this report;

4 All intervention materials including the Framework of Craft are available at http://socialsciences.exeter.ac.uk/education/research/centres/writing/projects/craftofwriting/interventionschools/. The Framework is available in Appendix M.

- residential weekend and CPD PowerPoints (one set per day, topics include 'Being an Author', 'Language choices', 'Writing process' and 'The reader–writer relationship'. These PowerPoints were made available to teachers following their delivery);

- web resources available to the teachers, including teaching sequences, worksheets and modelled feedback via the project website.

4. **What (procedures).** In each school, one Year 5 teacher self-nominated to participate in the CoW intervention. This self-nomination approach avoids headteachers forcing participation, and was also intended to ensure participants would be engaged with the programme if randomised to the treatment arm, although it does make participants less likely to be representative of the wider population of primary school teachers. This teacher attended two residential training weekends, which were led by professional authors selected by Arvon, as well as three CPD days.

   These weekend sessions (Friday–Sunday), known as 'Arvon Teachers as Writers residentials', took place six months apart. The first residential weekend took place before the beginning of the school year. The residential weekends comprise workshops and tutorials for teachers led by professional writers, with time and space for writing, plus structured sessions to consider pedagogical implications and establish clear expectations for follow-through. Topics covered during the first residential weekend include 'Starting points for poems' and 'Starting points for writing'. Both residential weekends include a focus on embedding the CoW intervention into classroom practice. The first residential weekend for this trial took place in June 2018 and the second one took place in January 2019.

   The intervention also included a programme of three CPD days for teachers. Two of them took place during the six months between the two residential weekends and one took place after the second residential weekend. Each CPD day included a further 'Arvon Teachers as Writers' experience, as well as providing a programmatic sequence. The CPD days were led by the Open University and Exeter implementation team, with support from the professional writer tutors. The CPD days linked teachers' experience as writers with focused consideration of pedagogical transfer to the classroom. The goal of the CPD days was to focus on changing how teachers teach writing in the classroom, with an emphasis on using their improved knowledge of the craft of writing and through fostering a community of creative writers. The CPD days for this trial took place in September 2018, November 2018 and May 2019.

   To deliver the CoW intervention with fidelity as intended, teachers needed to apply the CoW framework (see Appendix M for further details), introduced to them through the residentials and CPD sessions, in the classroom. The CoW framework is made up of five overarching elements: language choices; text-level choices; the reader–writer relationship; being an author; and the writing process.

5. **Who (recipients).** The CoW is targeted at teachers of KS2 (Year 5) pupils.

6. **Who (implementers).** CoW is delivered by a consortium of the University of Exeter, the Open University and Arvon, a creative writing charity.

   The intervention is delivered by tutors from Arvon, a charity which delivers residential, online and community-based writing courses tutored by professional writers. In the case of this trial, the two Arvon tutors were Alicia Stubbersfield and Steve Voake. Alicia's most recent collection of poetry, *The Yellow Table*, was published in 2013. Her other poetry collections include *The Magician's Assistant*, *Unsuitable Shoes* and *Joking Apart*. Before becoming a lecturer in Creative Writing at Liverpool John Moores University, Alicia was an English teacher in several large comprehensive schools, most recently Cleeve School in Cheltenham. She regularly tutors for The Poetry School and Ty Newydd, National Writers' Centre for Wales. Steve is the award-winning author of more than 20 books for young people, including *The Dreamwalker's Child, The Starlight Conspiracy* and *Blood Hunters.* He was longlisted for the University of Canberra Vice-Chancellor's International Poetry Prize. He has experience as a headteacher in Somerset and has taught a writing master class for the *Guardian*. He is Senior Lecturer in Creative Writing at Bath Spa University. They are both regular Arvon tutors who lead Arvon courses, and Arvon teachers' courses (Teachers as Writers) in particular. In the case of this intervention, both of the

tutors are former teachers with SLT experience (as is the case with most Arvon tutors). They had both participated in a previous research study: *Teachers as Writers.*

Sessions focusing on the pedagogic implications of the residential experience were delivered by Professor Teresa Cremin, Professor of Literacy in Education at The Open University, and Professor Debra Myhill, University of Exeter.

One Year 5 teacher participated per school.

7. **How (mode of delivery).** Sessions were delivered face-to-face by the CoW team to participating Year 5 teachers.

8. **Where (setting).** Residential weekend sessions were delivered at the Arvon residential weekends at Arvon's Lumb Bank Centre (**https://www.arvon.org/centres/lumb-bank/**). The CPD days were delivered at hotels in Leeds and Manchester.

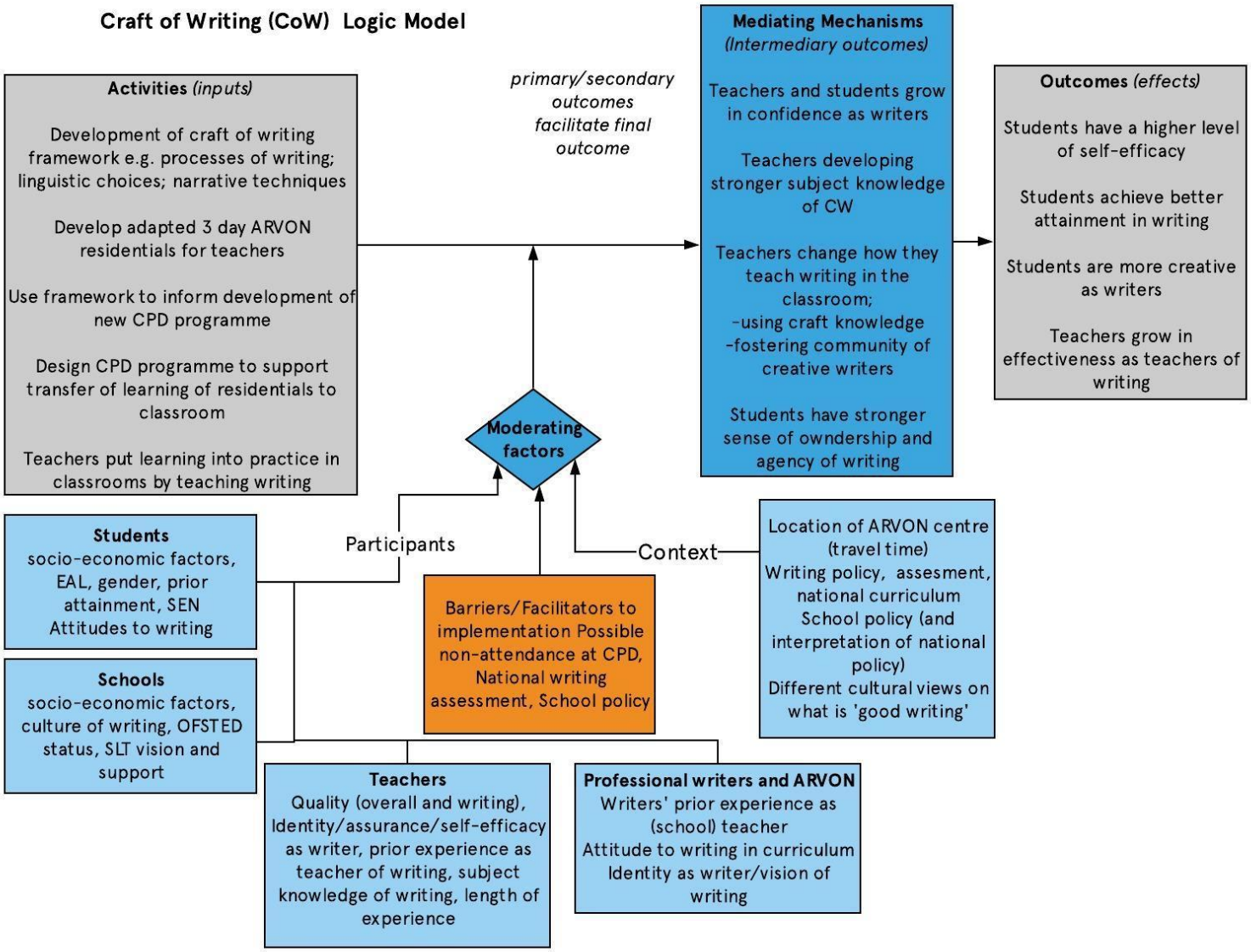
9. **When and how much (dosage).** The CoW intervention occurred over the course of an entire school year. There were two residential weekends (Friday–Sunday), which occurred six months apart, and three days of CPD, two that happened during this six-month gap, and one that followed the second residential weekend (i.e., in September 2018, November 2018 and May 2019).

10. **Tailoring.** The intervention was not planned to be personalised, meaning that all teachers received the same intervention.

11. **Modifications.** No modifications were made.

12. **How well (planned).** This evaluation examined two implementation dimensions: compliance and implementation fidelity. Compliance was conceptualised as a binary indicator of whether the teacher received a sufficient amount of the intervention (defined below using specific criteria) to be considered as having received the intervention (i.e., treatment). Compliance with this intervention was measured at the school level (only one teacher participated per school), which reflects the intervention delivery method. A school is considered to have complied *if and only if* the following two conditions are met: (a) the teacher must have attended both residential weekends, and (b) the teacher must have attended at least two of the three CPD sessions. In comparison, implementation fidelity was conceptualised as how the way in which the intervention was implemented in practice compares to the intended implementation of the intervention, as described in this section of the report. Case study data were used to examine the variation in how the intervention was implemented and any adaptations made in the case study schools, alongside identifying barriers and facilitators to implementing the intervention with fidelity.

Figure 1: Original logic model of the CoW intervention



Evaluation objectives

Impact evaluation

The evaluation sought to answer the following research questions (RQs). These RQs were reported in the evaluation protocol (initially published in May 2018 with revisions for clarity published in March 2019, available at **https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Protocols/Craft_of_Writing_evaluation_protocol.pdf**) and further details on the quantitative approach were provided in a statistical analysis plan (published in October 2018, **https://educationendowmentfoundation.org.uk/public/files/Projects/CoW_SAP_1.1_(amended).pdf**).

The primary objective of this evaluation was to estimate the effect of participating in the CoW over the course of one school year on pupils' writing skills.

Primary research question

- What is the effect of participating in the CoW over the course of one school year on pupils' writing skills?

Secondary research questions

1 What is the effect of participating in the CoW over the course of one school year on pupils' writing self-efficacy?

2 Does participating in the CoW over the course of one school year have an impact on pupils' perception of their own capacity to generate ideas?

Implementation and process evaluation (IPE)

The IPE focused on addressing RQs related to the implementation, delivery and perceived impact of the intervention. The IPE was designed to explore overarching implementation questions across all five Learning about Culture trials, as well as RQs specific to CoW. The four overarching questions were written based on cross-project similarities; however, not all questions apply to each programme, due to variations in training and delivery (see Appendix Q for the model depicting cross-project similarities). The relevant overarching implementation questions that are explored across all projects, as well as RQs specific to CoW, are detailed below, along with explanations for deviations from the evaluation protocol:

RQ1 In what ways was the programme implemented? What are the barriers and facilitators of delivery (Fidelity)? In particular:

a) Senior leadership team (SLT) buy-in;

b) Delivery of training: (i) the extent to which it is consistent across sites *[not answerable]*; and (ii) whether it appears to be effective in ensuring that teachers understand the aims and main features of the intervention;

c) Delivery of the intervention: (i) consistent across sites *[not answerable]*; (ii) whether it appears to be effective in supporting children's attainment; and (iii) whether it appears to facilitate children's engagement.

RQ2 To what extent did the schools engage with the intervention in line with the intervention aims? (Responsiveness).

RQ3 How was the quality of the intervention perceived by teachers, senior leaders and teaching assistants? (Quality)

RQ4 *[Not applicable]* To what extent is the knowledge of arts practitioners delivering the intervention integrated with the pedagogic knowledge of teachers involved? (Implementer support system) [*This question is not applicable because CoW does not involve direct delivery in schools by arts practitioners.]*

Both RQ1b and RQ1c contain sub-questions about the consistency of training and delivery. It was not possible to capture detailed insights on the consistency of training and programme delivery. As such, the findings on training focus on

teacher engagement and effectiveness. The findings on delivery focus on effectiveness in supporting attainment and engagement, as planned.

In addition, the IPE sought to answer these questions specific to the CoW intervention:

RQ5 How do the strategies and techniques from the CPD and residential training emerge as part of teaching, and in what ways does this lead to improved teaching practice?

RQ6 To what extent does the intervention improve teacher confidence as a writer?

RQ7 How does the intervention benefit the pupils: what are the mechanisms by which the teacher implements the intervention with their pupils?

RQ8 How does teachers' practice in relation to pupils revising their writing change?

We also planned to estimate the longer term effects of participating in the CoW after one further year, looking at results from the end of KS2 national curriculum tests in English grammar, punctuation and spelling. However, it will not be possible to carry out this analysis due to the cancellation of these assessments as a result of the COVID-19 pandemic.

Ethics and trial registration

The project's aims, methods and materials were reviewed through the processes laid out by the UCL Institute of Education Research Ethics Committee, and approved on 14 December 2017. While the application was approved, the ethics reviewers stressed the importance of ensuring ongoing pupil assent for participation in any evaluation activities throughout the research. As such, all research assistants (RAs) conducting assessments with pupils verbally described the activities to the pupils using age-appropriate language, informed them all activities were voluntary, and gave an opportunity for pupils to decline to participate.

Schools were informed about the trial through initial information from the developer and formally committed to participation by signing a memorandum of understanding (MoU). A template version of this document is included as Appendix L to this report.

This trial protocol has been pre-registered at www.controlled-trials.com and assigned an International Standard Randomised Controlled Trial Number (ISRCTN) of ISRCTN10546365.

Data protection

As part of this project, we processed pupils' and teachers' personal data. For this reason, it was important that we processed these data lawfully, following the principles laid out in the Data Protection Act 1998 (DPA) until May 2018, the EU General Data Protection Regulation (GDPR) from May 2018 until December 2020, and the UK General Data Protection Regulation from January 2021 (the project spanned these three regulatory periods). We explain the lawful basis below with respect to the GDPR, but there are equivalent regulations in the DPA for the justifications set out below.

BIT used Article 6(1)f of the GDPR as the lawful basis for processing personal data as part of this project. This is known as the 'legitimate interests' basis. For further information on this article see**: https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/legitimate-interests/.** BIT carried out a 'legitimate interests assessment' in support of this, identifying societal benefits to this processing of personal data. Specifically, the use of pupils' and teachers' personal data as part of this research were to understand the benefits to pupils of participating in this programme in terms of their academic attainment and other related benefits. This has public benefits that BIT believes are significant in terms of understanding whether this programme has the potential to benefit children in schools across England. Without processing these data it would not have been possible to provide this quality of new evidence.

UCL used Article 6(1)e of the GDPR as the lawful basis for processing personal data as part of this project. This is generally known as the 'public task' basis. UCL has reviewed current ICO guidance available here:

https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/public-task/, and has determined that this research forms part of its performance of a task in the public interest, as one of its core purposes provided for in its Charter and Statutes. This use of data was allocated the following UCL Data Protection Registration Number: Z6364106/2017/11/69 social research.

We do not believe that any of the data we processed falls within the definition of special category data under the GDPR. This would require an additional justification under Article 9(2) of the GDPR.

We informed pupils' parents of the proposed data processing and provided the opportunity to object to this. If parents objected, then the pupils' data were never passed to us by schools. If a parent chose to withdraw their child's data at a later stage, then the data were destroyed. The data controllers were named in the privacy information provided as part of this project, and contact details provided in case they had any queries about the data we hold about them, including provision and deletion of their data. The relevant letters and forms have been reproduced in Appendix P.

The information provided to parents explained in clear and plain language the lawful bases for processing, the purpose of processing the data, that they could object to this data processing and this would be respected, contact details of the organisation, and categories of data that have been processed.

Data will be kept until the end of the research project, including academic paper writing and dissemination (and certainly not longer than 10 years in line with UCL's policy on data retention). When they are deleted, they will be securely destroyed.

Data will be shared with the Education Endowment Foundation (who funded the trial), EEF's data contractor FFT Education (who manage the EEF's Data Archive) and (in a form that is truly anonymised) to the UK Data Archive. Details of this sharing were included in relevant Data Privacy Notices.

Project team

The project team comprised Prof. Debra Myhill and Sara Venner at the University of Exeter, Prof. Teresa Cremin and Dr Rebecca Coles at the Open University, and Becky Swain at Arvon. The intervention was designed by Myhill, Cremin and Swain, and was delivered by Arvon, a creative writing charity, with support from the rest of the project team.

The impact evaluation was led by Dr Jake Anders and Dr Nikki Shure at UCL Institute of Education with support from Prof. John Jerrim, and was led at the Behavioural Insights Team (BIT) by Kimberly Bohling and Dr Matthew Barnard. Data collection was managed by Faisa Abdi, Eleanor Collerton, Camilla Devereux, Amber Evans, Louise Jones, Alex Manby, Bridie Murphy and Juliane Wiese of BIT. Primary data collection was carried out by research assistants employed by BIT and marking of those data was also carried out by research assistants employed by BIT, drawn from finishing students at UCL Institute of Education. The IPE was led by Dr Matthew Barnard and Johanna Frerichs of BIT with input from Prof. Dominic Wyse (UCL IPE lead), Prof. Gemma Moss and Prof. Andrew Burn at UCL Institute of Education. The cost evaluation was led by Dr Matthew Barnard with support from Faisa Abdi of BIT. The evaluation design was also supported by Daniel Carr, Dr Florentyna Farghly, Dr Jessica Heal and Dr Pantelis Solomon of BIT.

Methods

Trial design

Table 2: Trial design

Trial design, including number of arms		Cluster randomised, two arms
Unit of randomisation		School
Stratification variable(s) (if applicable)		Time of sign up (two randomisation batches). • Intersection of proportion of English as an additional language (EAL) pupils (high/low split at sample median within randomisation batch) • proportion of free school meals (FSM) pupils (high/low split at sample median within randomisation batch)
Primary outcome	Variable	Writing attainment
	Measure (instrument, scale, source)	Writing assessment measure (WAM), 0–32, (Dunsmuir et al., 2015)
Secondary outcome(s)	Variable(s)	Writing self-efficacy Ideation
	Measure(s) (instrument, scale, source)	Writing self-efficacy measure (WSEM), 16–80, (adapted from Bruning et al., 2013) Ideation captured from sub-measure of the WSEM, 5–25.
Baseline for primary outcome	Variable	Planned to be: Phonics attainment Protocol deviation[5]: FSM status, EAL status, class FSM composition, class EAL composition
	Measure (instrument, scale, source)	Planned to be: phonics screening check (DfE) Protocol deviation: 0/1 indicator of current FSM eligibility status, 0/1 indicator of EAL status, proportion of class reported FSM-eligible [0, 1], proportion of class reported EAL [0, 1]. (All derived from school reports collected ahead of randomisation; see pp.17–18 for further details and justification).
Baseline for secondary outcome(s)	Variable	Planned to be: Personal, social and emotional development skills Protocol deviation: FSM status, EAL status, class FSM composition, class EAL composition
	Measure (instrument, scale, source)	Planned to be: EYFSP personal, social and emotional development skills (DfE) Protocol deviation: 0/1 indicator of current FSM eligibility status, 0/1 indicator of EAL status, proportion of class reported FSM-eligible [0, 1], proportion of class reported EAL [0, 1]. (All derived from school reports collected ahead of randomisation see pp.17–18 for further details and justification).

[5] See *Background* section.

This trial was designed as a two-arm stratified, clustered randomised efficacy trial with randomisation occurring at the level of the school and outcomes measured at the level of pupils. This level of randomisation was selected since entire classes participate in the programme and thus the risk of contamination within schools is very high. The two arms are as follows:

- participation in the CoW (treatment)
- business-as-usual (control)

The primary outcome of interest is improvement in pupils' writing, with secondary outcomes of their writing self-efficacy and ideation (as a sub-domain of writing self-efficacy). Further information on the approach taken to capturing these is provided below.

Participant selection

In line with the trial protocol, the project aimed to recruit 96 English state-funded primary schools principally from the following geographical regions: the North West and North East, with a small number from London. In the end, 94 schools were fully recruited and randomised.

The CoW is currently delivered to teachers of pupils across the primary school age range. Year 5 was chosen for the purposes of the evaluation, given the greater perceived potential to capture writing-based outcomes from pupils at older ages, but without attempting to deliver and evaluate in Year 6 classes because of the perceived tension with KS2 national curriculum tests at the end of the year. Furthermore, evaluation of Year 5 delivery was intended to allow for medium-term follow-up in those KS2 national curriculum tests, but this was ultimately impossible due to their cancellation for the relevant year group because of the COVID-19 pandemic.

One teacher of a Year 5 class in each participating school self-nominated for participation in the trial. All children in the teacher's class participated in the trial (other than where objections to processing personal data were received or refusal to participate in evaluation activities). Except in unforeseen and unavoidable circumstances (e.g., teacher moving school), the teacher (and therefore pupils) who participated were selected prior to randomisation to minimise the potential for this to introduce differences between the intervention and control groups; except in this small number of cases, all data on participating pupils was collected pre-randomisation in order to assure this.

In order to be considered, schools had to:

1 agree to distribute opt-out consent[6] forms to parents;

2 provide pupil data in order to identify a consistent relevant analysis sample and to allow for linking to the DfE's NPD (for pre-test data and long-term follow-up);

3 identify the teacher who will participate in the trial;

4 cooperate with the project and evaluation teams during the trial.

Further details of these requirements are outlined in the MoU with schools, available in Appendix L.

The project team advertised the trial and also approached schools through their existing networks. The team aimed to recruit schools that have larger populations of individuals receiving FSM than the national average of 15.3 percent of

[6] Note that this is opt-out consent from a research ethics point of view, not from a data protection point of view. We note that since the first version of this protocol was agreed the GDPR has been implemented. As such, UCL's legal basis for processing this data is now considered to be 'public task' and BIT's legal basis for processing personal data is now considered to be 'legitimate interest'. 'Consent' is not used by either party as a basis for the processing of personal data. Nevertheless, it remains the relevant term in respect of research ethics.

pupils aged 5–10 (DfE, 2016). This was successful, judging by the proportion of pupils eligible for FSM in the analysis sample.

The eligibility criteria for schools to participate were:

1. participating schools must be English state-funded primary schools (they were recruited principally from the North West, North East and Yorkshire, with a small number from London, but this was not a formal eligibility requirement);
2. schools had to agree to distribute study information sheets, data privacy information and data processing objection forms to parents;
3. schools had to agree that, if allocated to the control group, they would continue with 'business-as-usual' for the duration of the trial;
4. schools had to return a signed MoU, including committing to participate fully in the study – including in the collection of outcome measures in summer 2019 – regardless of which trial arm they are assigned to;
5. schools had to find one Year 5 teacher who self-nominated to participate in the intervention;
6. schools had to agree to allow time for each assessment phase and liaise with the evaluation team to find appropriate dates and times for assessments to take place;
7. schools had to agree that teachers in both trial arms cooperate with activities for the IPE, if requested.

Outcome measures

Baseline measures

Baseline measures for this research were planned to be drawn from the DfE's NPD. All participating schools were asked to provide personal information about pupils in the participating teacher's class that would allow a reliable link to be achieved, based on current guidance from the DfE and balancing this against personal data minimisation requirements set out in data protection legislation. Using this link, it was expected that we would obtain information on pupils' performance in the phonics screening check (using the NPD variable PHONICS_MARK) for the primary outcome analysis, and assessments of pupils' personal social and emotional development from the Early Years Foundation Stage Profile (EYFSP) (aggregated scores from NPD variables FSP_PSE_G06, FSP_PSE_G07 and FSP_PSE_G08) for the secondary outcome analyses.

Due to the data access issues described above, an alternative approach was taken, with its design informed by an intention to maximise the explanatory power of our analysis model, and hence maximise the precision of our treatment estimates, given the data available. As such, instead of including the planned baseline measures in the model, we substitute the available demographic information that was collected about pupils ahead of randomisation (initially intended solely for the purposes of stratification / blocking as part of the randomisation process), specifically eligibility for FSM and whether the child has English as an additional language (EAL). We include these in the model as predictors themselves, and also aggregated them to the class level to produce composition variables, given evidence that school-level aggregate predictors also provide explanatory power (Bloom et al., 2007). FSM and EAL status are both known to predict academic attainment (Strand et al., 2015; Sutherland et al., 2015) and, as such, we expected this to improve power compared to an empty model.

Nevertheless, the improvement in statistical power is still likely to have fallen short of that we would have expected from including a prior attainment measure, as was planned. It is necessary to understand the implications of this change. It is important to stress that there are no expected implications for bias in our impact estimates of not having our planned baseline measures: the unbiasedness of RCT estimates derives from the randomisation, not from statistically controlling for differences at baseline. Indeed, in principle, there is no need to include any baseline measures at all in the analysis to achieve an unbiased estimate from an RCT. Inclusion of inappropriate covariates in our analysis would have the potential to introduce bias – such inappropriate covariates are ones that could have been affected by the treatment,

which is why we are including pupil characteristics from prior to randomisation. The main implication of this change is a reduction in statistical precision (i.e., the uncertainty around estimates that is inherent in all evaluations is likely to be larger in this evaluation than it would have been), which is manifested as wider confidence intervals (CI) (or, equivalently, less likely to be statistically significantly different from zero for a given size of impact estimate). Given EEF policy to report impact estimates whether or not they are statistically significant, there is an increased risk that headline positive or negative effects are just due to this uncertainty rather than representing a true effect, which would have been the case in the presence of more explanatory power from baseline measures. As a result, we particularly stress the importance of statistical significance as a check on interpretation of the results in this report.

Primary outcome

Writing attainment

To measure the primary outcome, we used the WAM (Dunsmuir et al., 2015; Murphy et al., 2013). The WAM was developed in order to create a valid and reliable writing assessment measure, relevant within the context of the English educational system. This measure is designed to assess narrative writing in response to a written prompt, to which pupils are given 15 minutes to write. Previous evidence suggests that this measure is reliable (test–retest correlation $r = 0.82$ over 21 days with different prompts) and valid ($r = 0.786$ with Wechsler Objective Reading Dimensions – WOLD – Written Expression subtest) (Dunsmuir et al., 2015). The WAM prompt presented to pupils is included as Appendix N. We carried out further analysis of the performance of the measure given its lack of widespread use up to this point, which are reported below.

The WAM is based on the structure and format of the WOLD Written Expression subtest, with modified dimensions that incorporate descriptors from the National Curriculum writing attainment targets, including the following seven dimensions: ideas development, organisation and planning, vocabulary, sentence structure and grammar, spelling, punctuation and handwriting. The pupil can receive a mark of 1–4 for each of these sub-scales, with 4 being the highest. The WAM is unique as an assessment because it incorporates 'ideas development', which fits well with the aims of all the interventions being tested as part of the Learning about Culture project. Note in the logic model for this intervention (Figure 1) that increased creativity is an outcome. In support of this, we double-weighted the score on the 'ideas development' dimension. Final scores range from zero to 32 (after accounting for double weighting).

Robust assessment of writing is challenging, particularly during primary schooling. However, the centrality of understanding improvements in writing and, hence, the need for this to be the primary outcome measure was stressed in the initial project outlines (noting that previous trials had generally focused on reading, rather than writing, adding to the rationale for funding these projects) and set up meeting discussions with the EEF and programme teams. Use of the WAM (Dunsmuir et al., 2015) as a measure of KS2 age pupils' writing was not the initial proposal for this trial, but emerged from discussions held as part of the project setup meetings. The WAM is an analytic measure of writing based on equal weighting of the following criteria: handwriting, spelling, punctuation, sentence structure and grammar, vocabulary, organisation and overall structure, and ideas. There are, of course, some limitations of its use, largely stemming from the fact that it is a relatively new measure and we would, ideally, prefer to have used a measure with a longer track record. Nevertheless, we believe that it is a pragmatic measure for the context of this research. Dunsmuir et al.'s (2015) results are encouraging in terms of the measure's internal consistency (Cronbach's alpha = 0.87), inter-rater reliability (Cohen's kappa > 0.7 for all sub-scales except 'ideas', where kappa = 0.62; kappa > 0.6 is generally considered satisfactory), and test–retest reliability ($r = 0.82$). To supplement this evidence, the EEF provided funding for us to undertake a small, informal piloting of the WAM in a single class.

The Behavioural Insights Team (BIT) conducted a small-scale pilot of the WAM in October 2017 with approximately 50 pupils from one Year 5 class, one Year 6 class and one mixed Year 5/6 class. The aim was to understand how clearly pupils understood the prompt, how much they were able to write during the time allotted, and to act as a sense check of the measurement properties reported by Dunsmuir et al. (2015). Pupils were given the WAM prompt, one sheet of A4 paper, and 15 minutes to complete the task. The results of the pilot showed that pupils had little difficulty in completing this writing task but required some additional clarification on the prompt, and additional paper. In addition, the measurement diagnostics remained encouraging (albeit this may be helped by the small sample).

While we were keen to maintain consistency between the WAM as implemented by Dunsmuir et al. and this work, in order to ensure that what we do know about the WAM from their work carries forward to our context we have made some small adaptations based on concerns identified during the setup process, and from observations arising from the small-scale pilot. These were discussed with the WAM's lead developer, whom we gratefully acknowledge as having provided helpful informal advice as part of this process and also as part of the marking process. Concerns from developers were noted and discussed with mitigating actions as follows:

- **Concerns that the WAM may over-weight surface features of the language.** We agreed that this was a potential concern, particularly within the context of arts-based learning evaluations, while noting that we thought alternatives (such as using KS2 grammar punctuation and spelling tests) would be much worse affected by this problem. In partial mitigation, we double-weighted the ideas sub-scale. This double-weighting results in an outcome distribution which is slightly less normally distributed but not to an excessive degree (see Figure E1 in Appendix E).

- **Concerns regarding the content and face validity of the WAM, given its alignment with an earlier version of the English national curriculum.** The overall aims for the teaching of writing in primary schools, that are specified in England's national curriculum implemented since 2014, require that teachers develop pupils' competence in 'transcription (spelling and handwriting)' and 'composition (articulating ideas and structuring them in speech and writing)' (DfE, 2013, p. 16). The WAM is an appropriate measure of writing in the context of England's current national curriculum aims.

- **Concerns that aspects of the prompt may be confusing to pupils.** These were identified from the pilot. Revisions were made to the introduction to pupils of the WAM to provide increased guidance to pupils on the purpose of the writing sample we ask them to produce, given that we understand this to be normal practice for pupils of this age when taking part in a writing activity (see the WAM prompt used in this trial in Appendix N).

- **Concerns that 15 minutes is not long enough.** The observations from the pilot suggested that 15 minutes was sufficient for pupils to produce a writing sample that could be meaningfully assessed. However, clearly this was only small scale. We took this concern seriously while we aimed to maintain broad alignment with previous use of the WAM. As such, in a change from the WAM as previously used, we provided five minutes of planning time at the start of the activity, in which pupils can make notes but don't begin the writing activity itself. This also helped to make the activity more familiar to pupils, in line with the previous point that the prompt may be confusing.

The writing tasks for this evaluation and the other two KS2 evaluations (Power of Pictures and Young Journalist Academy), that were part of the EEF/RSA Learning about Culture project, were invigilated and collected in summer 2019 by a team of research assistants (RAs) coordinated by BIT as a combined exercise. Since outcome data were collected as part of a single exercise and consistent (in terms of both measure and timing) across these three evaluations, we report our analysis of the data collection and measurement with pooled WAM data collected across the three projects; pooling these data allows us to increase sample size for these analyses and, hence, reduce noise and risk of small sample bias in estimates from these analyses. The same goes for consideration of the WSEM measure, which follows.

Data collection RAs were kept blinded to trial arm assignment of schools in order to avoid the potential for this to bias the outcome measurement (e.g., by being more lenient on timing in treatment schools). A separate group of 25 RAs (17 of whom marked tasks on the CoW project), also blind to trial arm assignment of schools (this time in order to avoid the potential for this to introduce bias into the trial, e.g., through unconsciously being more generous in their marking of pupils in the treatment group), marked the writing exercises against the WAM scoring sheet. This blinding is important in supporting the evaluation's internal validity. The WAM scoring sheet provided detailed criteria for assigning scores on each of the seven dimensions. During training, markers were provided with examples of pupil writing that exemplified each rating within a given dimension and an explanation of why that sample achieved the rating. The samples, scores and explanations were all provided by UCL academics who developed the WAM. A random sample of the tests (approximately 3%) were independently second marked by one of the other RAs during this process (a minimum of two tests per RA per day), with a correlation of 0.75 between markers in this double-marked sample. Where discrepancies arose, these were used to feed back to markers in order to improve the consistency of marking over the exercise as a

whole – this continuous improvement process may inflate this correlation over the course of the process, but improves the reliability of the marking relative to the alternative. As a point of reference, Dunsmuir et al. (2015) report an average marker-level intra-class correlation (ICC) of 0.97 for the WAM (range 0.93–0.99 at 95% CI).

Furthermore, analysis of the basic statistical properties of the overall measure are encouraging: the distribution of the scores is normal (skewness = –0.27; see Figure E1 in Appendix E), which suggests there were minimal issues with floor or ceiling effects, and we calculate a Cronbach's alpha across the seven marking sub-domains of 0.84, suggesting these cohere sensibly into the overall score.

Secondary outcomes

Writing 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, writing, which may in turn have an effect on their sense of efficacy as a writer. For this reason, we consider writing self-efficacy as our secondary outcome. In addition, self-efficacy has been highlighted in EEF's review of non-cognitive skills: the evidence 'indicate[s] that self-efficacy for a particular task is malleable and that improved self-efficacy is associated with greater persistence, interest, and performance' (Gutmann & Schoon, 2017, p.11) and that 'the best predictors of specific academic performance are self-efficacy beliefs regarding those specific academic domains (Pajares, 1996 in Gutmann & Schoon, 2017, p.11).

To measure writing self-efficacy, we used a writing self-efficacy measure (WSEM), which was adapted from the self-efficacy for writing scale (SEWS) measure proposed by Bruning et al. (2013), in order to make it suitable for primary school pupils through some simplification (see Appendix O for the full WSEM questionnaire used in this trial). Bruning et al.'s original measure involves 16 statements capturing aspects of writing, including 'I can think of many ideas for my writing' and 'I can avoid distractions while I write', with pupils giving scores 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, rather than scores out of 100. These adaptations were based on consultation with experts in primary literacy pedagogy, and were piloted through the same process as outlined for the WAM, above, with resulting refinements to wording of the Likert categories, as the initial versions were found to encourage pupils to exaggerate their confidence as this was seen as the 'right' answer. Bruning et al. (2013) developed a multi-factor model of writing self-efficacy; however, since the intervention is not hypothesised to have a link with the specific factors (with the exception of ideation, which we discuss separately below) we used a simple aggregate of self-assessments across all sixteen statements (all are positively framed so there is no need for reverse coding). As such, possible scores range from 16–80 for each child.

As with the WAM, these tasks were administered and collected in summer 2019 by a team of RAs, coordinated by BIT, who also marked the WSEM. RAs were kept blinded to trial arm assignment of schools. Again, consistent with the WAM, we explored the statistical properties of this measure (carried out pooled with WSEM data collected concurrently for the Power of Pictures and Young Journalist Academy projects), given the adaptations that were made in order to use it for this project. The overall scores are rather negatively skewed (skewness = –1.16; see also Figure E2 in Appendix E) which could attenuate impact estimates for this outcome; we estimate a Cronbach's alpha of 0.90 across the individual items of the measure, suggesting they cohere sensibly into the overall score.

Ideation

The logic model also identifies the potential for increased creativity in the pupils whose teachers have participated in their programme. To explore this, we reported differences in the 'ideation' sub-measure of the WSEM as an additional secondary outcome measure. This measure was jointly chosen with RSA and allows us to address the research question on pupils' perception of their capacity to generate ideas [secondary research question]. This uses the first five questions of the WSEM (see Appendix O for these statements) and, as such, possible scores range from 5–25 for each child. As this is a sub-domain of the WSEM measure as a whole, data collection details are as per that measure. Also, like the WSEM, the scores are rather negatively skewed (see Figure E3 in Appendix E) with potential consequences for attenuating impact estimates for this outcome.

Sample size

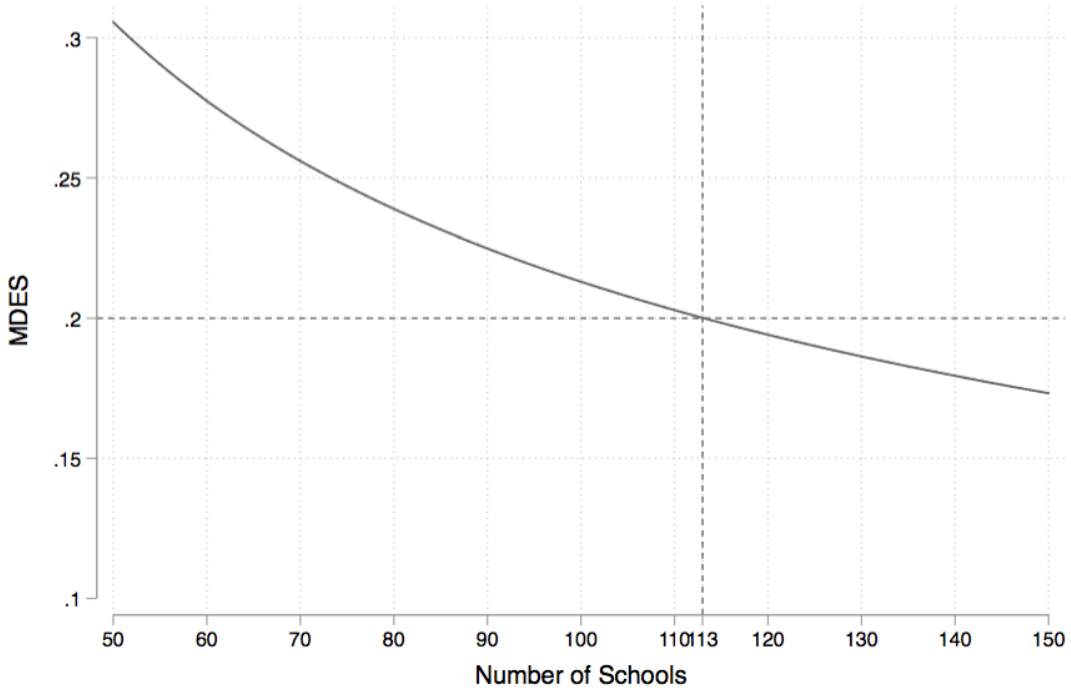
Sample size calculations were carried out for analysis of the WAM, since this is the primary outcome of interest, and were carried out with an initial baseline of achieving a minimum detectable effect size (MDES) of *d* = 0.20 with the following assumptions: power of 0.8 for a two-tailed 0.05 significance test, randomisation at school level, an ICC of 0.15[7] (EEF, 2015) and 25 pupils involved in the trial per school with 10 percent pupil-level attrition.

When estimating required effect sizes at the protocol stage an appropriate pre-test/post-test correlation assumption could not be estimated empirically directly for this trial, since correlation data between the planned baseline and outcome measures were unavailable. This is because the planned baseline (score in the Year 1 phonics screening check; consistent with EEF policy to use an administrative measure rather than an additional pre-test) has only been in place since 2012, and our primary outcome measure (the WAM) is an even newer measure. EEF guidance suggests that a pre-test/post-test correlation of 0.7 in education research is common (Torgerson & Torgerson, 2013); however, we see this as too optimistic in this case. The 21-day test–retest correlation coefficient of the WAM is reported to be 0.82 (Dunsmuir et al., 2015) but the time elapsed between the planned baseline and outcome measure in this trial would have been much longer, and it was never planned to use the WAM itself as a baseline. Our planned baseline measure (score in Year 1 phonics screening check) has less variance than would be ideal, due to a degree of bunching between the pass (32) and highest available mark (40). This bunching presumably occurs in order to push pupils over the pass line due to accountability concerns.

Nevertheless, given its closer temporal proximity to the outcome measure point, we expected (and continue to expect) that it would have been likely to explain more variance in our outcome than earlier measures available in the NPD (which would have to be measured at the Early Years Foundation Stage). While there is no direct measure of the correlation between the WAM and the phonics screening check available, a value of 0.52 was estimated using Year 1 phonics screening check scores and Progress in International Reading Literacy Study (PIRLS) scores (DfE, 2017) (taken in Year 5, the same year as the WAM was administered). Given the similar time period between baseline and outcome test administration, and the related domain, we used this estimate as likely to approximate the value that would have been expected in this trial. Based on this, our sample size calculations at the protocol stage assumed that 25% of post-test variance at both pupil- and school-level would be explained by the pre-test (equivalent to pre-test/post-test correlation of 0.5).

[7] EEF guidance on ICCs (EEF, 2015) is provided for NPD outcomes. In the absence of ICC data for our outcomes of interest we use this guidance, specifically for the reading fine points score, and, given uncertainty about the geographical spread of participating schools, we use the highest regional ICC (which happens to be Inner London), to the nearest two decimal places.

Figure 2: Minimum detectable effect size (MDES) estimate as a function of the number of schools at the design stage



These assumptions suggested a requirement of 113 schools to achieve an MDES of *d* = 0.2 (see Figure 2). Based on discussions between the evaluation team, the CoW team and the EEF at the setup meetings, a sample size of 96 schools was agreed. This number is due to the maximum figure of 16 participants per Arvon residential. The CoW team confirmed that recruitment of 96 schools and intervention delivery to 48 treatment schools were reasonable and achievable numbers, given their capacity. Conditional on the sample size of 96 schools and the assumptions outlined above, this trial was, hence, estimated to be able to detect an effect of 0.22 at the protocol stage. Based on an assumption that the FSM sub-group would be 15.3 percent of the total size of the sample (based on pupils aged 5–10 in data from DfE statistics (DfE, 2016) and ignoring that it may be higher if recruited schools are in more disadvantaged areas), and maintaining all other assumptions (which is likely to be a conservative approach, given the lower levels of within-group variation in this sub-group), the MDES for this group at the time of design was estimated to be approximately 0.33 standard deviations (SD).

Ultimately, 94 schools were fully recruited and randomised. As this was only a small reduction and no other assumptions for the power calculations were altered due to new information available at the point of randomisation, there was no change to the main sample MDES (at the level of two decimal places) at this point. However, there were a larger number of FSM pupils per school than expected (six pupils per school after attrition adjustment) and, as a result, the estimated MDES for this sub-group reduced from 0.33 to 0.27.

As noted above, there have been substantial changes to the analysis that we are ultimately able to carry out, due to data access issues with our planned baseline measures. These have substantial bearing on the assumptions that underpin estimations of MDES. Specifically, the change in baseline variables included in the analysis model means that the earlier assumptions about proportion of variance explained were optimistic (although, of course, this could still have been the case, even if we had been able to obtain the planned baseline data): our data suggest that with the substituted baseline measures pupil-level variance explained (R^2) is 0.02 and school-level variance explained is 0.26 (with overall variance explained being 0.07). Furthermore, our average cluster size is lower than anticipated at 21 due to testing challenges, the number of control schools 36 and intervention schools 39 (75 schools in total, with a treatment ratio of 0.52; schools lost to follow-up are discussed below), and the ICC is substantially higher than anticipated (0.15 under standard EEF assumptions) at 0.33. Altogether, this leads to an estimated MDES for the primary analysis of *d* = 0.34. This implies substantially reduced precision or, equivalently, more uncertainty (larger CI) around our treatment estimate

than would have been the case if our initial assumptions were met – this higher level of uncertainty should be borne in mind in interpreting the findings.

There were similar challenges for the FSM sub-group. Here the pupil-level variance explained is 0.01, the school-level variance explained is 0.12 (leading to an overall figure for variance explained of 0.02), the ICC is 0.30 and the average cluster size is four (the number of schools are as with the main sample exercise). This results in an estimated MDES for the FSM sub-group analysis of $d = 0.44$.

Randomisation

Schools recruited by the project team were randomly assigned by the evaluation team.

Stratification (referred to as blocking in the evaluation protocol) was used to attempt to improve cross-arm comparability of schools, to improve precision of estimates, and for the practical purpose of allowing schools that signed up early to receive their allocation sooner than they otherwise would (this was agreed as important with the project team because of the nature of the intervention requiring as much notice of allocation be given to teachers, given it requires activity outside of their normal working hours).

Eight strata were defined on the basis of:

- randomisation batch based on timing of sign-up (first 47 vs. last 47);
- class composition by EAL (high vs. low split at sample median);
- class composition by eligibility for FSM (high vs. low split at sample median).

This approach tried to ensure that our treatment and control groups are well balanced in terms of these characteristics, which are likely to be correlated with our outcome measures (EEF, 2018). High and low EAL and FSM in these definitions was defined as whether a school was above or below the sample median (calculated separately for each randomisation batch) in each case, to ensure that block sizes are approximately equal (which may not be the case if we used population, rather than sample, characteristics). Table D7 in Appendix D shows the number of treatment and control schools in each stratum for each randomisation batch.

Randomisation followed recruitment of schools, including the signing of Memoranda of Understanding (MoUs) and baseline data collection in March–April 2018. This randomisation process was conducted using a script run in Stata 14. Each of the two randomisation batches followed the process below:

1 The schools were stratified into four strata on the basis of proportion of FSM pupils (split across the median sample proportion) and proportion of EAL pupils (split across the median sample proportion).

2 Each school was assigned a randomly generated number (setting a stable seed for the random number generation).

3 The schools were sorted by block and random number.

4 Schools were assigned to the treatment arm and to the control arm in turn, meaning that even numbered strata would have 50 : 50 allocations to treatment and control, but odd numbered strata would have small differences in allocation.

The code that was used to operationalise these steps (including the stable seed to allow for replication of the process) is reported in Appendix F.

An important challenge that occurred during the recruitment and randomisation process was that a significant proportion of recruited schools stated that they could not confirm which class the self-nominated teacher would be teaching the following academic year, ahead of randomisation. As it is well-known that schools tend to become less responsive after randomisation (also potentially differentially since this is particularly true if allocated to the control group, which would then have posed a potential threat to internal validity), and also we needed to have pupil-level data FSM and EAL for

the randomisation process, we obviously could not accept collecting these class lists once this was confirmed (post-randomisation). Instead, an approach was agreed to protect the internal validity of the trial while allowing for this pragmatic consideration (which would otherwise have resulted in the trial having dramatically missed its recruitment target), in which 39 schools were permitted to submit full year group lists (subject to the usual ethics and data protection procedures having been carried out) as a backstop, with revised class lists obtained once teacher–class allocation was confirmed. There are obviously some trade-offs to taking this approach. Firstly, it could have affected the randomisation stratum a school was found to be in if the class composition differed substantially from the year group composition, which was not found to be the case. Secondly, there was the theoretical possibility that schools would alter their teacher–class allocation plans based on the randomisation outcomes; we judged this as extremely unlikely, given all the other considerations that schools face in such decisions – especially as these were schools that were unable to bring forward this allocation process ahead of randomisation as initially requested.

Statistical analysis

Primary analysis

Our primary analysis focused on the WAM score and was performed using Stata 15. All continuous variables were used in their 'raw' form (in line with EEF guidance) as there was no clear reason to transform the data.

An ordinary least squares (OLS) model was estimated in which our outcome variable was regressed on a treatment arm indicator, block indicators (based on proportion of the class eligible for FSM, proportion of the class identified as EAL, and whether the school was randomised as part of the first or second batch), and, in a deviation from the evaluation protocol, the following baseline variables:

1 indicators of FSM eligibility;

2 EAL status;

3 class-level FSM composition;

4 class-level EAL composition (further details below).

As noted by EEF guidance, in a model that does not account for clustering, when this is a feature introduced by the experimental design, 'the point estimates will be accurate, but the standard errors will be downward biased' (EEF, 2018, p.3). However, we accounted for the potential effects of the experimental design in this respect by calculating standard errors, taking into account clustering (Angrist & Pischke, 2008) at the school level, which allows for the correlation of pupil outcomes within schools. We prefer this to the use of a hierarchical linear model which makes additional assumptions about the school-level effects that may not be justified.[8] We also estimated randomisation–inference *p*-values accounting for the clustering and stratification of the design, which were consistent with those based on clustered sampling inference (see Table D4 in Appendix D). The estimated impacts are intention-to-treat (ITT) effects and are reported with 95% CI. ICCs are also reported (as required by the EEF; further information below).

In the evaluation protocol and SAP, we stated our intent to estimate the following model in order to estimate the ITT impact of the intervention:

$$Y_{ij} = \alpha + \beta_1 Treat_j + \beta_2 PreTest_{ij} + \gamma' X_j + \varepsilon_{ij},$$

where individual i is nested in school j, $Y_{ij}$is the WAM score, $PreTest_{ij}$ is the value of the phonics screening check score (using the NPD variable PHONICS_MARK) used as a pre-test, $Treat_j$ is our school-level treatment indicator, X_j are

[8] Use of a hierarchical linear model to account for the clustering of the data assumes that the school-level random effects are normally distributed and uncorrelated with the pupil-level error term.

vector of stratification variables and ε_{ij} is an error term. Standard errors are calculated allowing for clustering at school-level (j).

However, due to data access issues discussed above, we are unable to estimate this model due to the unavailability of PHONICS_MARK as $PreTest_{ij}$. Instead, we estimate the following model in which $PreTest_{ij}$ has been replaced with FSM eligibility, EAL status, class-level FSM composition and class-level EAL composition (as discussed above):

$$Y_{ij} = \alpha + \beta_1 Treat_j + \beta_2 FSM_{ij} + \beta_3 EAL_{ij} + \beta_4 FSMProp_j + \beta_5 EALProp_j + \gamma' X_j + \varepsilon_{ij},$$

where everything is as per the planned model, except that FSM_{ij} is whether individual i is eligible for FSM in the current school year and, similarly, EAL_{ij} is whether individual i is recorded as having EAL, while $FSMProp_j$ is the FSM composition of treated class in school j and $EALProp_j$ is the same for its EAL composition.

As such, this report's primary ITT estimate is recovered from the estimate of β_1 in this latter model, when it is estimated on the full sample at randomisation.

Note that while this model is a deviation from the evaluation protocol and SAP, it was planned and reported to the EEF ahead of analysis being carried out. The model has not been altered depending on the significance of any variables included (i.e., no variables were removed due to being statistically insignificant), including the vector of blocking variables (X_j). Syntax for this primary analysis model is reported in Appendix F.

Secondary analysis

We conducted two secondary outcome analyses:

Writing self-efficacy

Same as the revised primary outcome analysis except replacing Y_{ij} with the writing self-efficacy measure (WSEM) score. Note that this is a deviation from protocol, which stated that this would be the same as the planned primary outcome analysis except replacing Y_{ij} with the WSEM score and $PreTest_{ij}$ with assessment of pupils' personal, social and emotional development skills from the EYFSP (aggregated scores from NPD FSP_PSE_G06, FSP_PSE_G07 and FSP_PSE_G08). This change was made due to data access problems rendering these NPD variables unavailable.

Ideation

Same as the revised primary outcome analysis except replacing Y_{ij} with the Ideation sub-score from the WSEM. Note that this is a deviation from protocol, which stated that this would be the same as the planned primary outcome analysis except replacing Y_{ij} with the ideation sub-score from the WSEM and $PreTest_{ij}$ with assessment of pupils' personal, social and emotional development skills from the EYFSP (aggregated scores from NPD FSP_PSE_G06, FSP_PSE_G07 and FSP_PSE_G08). This change was made due to data access problems rendering these NPD variables unavailable.

It was also planned, potentially as part of a separate report, to estimate the impact on KS2 grammar, punctuation and spelling test attainment. Unfortunately, the relevant KS2 national curriculum tests in summer 2020 that would have collected these data were cancelled as a result of the COVID-19 pandemic. As such, this medium-term follow-up is no longer possible.

Analysis in the presence of non-compliance

The following criteria have been defined in the trial protocol as variables that can be used to assess compliance with the intervention. This draws principally on attendance data collected from the project team. Compliance was measured at the teacher level, which reflects the intervention delivery method. A school is considered to have complied *if and only if* the following two conditions were met:

- participating teacher attended both residential weekend workshops;

- participating teacher attended at least two out of the three CPD sessions.

We used complier average causal effect (CACE) analysis (Gerber & Green, 2012) to estimate intervention effects on treated children. We estimated the CACE using two-stage least squares (2SLS) regression by estimating a (first stage) model of compliance, as follows:

$$Comply_j = \alpha + \beta_1 Treat_j + \gamma' X_j + \xi_{ij},$$

where $Comply_j$ is the binary compliance variable defined above and ξ_{ij} is an error term. The predicted values of $Comply_j$ from the first stage are used in the estimation of a (structural) model of our outcome measure Y_{ij}. In other respects, the specification remains the same as the revised primary outcome ITT model. This second stage model is specified as follows:

$$Y_{ij} = \alpha + \beta_1 \widehat{Comply}_j + \beta_2 PreTest_{ij} + \gamma' X_j + \omega_{ij},$$

where $\widehat{Comply}_j$ are the predicted values of treatment receipt derived from the first stage model and ω_{ij} is an error term. Our primary outcome of interest is β_1, which should recover the effect of the intervention among compliers.

We conducted this analysis using the 'ivregress' functionality of Stata to make necessary adjustments to standard errors (which have also been clustered at school level) due to the instrumental variables approach. We note the deviation to protocol due to these models being based on the revised primary outcome ITT model, rather than the planned primary outcome ITT model, which is for the same underlying reasons of data access.

Missing data analysis

We describe and summarise the extent of missing data in the primary and secondary outcomes, and in the model associated with the analysis. Reasons for missing data are also described.

For all models we planned to implement a missing data strategy if more than 5% of data in the model was missing or if more than 10% of data for a single school was missing. The strategy would be followed separately for each instance of model and variable for which the threshold was exceeded.

We first assessed whether the missing data was missing at random (MAR), since this is a prerequisite for missing data modelling to produce meaningful results. To do this we created an indicator variable for each variable in the impact model, specifying whether the data was missing or not. We then used logistic regression to test whether this missing status could be predicted from the following variables: all variables in the analysis model plus eligibility for FSM (and proportion eligible for FSM in the school) and EAL status (and proportion EAL in the school). Where predictability was confirmed, we proceeded to the appropriate next step of this strategy.

For situations for which the MAR assumption appeared to hold and only the outcome variable in the model was missing, we re-estimated the treatment effect using our pre-specified model with the addition of the covariates found to be statistically significantly predictive of missingness of the outcome.

For situations for which the MAR assumption appeared to hold and any variable other than the outcome variable in the model was missing, we used all variables in the analysis model plus eligibility for FSM (and proportion eligible for FSM in the school) and EAL status (and proportion EAL in the school) to estimate a multiple imputation (MI) model using a fully conditional specification, implemented using Stata MI to create 20 imputed data sets. We re-estimated the treatment effect using each data set, and took the average and estimated standard error using Rubin's combination rules (Rubin, 2004).

Analysis that is altered following the missing data strategy (either on a multiply imputed data set or with additional variables) would only ever be viewed as a sensitivity analysis. As such, the main estimates of the effectiveness of the treatment are derived from complete case analysis only. However, the sensitivity of the estimates to missingness would be assessed by comparing this main analysis with those altered following the missing data strategy. For example, if the complete case analysis model were to imply effectiveness but the imputed estimate did not, we would assume that the

missing data is missing systematically to such an extent as to invalidate the initial conclusion of effectiveness, which would then be stated in the reporting of the evaluation.

Sub-group analyses

Following EEF guidance, we first tested for an interaction of the treatment and FSM status. This was originally planned to be carried out using the NPD variable EVERFSM_6_P (in line with EEF guidance) and the following model:

$$Y_{ij} = \alpha + \beta_1 Treat_j + \beta_2 FSMEver_{ij} + \beta_3 Treat_j * FSMEver_{ij} + \beta_4 PreTest_{ij} + \gamma' X_j + \nu_{ij},$$

where individual i is nested in school j, $Y_{ij}$is the WAM score, $PreTest_{ij}$ is the value of the phonics screening check score used as a pre-test, $Treat_j$ is our school-level treatment indicator, $FSMEver_{ij}$ is an indicator of FSM eligibility (EVERFSM_6_P), $Treat_j * FSMEver_{ij}$ is an interaction between these two terms, X_j is a vector of stratification variables and ν_{ij} is an error term.

However, in a deviation from protocol, the NPD derived indicator of FSM eligibility (EVERFSM_6_P) was unavailable, as was the phonics screening check score (for data access reasons, as discussed above), so our model was adapted as follows:

$$Y_{ij} = \alpha + \beta_1 Treat_j + \beta_2 FSM_{ij} + \beta_3 Treat_j * FSM_{ij} + \beta_4 EAL_{ij} + FSMProp_j + EALProp_j + \gamma' X_j + \nu_{ij},$$

where all terms are defined as per the planned FSM interaction model above or the planned primary analysis model but where we use school-supplied FSM indicators for whether pupils were FSM-eligible in the current school year.

Standard errors have been calculated allowing for clustering at school-level (j). Syntax for this interaction model is reported in Appendix F.

If a significant interaction was found (i.e., the absolute value of the point estimate of β_3 divided by the school-level clustered standard error is greater than 1.96), then we would proceed to conduct a specific sub-group analysis for those who are identified by schools as eligible for FSM ahead of randomisation, using the same model as our revised primary analysis. We note that this is also a deviation from protocol as we would have defined this sub-group as those who have ever been registered for FSM in the NPD (identified using the variable EVERFSM_6_P), and used our planned primary analysis model.

But for the small difference in definition of FSM (which investigation with the DfE suggests is unlikely to be material as no cleaning of the data submitted by schools is carried out before it is made available in the NPD), this sub-group was identified in the trial protocol and FSM pupils are a key subgroup to be analysed in all EEF trials.

This FSM analysis was conducted for both the primary and secondary outcomes.

Additional analyses and robustness checks

No additional analyses were planned as part of the project's SAP. All additional analyses and robustness checks carried out should be considered exploratory only.

In addition to carrying out inference through school-level clustered standard errors, we also estimated randomisation inference *p*-values, in order to check the robustness of inference to this approach. As this was not planned in the SAP, this should be considered exploratory and will not be used to guide interpretation of the results as a result. However, it will provide useful information on the extent to which there is variation between these different approaches to statistical inference.

We ran three exploratory robustness check models based on potential issues identified in the course of analysis:

- Due to delays in the testing of some schools, we ran a robustness check model in which we replicated the primary analysis model but added a control for the number days between the date that the first school was

tested and the date that the school in question was tested. The logic for this is that delays to testing could have affected the dosage of the intervention.

- It is possible that variation in the approach of different markers (e.g., a degree of leniency by some markers, despite the steps documented above to maximise consistency) who marked the WAM (primary outcome measure) could affect the treatment estimate. While marking was blind to treatment assignment, the relatively small number of markers could lead to imbalance in such approaches by chance. As such, we ran a robustness check model in which we replicated the primary analysis model but added marker fixed effects.

- Imbalance in school level KS1 scores between treatment and control group schools in the analysis sample is identified as part of our balance checks. To check whether such imbalance might explain our findings, we ran a robustness check model which replicates the primary analysis model but includes average school level KS1 scores as an additional covariate. As it was not possible to link in average KS1 scores for all schools, MI was used for this analysis, carrying out 20 regression-based imputations of average KS1 score using all covariates in the primary analysis model.

Estimation of effect sizes

Hedges' *g* effect size was calculated as set out by Hedges (1981):

$$g = J(n_1 + n_2 + 2)\frac{\underline{\underline{x_1}} - \underline{\underline{x_2}}}{\widehat{s*}},$$

where our conditional estimate of $\underline{\underline{x_1}} - \underline{\underline{x_2}}$ is recovered from $\beta_1$in the primary ITT analysis model. $\widehat{s*}$ is estimated from the analysis sample as follows:

$$\widehat{s*} = \sqrt{\frac{(n_1-1){s_1}^2 + (n_2-1){s_2}^2}{n_1+n_2-2}},$$

where n_1 is the sample size in the control group, n_2 is the sample size in the treatment group, s_1 is the standard deviation (SD) of the control group and s_2 is the SD of the treatment group (all estimates of SD used are unconditional, in line with the EEF's analysis guidance to maximise comparability with other trials. $J(n_1 + n_2 + 2)$ is calculated as follows:

$$J(n_1 + n_2 + 2) = \frac{\Gamma\left(\frac{n_1+n_2+2}{2}\right)}{\sqrt{\frac{n_1+n_2+2}{2}}\Gamma\left(\frac{n_1+n_2+2-1}{2}\right)},$$

where n_1 is the sample size in the control group and n_2 is the sample size in the treatment group, or if calculating $J(n_1 + n_2 + 2)$ proves computationally intractable[9] using the above method, we instead fall back on the following approximation:

$$J(n_1 + n_2 + 2) \approx \left(1 - \frac{3}{4(n_1+n_2)-9}\right).$$

Ninety-five percent confidence intervals (95% CI) of the effect size have been estimated by inputting the upper and lower confidence limits of $\widehat{\beta_1}$ from the regression model into the effect size formula.Estimation of intra-cluster correlation (ICC)

In order to estimate the ICC of the outcome measure at school-level we employed an empty variance components model, as follows:

[9] The output of the gamma (γ) function in the Hedges' *g* correction factor (*J*) becomes large quickly, making this method of computation intractable where $n_1 + n_2$ is not small. As such, it can quickly become intractable. Thankfully, the approximate method tends towards the full correction factor quickly. As such, where the computational intractability is an issue, the approximate method is appropriate. In any event, the correction factor is likely to be small in this trial.

$$Y_{ij} = \alpha + \eta_j + \varepsilon_{ij},$$

where individual i is nested in school j, $Y_{ij}$is the WAM score, η_j is a school-level random effect and ε_{ij} is an individual-level error term. The school-level random effect is assumed to be normally distributed and uncorrelated with the individual-level errors. An empty variance components model is used to facilitate comparability between trials (and in line with EEF guidance).

The ICC itself was then estimated from this model using the following equation:

$$\rho = \frac{var(\eta_j)}{var(\eta_j)+var(\varepsilon_{ij})}.$$

In the SAP, we also intended to estimate the ICC of the planned baseline measure, the phonics screening check score. In a deviation from this plan for reasons of data access, as discussed above, this analysis was not conducted.

Longitudinal analysis

We had planned to estimate the effect of the intervention on pupil performance in nationwide KS2 national curriculum tests in English grammar, punctuation and spelling. Unfortunately, the relevant KS2 national curriculum tests in summer 2020 that would have collected these data were cancelled as a result of the COVID-19 pandemic. As such, this medium-term follow-up is no longer possible. It may ultimately be possible to consider analysing impacts on GCSE results (e.g., GCSE English, taken in 2025); however, planning for this is beyond the scope of this report.

Implementation and process evaluation (IPE)

As part of the mixed method design of this evaluation, an IPE was conducted to complement the findings from the impact evaluation. The IPE involved members of the team with expertise and knowledge of the arts and education, which they fed into the design, conduct and analysis of the IPE. This section describes the IPE aims and approach, sampling and recruitment, data collection and analysis methods used.

IPE aims and approach

The purpose of the IPE was to understand how teachers implemented the CoW intervention in the classroom and, in particular, to better understand the barriers and facilitators to implementation and delivering the intervention with fidelity. The IPE was also used to understand how the teaching of writing through CoW compared to teachers' usual practice, and to determine the cost to schools of delivering the intervention.

Multiple sources of data were triangulated to address the IPE RQs. The primary approach to IPE data collection consisted of collecting case study data from six schools in the intervention-arm of the study. Each case study involved the following methods:

- a semi-structured 30-minute interview with the teacher involved in delivering the intervention;
- a semi-structured 20-minute interview with a member of the SLT;
- an observation of an English lesson delivered by the interviewed teacher (40–60 minutes, depending on length of lesson).

A survey was also sent to all intervention and control group schools, and administrative data were collected. In addition, one residential training weekend (June 2018) and two CPD training sessions (September and November 2018) were observed. Intervention manuals and guidance were also reviewed to inform interviews and data analysis.

The research questions (RQs) and the data collection methods used to address them are shown in Table 3.

Table 3: Implementation and process evaluation (IPE) methods overview

Data collection method (case study)	Data analysis method	Participant groups	Target number of participants per case study	Actual number of participants / activities per case study	Total number of participants / activities	Research questions addressed
Semi-structured Interviews	Framework approach	SLT	1	1	6	RQ1 (Fidelity); RQ5 (Training transfer to practice), RQ6 (Teacher confidence); RQ7 (Mechanisms); RQ8 (Revision practice)
	Framework approach	Teachers	1	1	6	RQ1 (Fidelity); RQ2 (Responsiveness), RQ3 (Quality), RQ5 (Training transfer to practice), RQ6 (Teacher confidence), RQ7 (Mechanisms), RQ8 (Revision practice)
Observations	Framework approach	Pupils, teachers	1	1	6	RQ5; RQ1 (Fidelity); RQ5 (Training transfer to practice); RQ7 (Mechanisms); RQ8 (Revision practice)
Data collection method	Data analysis method	Participant groups	Target number of participants/activities	Actual number of participants / activities		Research questions addressed
Observations	Framework approach	Trainers, Teachers	1 observation of residential training, 1 observation of midpoint CPD training	1 observation of residential training, 2 observations of midpoint CPD training		RQ1 (Fidelity), RQ5 (Training transfer to practice)
Baseline survey	Descriptive statistics	Treatment teachers	47	19		RQ1 (Fidelity), RQ5 (Confidence)
Follow-up surveys	Descriptive statistics	Teachers	47 treatment, 47 control	47 treatment, 49 control (prior to data cleaning); 36 treatment, 37 control (post-cleaning)[10]		RQ1 (Fidelity), RQ3 (Quality), RQ5 (Training transfer to practice), RQ6 (Teacher confidence), RQ7 (Mechanisms)
Cost interview	Framework approach	Teachers	4	4		Cost

[10] Some schools submitted more than one survey; see *IPE data analysis* section for more detail on data cleaning and reconciliation.

IPE sampling and recruitment

Case studies

Table 4 sets out the characteristics of the six case study schools that were recruited. Case study schools were selected using a purposive sampling approach to capture the range of Year 5 classes that had received the CoW intervention during the 2018/2019 academic year. The primary sampling criteria were:

1. level of attendance at training (defined as high where a teacher from the school had attended the two CPD and two residential sessions run when the case studies were undertaken, and low where the school's teacher had missed one or more of these training sessions);
2. the proportion of pupils receiving FSM (defined as high or low, depending on whether the school's proportion of pupils receiving FSM was above or below the median the median for all CoW intervention schools, which was 18.5%; information obtained from UK Government, 2019).

The secondary sampling criteria were:

1. geographical region (categorised as East Midlands, Greater London, North East, North West, South West or Yorkshire and the Humber);
2. Ofsted rating (recorded as Outstanding, Good or Requires improvement [information obtained from UK Government, 2019]);
3. training group (teachers were placed in group A, B or C).

There was some difficulty with recruiting case study schools. Of the first six schools sampled, four declined to participate for reasons such as staff shortages. Two further schools also declined to participate from the second round of sampling. As a result, the achieved sample included schools with higher levels of training attendance than was planned.

Table 4: Characteristics of case study schools

Characteristic	Number
Attendance at training	
Attended all residential and CPD training sessions	4
Missed one CPD session	2
Free school meals (FSM) rate	
0–5%	2
6–10%	1
11–15%	0
16–20%	0
21–25%	2
26–30%	1

Geographical region	
North West	4
Yorkshire and the Humber	1
East Midlands	1
Ofsted rating	
Outstanding	2
Good	3
Requires improvement	1
Training group	
A	2
B	2
C	2

Surveys

Baseline and follow-up surveys were administered to schools taking part in the trial. In intervention schools, the survey was to be completed by the teacher who had been involved in the CoW intervention, while in control schools, we asked that a Year 5 teacher complete the survey. The baseline survey was distributed to schools randomised to the intervention-arm of the trial in October 2018. Unfortunately, the response rate (calculated using the number of schools originally randomised to the intervention arm of the trial) was very low, with 19 respondents (40%), so it did not produce meaningful data. Instead, questions from the baseline survey were incorporated into the follow-up survey.

For the follow-up survey, all schools involved in the trial were invited to respond. Depending on which arm of the trial they were randomised to, schools were sent either an intervention survey or a control survey. Because we are not conducting any statistical inference with the survey results, CI are not given for response frequencies (see Appendix J and Appendix K); in addition, as the schools taking part in the trial were not a representative sample of all primary schools in the UK, it is not appropriate to generalise the findings beyond this group of schools.

IPE data collection

Case studies

Sampled schools were contacted by email and, where schools agreed to take part, a date was arranged for a researcher to visit. The researchers did not inform the delivery team about which schools they intended to visit. All visits took place between June and July 2019. At the visits, the following data were collected: a semi-structured, audio-recorded interview with:

- the Year 5 teacher involved in delivering the intervention

- a member of the school's SLT.

There was also an observation of a writing lesson delivered by the interviewed teacher (see Appendix I).

The exception to this was one school where the Year 5 class teacher was also a member of the school's SLT and no other member of the SLT had been involved in the intervention.

It was also set out in the protocol that interviews would be conducted with teachers both before and after the observation. The first interview was intended to focus on their broader experience of the programme; the second interview was intended to discuss the observation. However, when arranging interviews, teachers expressed concern at leaving their classroom twice or had a timetable that did not support conducting interviews before and after instruction; instead it was agreed with teachers that they would attend one summative interview. Given the design of the interview schedule, this change did not pose any risk to the quality of the data being collected.

The interviews were conducted using guides that focused on exploring the following:

- usual practice in relation to teaching writing;
- the context in which the intervention was implemented;
- the facilitators and barriers to implementation;
- children's engagement in intervention lessons;
- the perceived impact of the intervention;
- the mechanisms underlying this change.

Interviewees were informed that the interview was anonymous, that they could withdraw at any time and that they did not have to answer any questions they did not want to answer. Full interview guides can be found in Appendix G.

Observations of teaching focused on the teacher's approach to delivery; the content of the lesson, including use of CoW principles; and children's engagement with the lesson. The observations proforma was developed collaboratively in partnership with the UCL team, who brought their respective subject matter expertise. Specifically, Dominic Wyse has writing, music and mixed-methods research expertise; Gemma Moss has expertise in early literacy development; and Andrew Burn has expertise in English, media and drama. Observation notes were recorded first in field notes, then transferred to the structured proforma (which can be found in Appendix I). The observation data were used to help researchers probe effectively during the interviews and to deepen understanding of observed practice. Strategies to reduce bias were implemented, including having two researchers (e.g., a BIT and UCL team member) conducting the initial observations, discussing and agreeing upon the final observation notes that were recorded; one of those researchers continued to conduct the subsequent observations in the other case study schools.

Surveys

The RQs and programme logic model were used to inform the design of the baseline survey. Data from interviews, as well as feedback from the delivery team, were used to inform the design of follow-up surveys. The survey administered to intervention schools covered:

- barriers to attending training sessions;
- translation of the training into classroom practice;
- confidence in delivering CoW-style lessons;
- the perceived impact and quality of the intervention.

The control survey focused on understanding usual practice in terms of teaching writing. Full details on the intervention and control group surveys can be found in Appendix J and Appendix K.

A link to the online survey platform SmartSurvey was sent to the key contacts at each school in June 2019. For the intervention survey, the email stated that the survey should be completed by the teacher who had been involved in the CoW intervention, while the control survey email stated that a Year 5 teacher should complete the survey. Reminders were sent to schools who did not initially complete the survey. Those schools who did not respond to reminders were followed up with a phone call and given the opportunity to complete the survey over the phone. All data was collected by the end of July 2019.

Administrative data

Data were collected on the number of residential training sessions attended (out of a total of two) and the number of CPD sessions attended (out of a total of three) by each school, together with reasons for non-attendance (where provided by the school). These data were sent by the delivery partner to the evaluation team once in March 2019 (at which point one CPD session was yet to take place) to inform case study sampling. At the end of the trial, this data was provided again, when a complete data set was available, to measure compliance.

IPE data analysis

Case studies

Verbatim transcripts of the interviews and notes from the observations were analysed using the Framework approach (Ritchie, et al., 2013). Firstly, emerging themes were identified through familiarisation with the data. The analytical framework was then created using a series of matrices in Excel, each relating to an emergent theme (see Appendix R). The columns in each matrix represented the key sub-themes drawn from the findings, and the rows represented individual participants interviewed or schools observed.

The interview and observation data were then summarised in the appropriate cells, which meant that all data relevant to a particular theme were noted, ordered and made accessible, facilitating a systematic approach to analysis that was grounded in participants' and schools' accounts, as well as the observers' accounts. Analysis involved working through the charted data to draw out the range of schools' experiences and participants' views, and identifying similarities, differences and links between them. Thematic analysis (undertaken by looking down the theme-based columns in the Framework) identified concepts and themes, and the case-based analysis (undertaken by comparing and contrasting rows in the Framework), enabled links within cases to be established and cases compared and contrasted with each other.

During the analytical process a balance was maintained between deduction (using existing knowledge and the RQs to guide the analysis) and induction (allowing concepts and ways of interpreting experience to emerge from the data). The *IPE results* section is organised based on the identified themes and sub-themes, which are outlined and described in their respective subsections. Multiple strategies were employed by the researchers to increase the credibility (i.e., accurate representation of the data), transferability (i.e., potential to apply the findings to other settings), dependability (i.e., traceable, logical analytical process) and confirmability (i.e., being grounded and traceable to the raw data) of the findings, with the ultimate aim of reducing bias during the analytical process (Hannes, 2011).

Firstly, in terms of striving to increase credibility, the researchers conducted peer debriefing meetings with the senior qualitative research lead (Matt Barnard, Head of Evaluation at BIT) and qualitative researchers who were not directly involved in the data collection or analysis process for the respective intervention. In addition, in accordance with the chosen approach to data analysis, the researchers focused on describing range and diversity, including the noting of any disconfirming cases. Verbatim participant quotations are used to provide evidence and exemplify the theme(s) discussed in the paragraph before the quotation. Quotations were selected by the qualitative researchers who conducted the data analysis, by considering multiple factors including how well they exemplify the theme(s) discussed. The researchers also sought to ensure that the quotations used in the *IPE results* section capture the variation in terms of points of view and experiences, as well as types of participant being interviewed (e.g., SLT, teachers) and the associated schools. Further, as qualitative data can only be generalised in terms of range and diversity, and not in terms of

prevalence, the analytical outputs focus on the nature of experiences, avoiding numerical summaries or language such as 'most' and 'majority'.

Secondly, to increase the potential for transferability and assessment of applicability to other contexts, the *Context* subsection in the *IPE results* describes the key details of the case study schools and the selection criteria. The findings also include descriptions of the co-facilitators (e.g., number of years of teaching and/or drama experience) and important contextual details about the case study schools (e.g., motivations for choosing to implement the intervention).

Thirdly, to increase dependability and confirmability of the findings, the researchers maintained a detailed audit trail and triangulated the data by comparing the findings from multiple types of participants (e.g., teachers, DPs) and sources of data (e.g., survey, interviews, observations). Researchers adhered to the key principles of the Framework approach, which includes ensuring that data management and analysis is systematic, comprehensive, transparent and grounded in the participants' accounts. Doing this was facilitated by the creation of the series of matrices in Excel that contained descriptive summaries of data that can be easily traced back to the verbatim quote on the relevant page of the transcript being described.

Surveys

Follow-up data were first cleaned by ensuring that all responses received came from a school in the relevant arm of the trial. Following this, data were checked to identify schools that had returned two or more survey responses. In these instances, the most recent survey completed by a respondent who identified as a 'teacher' was used for analysis. For instance, if there were two responses both from teachers, then the most recent one was kept, whereas if one response was from a member of the SLT and one was from a teacher, the SLT response was removed even if this was more recent. Data from teachers were prioritised where there were multiple responses from the same school, because it was assumed that the class teacher would have been more involved in the delivery of the intervention (for the intervention survey) and know more about usual classroom practice (for the control survey), and therefore, their views and experiences were most relevant. Data were also removed from three schools that were randomised to the intervention group but had not attended any of the residentials or CPD sessions, as they would not be able to provide information about the intervention. Due to low response rate, baseline survey data were not analysed.

Prior to cleaning the survey data set, there were 47 intervention school responses and 49 control school responses. Following cleaning, there were 36 responses from intervention schools (out of 47 schools randomised to the intervention group) and 37 responses from control schools (out of 47 schools randomised to the control group), giving a response rate for intervention schools of 77% and for control schools of 79%. Stata (version 14) was used to conduct descriptive analyses of the data. Percentages scores are reported, where relevant, in the *IPE results* section. Complete survey findings are provided in Appendix J and Appendix K.

Costs

The evaluation gathered three key categories of data:

1 direct marginal costs (which forms the basis of the cost per pupil);

2 pre-requisites (which is reported separately from the cost per pupil);

3 school staff time.

The data was gathered in two ways. Firstly, the evaluators requested from delivery partners information on how much they charged schools for delivering the intervention as part of the evaluation, and how much they will charge schools in the future excluding any funding or subsidy that is associated with delivering the intervention, as part of this evaluation. The latter data is used in calculating the cost per pupil; the former data is to ensure there is clarity about the precise nature of the data that is being requested and transparency of the approach.

The second mode of data collection was the use of case study interviews, as specified in the protocol. IPE interviews were used to determine whether questions about costs would be included in surveys, with the decision taking into account survey length and risk of damaging response rates. Based on low responses to the baseline survey, it was

judged appropriate to omit cost-related questions, so as to keep survey length down and not potentially dampen response rates. Instead, costs were further explored through interviews, which were also deemed a more appropriate method to gather detailed data, as they allow for follow-up questions to clarify responses and probe for more information.

The evaluators felt the programme cost was best estimated by having a good sense of the range and diversity of costs incurred by schools, which is facilitated by using a case study approach supported by purposive sampling (Ritchie et al., 2013). Cost case study schools were selected from those who had good engagement with the programme, as they were more likely to give the best indication of the resources needed to implement the programme fully; including schools with little engagement was likely to artificially deflate costs. As an indicator of this, we selected schools from the pool who had completed the outcome data collection (as fidelity data was not available at that stage).

The resources required to deliver the intervention were most influenced by staff time and any related marginal costs (such as travel and subsistence). We assumed these things were most likely to be related to the amount a school has to spend per pupil and the nature of the local area and school population, the most relevant indicator for which is the percentage of pupils eligible for FSM. That is, school spending on the programme was likely related to school financial resources. Therefore, we purposively selected one school in each of the following four categories:

- School with percentage of FSM in top half of participating schools and spend per pupil in top half;
- School with percentage of FSM in top half of participating schools and spend per pupil in bottom half;
- School with percentage of FSM in bottom half of participating schools and spend per pupil in top half;
- School with percentage of FSM in bottom half of participating schools and spend per pupil in bottom half.

The case study data was collected by RAs employed and trained by BIT. The RAs conducted interviews with teachers via telephone using a structured interview guide designed by BIT for this purpose. RAs estimated it took approximately 20 minutes to complete the discussion. Teachers were asked to report on direct costs of the intervention to the school, materials purchased, travel and subsistence, the cost of covering staff at training and the cost of any new physical materials purchased to improve the classroom environment. Teachers also reported on time spent embedding the intervention in their school, time at training, as well as time spent preparing to deliver the intervention. Staff were also asked to report on time taken to organise supply cover and the amount of supply cover. Data from these interviews were used to calculate the financial and time costs outlined in this report.

Timeline

Table 5: Timeline

(a) Overall evaluation timeline

Dates	Activity	Staff responsible / leading
June 2015	Participants sat Year 1 phonics screening check, which was planned to be used as baseline measure.	N/A
October 2017–February 2018	Recruitment: The CoW team began recruitment halfway through the autumn term 2017 using its existing network of schools and continued into early 2018.	CoW team
October 2017–February 2018	Pre-randomisation data collection.	BIT and CoW team
1 March 2018	Randomisation of first batch.	UCL
17 April 2018	Randomisation of second batch.	UCL
June 2018	First retreat for intervention group.	CoW team
September 2018–July 2019	Intervention in schools.	CoW team
September 2018–July 2019	Implementation and process evaluation (IPE) fieldwork (see further detail on IPE timeline below).	BIT and UCL
May–July 2019	Outcome testing: Pupils' writing and self-efficacy outcomes were collected by BIT. These tasks were then marked by PGCE students at UCL in a process overseen by BIT.	BIT and UCL
July–September 2019	Collation and cleaning of outcomes and compliance data in readiness for upload to ONS SRS for linkage with DfE National Pupil Database (NPD) extract.	UCL and BIT
January–July 2020	Project paused awaiting conclusion of data sharing agreement necessary for upload of project data to the ONS SRS for linkage with NPD.	
August–October 2020	Project resumed with planned revisions as a result of delays in achieving data sharing agreement. Impact analysis and report writing. UCL led on the data analysis with agreed deviations from published statistical analysis plan (SAP).	UCL and BIT

(b) IPE timeline

Date	Activity
Autumn term 2018	Observed training
	Collected baseline survey to measure school buy-in and teacher attitude towards intervention
	Collected school characteristics
Spring term 2019	Observed mid-point training
	Conducted in-school case studies
	Collected fidelity data to inform case study sampling
	Finalised sampling strategy
Summer term 2019	Conducted in-school case studies
	Administered end of intervention survey
	Conducted analysis

Impact evaluation

Summary

- There was no significant impact of participating in the CoW on writing attainment (WAM).
- There was no significant impact of participating in the CoW on writing self-efficacy (WSEM) or on ideation.
- There was no differential effect of the intervention on the primary outcome measure (WAM) or on writing self-efficacy (WSEM) for FSM-eligible pupils
- Participating in the CoW had a positive and statistically significant impact on the FSM-eligible pupils' ideation (equivalent to three months' progress). This provides some indication that the intervention may be beneficial for increasing the creativity of children from disadvantaged backgrounds.

Participant flow including losses and exclusions

The flow of participants is detailed in Figure 3. Of the original 3285 schools that were approached, 100 schools agreed to participate in the trial and met the eligibility criteria (3%). Six schools dropped out prior to randomisation. The remaining 94 schools were randomly allocated to the intervention and control groups using a stratified randomisation, as described above. At randomisation, 1273 pupils were allocated to the intervention and 1331 to control.[11]

Figure 3 shows that there was a substantial number of schools and pupils who could not be reached for follow-up (eight schools and 400 pupils in the treatment arm, and 11 schools and 507 pupils in the control arm). As is the case in many trials, it was easier to maintain contact with treatment schools and the developers were able to help ensure that fewer treatment schools dropped out than control schools. Nevertheless, primary data collection can be challenging, and this was the case in this evaluation. A total of 873 treatment pupils in 39 schools and 824 control pupils in 36 schools were analysed for this report.

The MDES estimated at various points of the trial is reported in Table 6. This was 0.22 at the design stage and remained at this level at randomisation. It increased to 0.34 at the analysis stage due to a combination of being unable to access the planned baseline measure (discussed further above), attrition rates and a higher ICC of the outcome measure than was anticipated based on previous EEF guidance.

[11] It should be noted that these numbers include pupils from schools that submitted entire year group lists instead of class lists prior to randomisation. As outlined in the SAP and discussed further in the *Randomisation* section, 41% of schools did not provide a class list prior to randomisation. As such, all pupils were included in the randomisation process and class lists were obtained once the intervention had begun. The randomisation numbers reported in the SAP were adjusted to assume that on average 26 pupils would participate in the intervention from each school that did not submit a class list. This assumption was made on the basis of the average number of pupils per class from the 59% of schools that submitted class lists.

Figure 3: Participant flow diagram

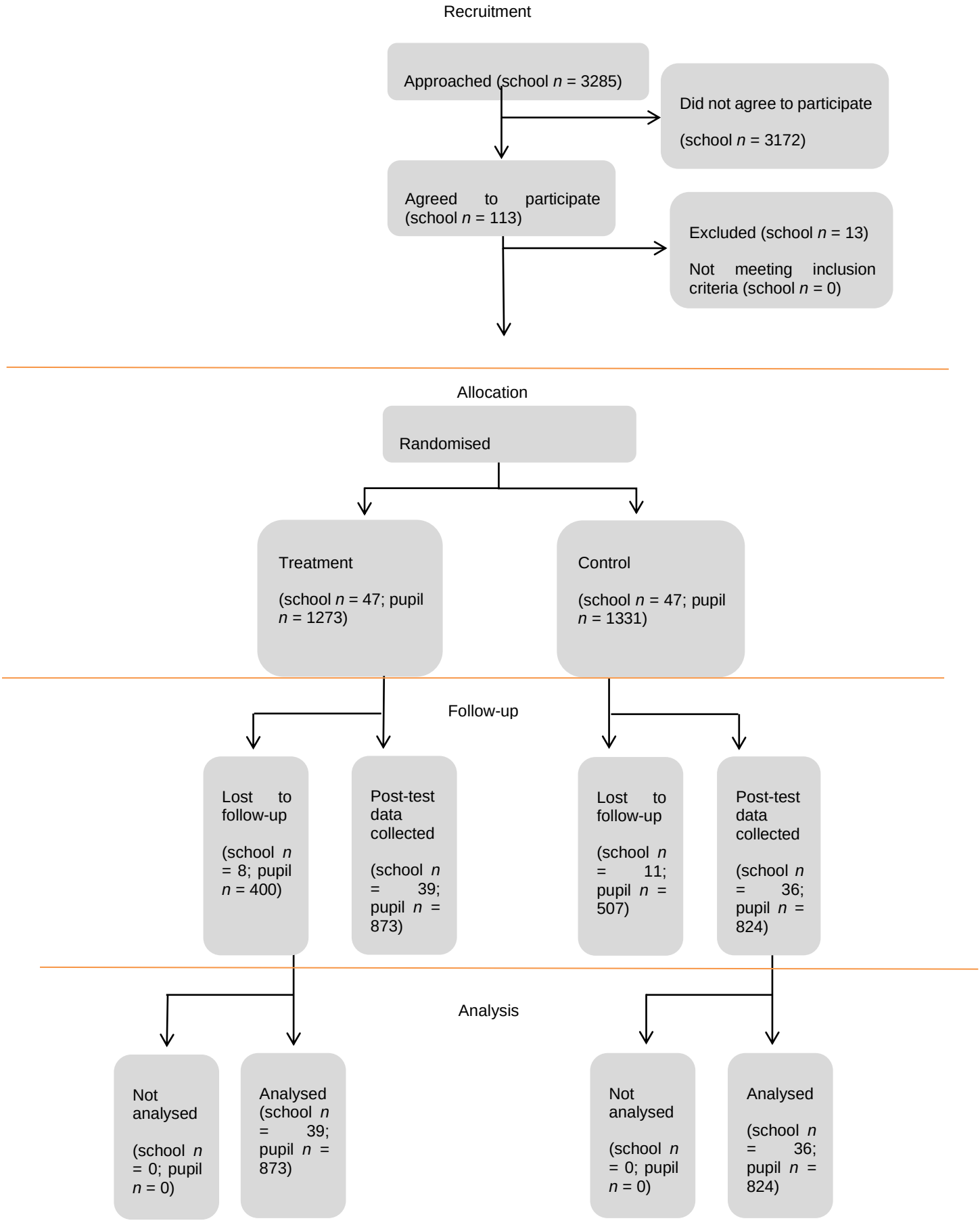


Table 6: Minimum detectable effect size at different stages

		Protocol		Randomisation		Analysis	
		Overall	FSM	Overall	FSM	Overall	FSM
MDES		0.22	0.33	0.22	0.27	0.34	0.44
Pre-test/post-test correlations	Level 1 (pupil)	0.50	0.50	0.50	0.50	0.14	0.10
	Level 2 (class)	0.50	0.50	0.50	0.50	0.51	0.35
	Level 3 (school)	N/A	N/A	N/A	N/A	N/A	N/A
Intra-cluster correlations (ICCs)	Level 2 (class)	0.15	0.15	0.15	0.15	0.33	0.30
	Level 3 (school)	N/A	N/A	N/A	N/A	N/A	N/A
Alpha		0.05	0.05	0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8	0.8	0.8
One-sided or two-sided?		two-sided	two-sided	two-sided	two-sided	two-sided	two-sided
Average cluster size		23	3	23	6	21	4
Number of schools	Intervention	48	48	47	47	39	39
	Control	48	48	47	47	36	36
	Total	96	96	94	94	75	75
Number of pupils	Intervention	1200	144	1222	329	873	288
	Control	1200	144	1222	329	824	207
	Total:	2400	288	2444	658	1697	495

Notes. As no pre-test data was ultimately available, the pre-test/post-test correlations as analysed report the multiple correlation between the covariates included in the analysis model and the primary outcome measure. Average cluster sizes are harmonic means of cluster sizes, which is more conservative than using the arithmetic mean in the presence of unequal cluster sizes (Bulus et al., 2019).

Attrition

Eight treatment and 11 control schools declined to participate in testing at the end of the intervention, and a further 166 pupils in treatment schools and 203 pupils in control schools were absent at the point of testing (despite return 'mop-up' testing in schools where three or more pupils were not available on the day of main testing). Reasons for schools not participating in testing included withdrawal from testing for various reasons (e.g., staff capacity and turnover), non-response to requests to schedule a testing date, scheduling challenges and RA illness (and school was unable to reschedule before end of year). Ultimately, 873 pupils in 39 schools allocated to treatment and 824 pupils in 36 schools allocated to control were analysed. This led to an attrition rate of 28.6 percent for the intervention group and 32.6 percent for the control group, which amounted to 30.6 percent of the total randomised sample (see Table 7).

Table 7: Pupil level attrition from the trial (primary outcome)

		Intervention	Control	Total
Number of pupils	Randomised	1273	1331	2604
	Randomised (SAP)	1222	1222	2444
	Analysed	873	824	1697
Pupil attrition (from randomisation to analysis)	Number	400	507	907
	Percentage[1]	31.4	38.1	34.8
	Percentage[2]	28.6	32.6	30.6

Note. The 'Randomised' numbers include the data on all pupils submitted by participating schools prior to randomisation. However, this was not the final class list in all cases as documented in the *Randomisation* section. As such, while 59 percent of schools provided class lists prior to randomisation, the rest provided whole year group lists. As a result, the 'Randomised' numbers are higher than those reported in the 'Randomised (SAP)', which were adjusted based on the reality that any school that had provided an entire year group would only have one class participate (an assumption on class size was made using the average of 26 pupils per class reported by schools that supplied class lists). The 'Percentage[2]' attrition numbers are calculated using the 'Randomisation (SAP)' numbers, since it was never possible for all pupils randomised to participate in the evaluation.

Pupil and school characteristics

Table 8 presents the baseline characteristics of treatment and control schools and pupils as randomised. In general, it shows that treatment and control schools are similar to each other and similar to the national average for a range of characteristics. Both intervention and control schools were more likely to be located in urban settings than the national average (100 percent of control schools and 94 percent of intervention schools as compared to 87 percent of all schools nationally). Both groups were also more likely to be rated 'Outstanding' by Ofsted, and somewhat less likely to be rated as 'Good' than national schools on average.

In terms of pupil characteristics, both treatment and control schools had very similar KS1 performance (15.45 for treatment and control), which was also very similar to the national average (15.9). Intervention schools had a slightly higher proportion of EVERFSM pupils (34 percent) vs. control schools (29 percent), which was also slightly higher than the national average (31 percent). The standardised difference between treatment and control groups is –0.109 (Imbens & Rubin, 2015).

The treatment and control pupils were much more likely to have EAL (25 percent of intervention pupils and 20 percent of control pupils) as compared to only 15 percent at the national average. Here, however, the absolute standardised difference between the intervention and control groups is small (–0.121). Table D1 in Appendix D presents the analogous balance characteristics for the groups as analysed. The descriptive statistics in this table show somewhat larger

imbalances between the treatment and the control groups. As a result of these differences, we control for FSM and EAL status in all of our analysis.

Table 8: Baseline characteristics of groups as randomised (n = pupil, N = school)

School-level (categorical)	National-level percentage	Intervention group		Control group		
		N (missing)	%	*N* (missing)	%	
Setting: Urban	87.3	44 (0)	94.5	47 (0)	100.0	
Setting: Rural	12.7	3 (0)	5.5	0 (0)	0.0	
Ofsted: Outstanding	17.1	12 (0)	25.5	12 (4)	29.1	
Ofsted: Good	69.4	30 (0)	62.3	27 (4)	61.7	
Ofsted: Requires improvement / Inadequate	13.4	5 (0)	12.2	4 (4)	9.2	
School type: Academy	23.6	14 (0)	27.8	11 (0)	27.3	
School type: Community	41.2	21 (0)	48.7	23 (0)	48.2	
School type: Other	35.2	12 (0)	23.5	13 (0)	24.5	
School-level (continuous)	National-level mean	*N* (missing)	Mean (SD)	*N* (missing)	Mean (SD)	Standardised difference
KS1 average performance	15.9	42 (5)	15.45 (1.13)	44 (3)	15.45 (1.16)	0.000
Pupil-level (categorical)	National-level percentage	*n* (missing)	%	*n* (missing)	%	
FSM	30.9	437 (0)	34.3	380 (0)	28.5	
Non-FSM	69.1	836 (0)	65.7	951 (0)	71.5	
EAL	15.3	321 (0)	25.2	262 (0)	19.7	

Non-EAL	84.7	952 (0)	74.8	1069 (0)	80.3	

Note. School-level imbalance is calculated applying weights for the size of the school, so that schools that are relatively more important in the pupil-level impact estimation are afforded the same importance in understanding imbalance.

Outcomes and analysis

Primary analysis

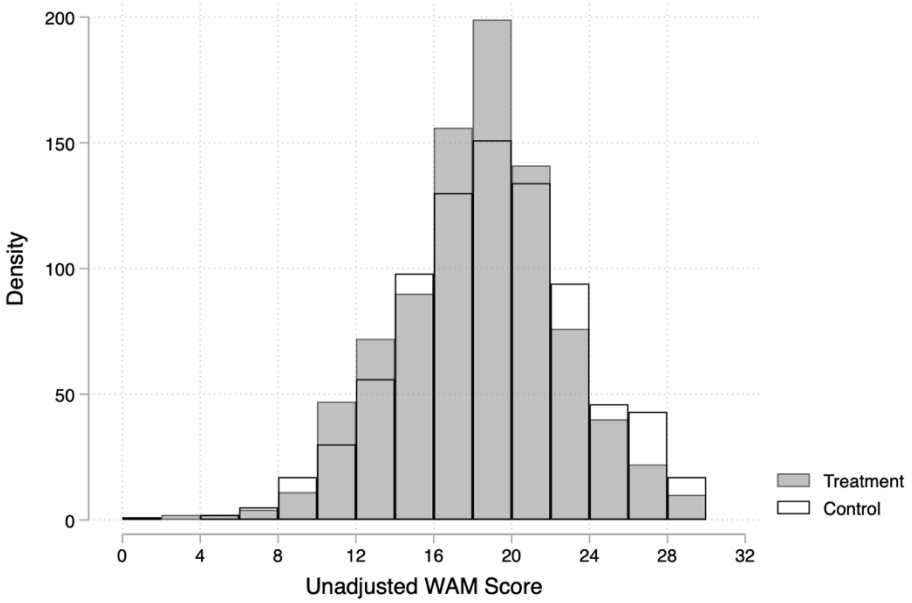
Table 9 presents the results of the analysis for the primary outcome measure. It shows the unadjusted mean for the CoW intervention group (17.80) and the unadjusted mean for the control group (18.39). After adjusting for covariates in the analysis model, we find an adjusted mean difference of 0.15. Based on this, we calculate a Hedges' *g* effect size of –0.03, which is not a statistically significant difference between the two groups ($p = 0.68$).

The same unadjusted WAM scores for the intervention and the control group may be seen in Figure 4. The overall mean for the WAM across both treatment arms is 18.09 and the median is 18, which is reflected in the normal distribution of this outcome measure.

Table 9: Impact estimates

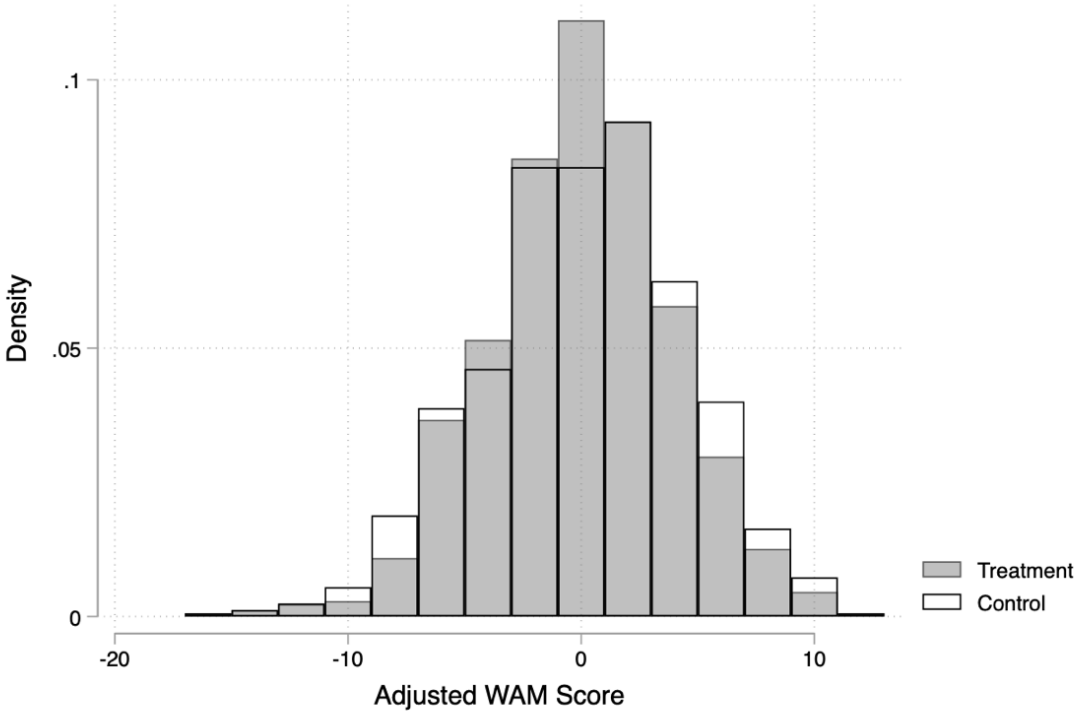
	Unadjusted means				Effect size		
	Intervention group		Control group				
Outcome	*n* (missing)	Mean (95% CI)	*n* (missing)	Mean (95% CI)	Total *n* (intervention; control)	Hedges' *g* (95% CI)	*p*-value
Primary outcome							
WAM score (ideas scale double weighted)	873 (400)	17.80 (17.21, 18.40)	824 (507)	18.39 (17.62, 19.17)	1697 (873; 824)	–0.03 (–0.19, 0.12)	0.68
Secondary outcomes							
WSEM score	861 (412)	64.74 (63.92, 65.56)	813 (518)	64.40 (63.46, 65.33)	1674 (861; 813)	0.04 (–0.08, 0.15)	0.55
Ideation score	861 (412)	20.09 (19.79, 20.39)	813 (518)	20.00 (19.71, 20.28)	1674 (861; 813)	0.02 (–0.09, 0.14)	0.69

Figure 4: Histogram of unadjusted writing assessment measure (WAM) scores by treatment arm



We also present the impact analysis results for the primary outcome measure graphically through a second kernel density plot by treatment arm displayed in Figure 5. The adjusted WAM scores for this plot are obtained from regressing the WAM writing score on pre-test measures (planned to be phonics score but, due to data access issues, instead FSM and EAL status plus class composition of these) and the variables used for stratification (i.e., the analysis model other than the treatment indicator).

Figure 5: Histogram of adjusted writing assessment measure (WAM) scores by treatment arm



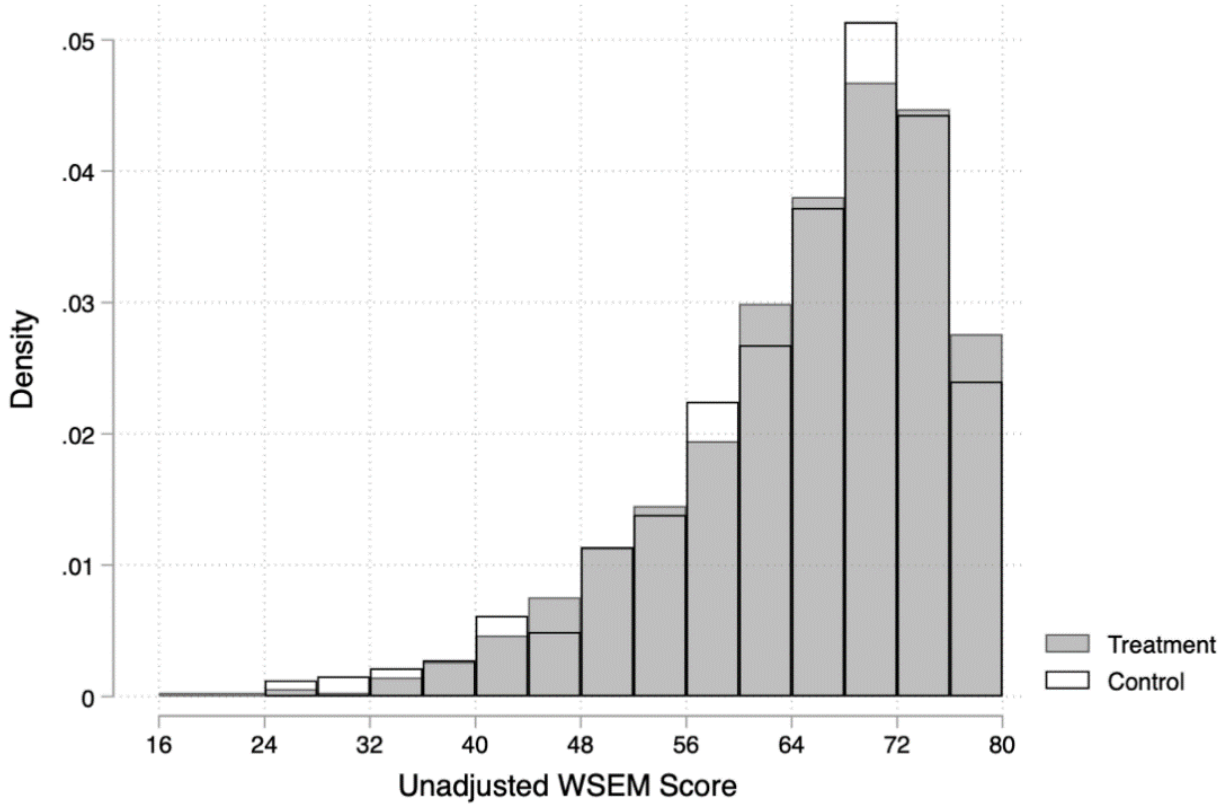
Unsurprisingly, conditioning the WAM scores on the variables used for stratification does not significantly alter the distribution for the treatment and control arms of this trial. The small mean difference between the two arms in our main analysis is evident in the plot.

Secondary analysis

The results of the secondary analysis are also presented in Table 9 above. The secondary outcome measures for this trial are writing self-efficacy (WSEM) and ideation. These outcomes are more closely aligned with the content and aims of the CoW intervention, and so one might expect them to be more likely to register a larger effect.

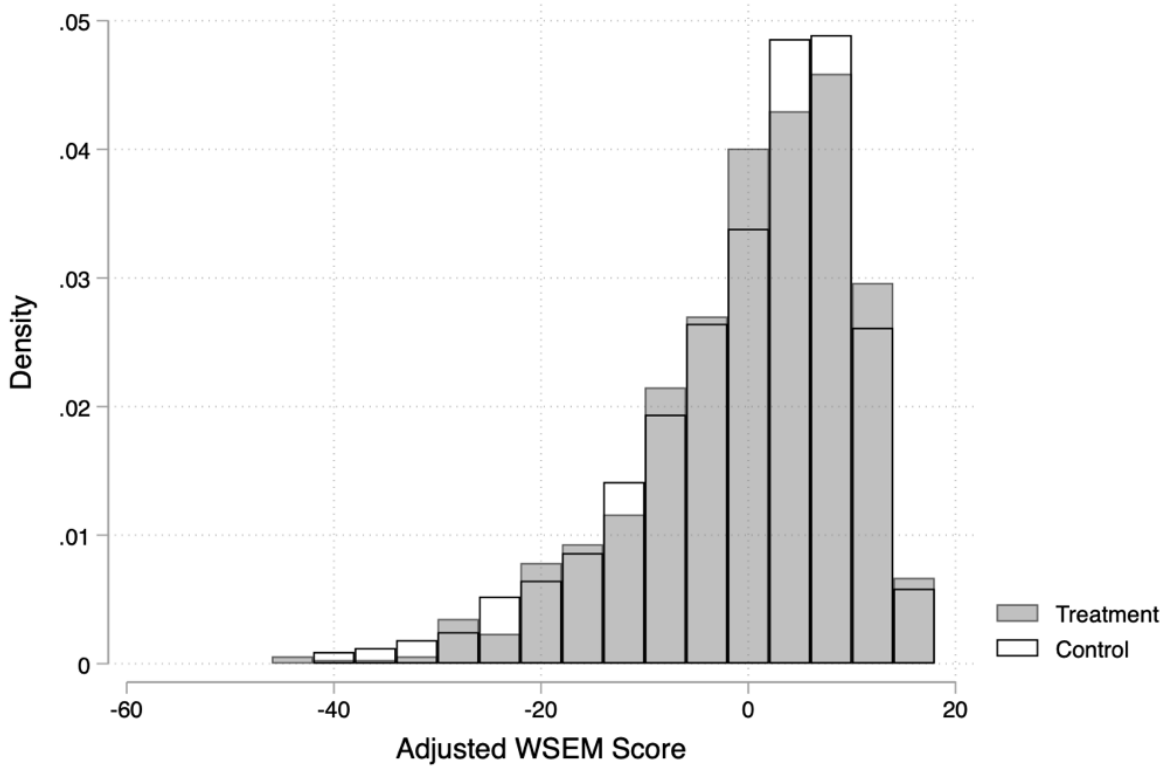
For the WSEM, there is only a very small difference in the unadjusted means between the intervention group (64.74) and the control group (64.40). Figure 6 shows the histogram of the unadjusted WSEM scores for the intervention and control groups. It shows a right skewed distribution, indicating a high proportion of pupils giving positive responses on how they view themselves as writers, introducing some risk of ceiling effects in the analysis of this outcome measure which could attenuate the impact estimate.

Figure 6: Histogram of unadjusted writing self-efficacy measure (WSEM) scores by treatment arm



After adjusting for pre-intervention covariates, the adjusted difference in means is *b* = 0.37, which translates into a Hedges' *g* effect size of 0.04. This effect size is equivalent to zero months of progress and again, there is no evidence that this is a statistically significant difference (*p*-value is 0.55). This may be seen in Figure 7, which shows the adjusted kernel density plots of the WSEM scores by treatment arm.

Figure 7: Histogram of adjusted writing self-efficacy measure (WSEM) scores by treatment arm



A similar result emerges for the ideation score. Both intervention and control groups have a nearly identical unconditional mean of the ideation score (20.09 for the intervention group and 20.00 for the control group; see Table 9 for further details and Figures E4 and E5 in Appendix E), which yields an effect size of 0.02 and is not statistically significant (*p*-value is 0.69). Again, this shows that the CoW intervention did not have a substantive impact on the secondary outcome measures in this trial.

Analysis in the presence of non-compliance

In order to examine the issue of non-compliance, we estimate the CACE. There are 21 intervention schools (45%) that did not comply based on the compliance criteria outlined in this report. Given this proportion of schools deemed to be non-compliant with the intervention and that non-compliant implementation may be diluting an underlying treatment effect in compliant schools, we want to account for this in calculating the effect size for the primary outcome measure.

To calculate the CACE, we use the 'ivregress' functionality of Stata to make necessary adjustments to standard errors (which are also clustered at school level) due to the instrumental variables approach. (Further details of this approach are available in the syntax file in Appendix F.) As compliance data are available for the full primary analysis sample, this analysis is also carried out for this full sample (n = 1697).

The *p*-value of the exogeneity test (0.35) suggests that there will be little difference in findings between the ITT and CACE analyses. Indeed, the results of the complier analysis produces an effect size of –0.05 (95% CI: –0.30, 0.20), which is qualitatively similar to that of the ITT analysis; furthermore, the *p*-value of the CACE treatment estimate (0.67) is not statistically significant. As such, we interpret these findings as not providing evidence of differential treatment effects among schools with higher levels of compliance for the primary outcome measure.

Missing data analysis

In the SAP for this trial, we outlined a missing data strategy. We noted that this strategy would be implemented if more than 5% of data in the model is missing or if more than 10% of data for a single school is missing. Due to issues with

testing, a substantial number of pupils in treatment and control schools do not have outcome data. In order to test whether or not this data is MAR, we ran a logistic regression model to predict missingness in outcome data using all variables in the analysis model plus eligibility for FSM (and the proportion eligible for FSM in the school), and EAL status (and the proportion with EAL in the school).

The results of this analysis show that the treatment indicator, FSM eligibility, and one of the randomisation stratum indicators are statistically significant predictors of missing outcome data (FSM eligibility is only statistically significant at the 10 percent significance level). This implies the potential for bias to complete case analysis ignoring these factors since, for example, FSM pupils might have worse outcome data, which we would be more likely not to observe due to missingness. However, since all of these variables are already included in the (revised) primary analysis model, there are no further actions possible in relation to systematic predictors of missing outcome data, and our existing model should address concerns about systematic missing outcome data associated with these factors (i.e., we believe, based on our analysis, that conditional on these variables, outcome data are MAR). If any additional predictors of missingness had been discovered in this analysis, we would have run an additional ITT model controlling for them as well.

In terms of missing predictor data, none of our analysis sample is missing these data (0.0% for EAL and 0.0% for FSM). Since no one in our sample is missing this information, there is no reason to undertake MI.

Sub-group analyses

As is standard in all EEF-funded evaluations, we considered whether there is evidence of differential effects among pupils eligible for FSM as a separate subgroup (n = 495). We started by considering an augmented version of the primary analysis model, including an interaction term between the treatment variable and membership of the FSM sub-group. The estimate on this interaction is small and not statistically significant (p = 0.66), providing little evidence of a differential effect among the FSM sub-group on the primary outcome. The results for the WSEM for the FSM sub-group are similar, with the interaction term not providing evidence of a differential effect of the intervention on the writing self-efficacy (p = 0.07) of FSM-eligible pupils. There is, however, evidence that the CoW intervention had a positive and statistically significant impact on the FSM sub-group's ideation. Here, the interaction term is positive, statistically significant (p = 0.02) and equivalent to three months' progress. This provides some indication that the intervention may be beneficial for increasing the creativity of children from disadvantaged backgrounds. Given the underpowered nature of the trial and the fact that this is not the primary outcome, but rather one of several analyses, there is the chance that this is a false positive. We are therefore cautious in our interpretation of this result. See Table D2 (Appendix D) for full details.

Additional analyses and robustness checks

As outlined in the SAP for this trial, no additional analyses were planned for this trial. However, in light of ongoing discussions about appropriate interpretation of classical statistical inference, we carried out alternative statistical inference using randomisation inference to provide useful information on the extent to which there is variation between these different approaches to statistical inference. Randomisation inference is a method of conducting statistical inference using the uncertainty inherent in the randomisation process regarding the assignment of units in the trial to the treatment arms, rather than any appeal to an external sample and sampling variation (i.e., focusing on internal rather than external validity; see Cunningham, 2021). These comparative findings are reported in Table D4 in Appendix D. They result in very similar p-values, despite their differing conceptual underpinnings.

Furthermore, as outlined in the *Methods* section, we decided to run three exploratory robustness checks to probe the main findings. The results of these analyses are reported in Table D6 in Appendix D.

Due to delays in testing some schools, we ran an additional robustness check controlling for the date of testing, since this could have affected the dosage of the intervention. The results obtained from this analysis were not substantially different from the overall impact evaluation results.

We also ran a model accounting for differences in the markers who marked the WAM (the primary outcome measure), since it is possible that they introduced a new source of bias due to different degrees of leniency in marking. The inclusion of these marker fixed effects slightly attenuated the effect size towards zero, but also did not change the overall results from the impact evaluation.

Finally, due to the imbalance in school level KS1 scores of the sample as analysed (see Table D1), we also ran a model which included average school level KS1 scores. Again, this analysis did not affect the effect size calculated in a way that alters our substantive conclusions.

Estimation of effect sizes

As previously outlined, effect sizes are calculated using Hedges' *g*. These have been presented in Table 8 for the primary and secondary outcomes. Table C1 in Appendix C contains additional information used in the estimation of effect sizes (e.g., the SD). None of the effect sizes estimated for this trial are of a substantive magnitude, nor are they statistically significant. This is true for both the primary and secondary outcomes.

Estimation of intra-cluster correlation (ICC)

As this is a clustered trial, we estimated the ICC at the class level. At the time of randomisation, the ICC was estimated to be 0.15, which we based on EEF guidance drawing on analysis of NPD data from across England. However, in the event, the ICC of the primary outcome measure (the WAM) has turned out to be considerably higher, at 0.33. There are aspects of the sample which may explain some or all of this difference, including that the schools participating in the trial are disproportionately drawn from urban areas (although this was also the case for the estimate using EEF guidance, since we used the highest regional ICC, which happened to be for London), and that our missing data analysis provides weak evidence of the potential for our analysis sample to differ in terms of FSM composition. That said, we doubt these explain the extent of the difference, which is likely to be attributable to the outcome and/or outcome measure itself; for example, reflecting within-class shared writing practices that are consistently captured by the measure. This should be borne in mind in the design of future trials making use of measures of this type.

Implementation and process evaluation

Summary

- Administrative data showed that teacher compliance with training was fairly low, with only 55% of teachers meeting or exceeding the compliance threshold. However, despite relatively low compliance rates, the teachers in case study schools who attended felt that the training sessions were worthwhile.
- Some teachers also highlighted a perceived lack of buy-in from SLTs, which made it more challenging for teachers to attend training and implement the CoW intervention.
- Some teachers in case study schools found it challenging to implement the CoW intervention due to time constraints and pressure to teach the national curriculum, in light of SATs.
- Teachers in case study schools described the CoW as a less structured approach to writing, highlighting a potential adjustment period required for pupils and teachers to get used to this new approach.
- The survey data showed that teachers in both intervention and control schools reported similarly high rates of teaching about some elements important to the CoW intervention in the last academic year.
- Teachers were very positive about the intervention, specifically the way they perceived it impacted children's enthusiasm for writing. This was added to the updated logic model (see Figure 8).

This section contains the findings of the IPE, which are based on data collected from case studies and the supplementary post-intervention survey (see *Methods* for more details). The subsections are structured by theme and not RQs, in line with what Nowell et al. (2017) propose as best practice qualitative reporting; however, the RQs addressed by each subsection are outlined in Table 10. The first subsection provides background *Context* on the case study schools, as well as information on usual practice in relation to teaching writing, while the next two subsections set out *factors* affecting the *implementation* of CoW and its delivery with fidelity. The final subsection details evidence on the *Mechanisms* underpinning the intervention and its *perceived impact*.

Table 10: Research questions (RQs) addressed

Subsection in *IPE results*	Research questions
Implementation	RQ1**.** In what ways was the programme implemented? What are the barriers and facilitators of delivery? In particular: **a** SLT 'buy-in'; **b** delivery of training and resources: (i) whether it appears to be effective in ensuring that teaching staff understand the aims and main features of the intervention; **c** delivery of the intervention: (i) whether it appears to facilitate children's engagement.
	RQ2. To what extent did the schools engage with the intervention in line with the intervention aims?
	RQ5. How do the strategies and techniques from the CPD and residential training emerge as part of teaching, and in what ways does this lead to improved teaching practice?
Mechanisms and perceived impact	RQ1c. Delivery of the intervention: (ii) whether it appears to be effective in supporting children's attainment; and c) whether it appears to facilitate children's engagement. RQ6. To what extent does the intervention improve teacher confidence as a writer?

	RQ7. How does the intervention benefit the pupils: what are the mechanisms by which the teacher implements the intervention with their pupils?
	RQ8. How does teachers' practice in relation to pupils revising their writing change?

Context

This section discusses three contextual factors in case study schools that influenced the implementation of CoW: teacher attendance at training; the school's rate of FSM; and the teacher's experience of writing. Each is discussed in turn, followed by a description of how CoW compares to usual practice.

Compliance: Teacher attendance at training was relatively low

Overall, 26 (55%) intervention schools met the minimum trial compliance criteria of attending both weekend residentials and at least two CPD sessions, 21 (45%) did not (further detail on barriers to attendance and training experience are provided in the next section (*Implementation*). Within the six case study schools, attendance at training was higher. Of the six case study schools, teachers from four schools attended all training sessions. In the two other case study schools, only one CPD session was missed. Training attendance is not a perfect proxy for engagement, as training was missed for a variety of reasons. However, it is possible that the case study schools may represent some of the more engaged schools. Case study data is supplemented by data from the survey, which all schools were asked to complete, to provide a broader picture of experiences with implementing the CoW intervention.

The school's rate of free school meals (FSM)

As well as the characteristics presented in the *Methods* section, variation in case study schools was also captured qualitatively. Teachers at the three schools with high (above 20%) FSM rates described factors that made it more difficult to teach their pupils writing. They described having classes with high numbers of children with special educational needs (SEN), as well as more generally, children reading less at home, and having more limited vocabularies and life experiences upon which to base their writing.

> *'I think our children's experiences – life experiences – are not as rounded as a lot of other children. So they find it very difficult to write about things that they have not experienced that you would take for granted. Vocabulary, our children on entry, vocabulary is extremely low and that then is a pattern through school.' (SLT member 01)*

At the school with a fairly low FSM rate (between 6 and 10%), issues with writing attainment were also reported, particularly among boys, who were seen to write the minimum and less creatively. In contrast, members of staff from the two schools with very low rates of FSM (under 5%) did not describe any specific challenges with teaching writing; indeed, one described their planning of writing lessons as one of the school's key strengths.

Teacher experience

The teachers involved in implementing the CoW programme at case study schools had a range of teaching experience, including those who were recently qualified with less than four years of teaching experience, a teacher who had taught for nine years, and a teacher who was a member of the school's SLT. Among these teachers, there were those who had a passion for writing, including a teacher who had specialised in English at university. The teachers with a passion for writing really welcomed the opportunity to take part in the programme, seeing it as an opportunity to improve their own writing and their teaching of writing. There were also teachers who did not have a particular inclination for writing themselves or did not see teaching English as one of their strengths. These teachers saw CoW as an opportunity to improve their teaching, for example:

> *'When I first saw Craft of Writing I thought this is perfect for me. I love the idea of going away and actually finding some time to be a writer myself.' (Teacher 04)*

One of these teachers was concerned that CoW was something additional they would have to get used to, among the other demands on them as a relatively inexperienced teacher.

> *'I was an NQT, I was focusing on this English, planning writing and getting my head around it so that I know what I'm doing. Then to sort of say, right we are going to start playing around with this idea, at first I was a little bit that I'm just getting my head around it.' (Teacher 02)*

Usual practice: Some Craft of Writing (CoW) techniques were already in use

In terms of how English, particularly writing skills, are typically taught in Year 5 classes, some interviewees, particularly SLT members, commented that their school already used CoW techniques, including free writing (i.e., periods where children are encouraged to write focusing on *composition*, rather than elements of writing such as spelling, grammar and handwriting), and phases of drafting and editing work. The control school survey also found that similar proportions of intervention and control school teachers reported that in the last academic year they had taught about a range of elements important to CoW including: writing for different audiences (89% intervention to 97% control); the effects on readers of choosing specific words and phrases (97% intervention to 92% control); and had provided opportunities for children to write in a way that involved multiple phases of drafting (86% intervention to 89% control) (Full survey responses are available in Appendices J and K.)

Some of the case study teachers also described key differences between usual classroom practice and CoW-style teaching. Case study teachers said that the way they typically taught writing was more structured and directive. For instance, they tended to provide children with an example of the type of writing they should produce, including instructions on the features to include in it, particularly grammatical structures. It was also commented that typical writing lessons often concentrated on story writing and presenting events using a 'beginning, middle and end' structure, rather than on providing rich descriptive detail as in CoW lessons.

> *'...it's just very, very different to what they have done before... Because before, a lot of the times teaching is you read a bit of a book and you discuss the book, you rewrite the story and then you move on to your next unit.' (Teacher 01)*

> *'... they [pupils] are trying to use the same grammar functions over and over again and they would have seen examples, and either replicating the same examples or they'll be coming up with something that's very near to those things.' (Teacher 06)*

These differences were picked up by the survey, which found that a much higher proportion of intervention teachers said that they had allowed children to choose their own topic to write about (75% intervention to 35% control) and provided opportunities for free writing (94% intervention to 41% control). In addition, while 68% of control school teachers said that they had received training on teaching writing skills in the last academic year, only 11% of control school teachers said they had received training on developing their own writing skills (in other words, going beyond training on how to teach writing to others), which is a core element of the CoW programme.

Implementation

Delivering the intervention with fidelity

To deliver the CoW intervention with fidelity, in other words, to deliver it as intended, teachers needed to apply the CoW framework, introduced to them through the residentials and CPD sessions, in the classroom. The CoW framework is made up of five overarching elements: language choices; text-level choices; the reader–writer relationship; being an author; and the writing process. From interviews and notes taken during lesson observations, there was evidence of all five elements being applied in multiple schools, with use of techniques relating to the writing process being particularly evident. Examples of the ways in which the framework was observed as translated into the classroom by teachers are provided in Table 11.

Table 11: Translation of Craft of Writing (CoW) framework to the classroom

Element from Craft of Writing (CoW) framework	Examples of translation to classroom
Language choices	Emphasis on language choices (e.g., using 'powerful' words and phrases) (Interview numbers: 03, 04; Observation numbers: 01, 03, 04, 05, 06)
	Discussion of why authors use particular language (Observation: 02)
Text-level choices	Impact of shorter and longer sentence lengths (Observations: 01, 05)
	Use of dialogue to enhance mood (Observation: 01)
	Use of pace to enhance mood (Observation: 01)
	Use of tense (Interview: 03; Observation: 02)
Reader–writer relationship	Helping the reader to feel specific emotions (Interview: 01, 02, 03; Observations: 03, 04)
	Use of 'show, not tell' (Interview: 02, 04; Observations: 01, 02, 03, 04, 05)
	Importance of writing to engage the reader (Interview: 01; Observations: 03, 05)
Being an author	Emphasising that children are authors and have ownership over their writing (Interview: 03; Observation: 03)
	Emphasising children having choice over their writing (Interview: 03; Observations: 03, 04, 06)
Writing process	Free writing Time for free writing (Interviews: 01, 02, 03, 04, 05, 06; Observations: 01, 03, 04, 05, 06) Separate exercise books for messier and neater writing (Interview: 06; Observations: 01, 03)
	Peer feedback on writing (Interviews: 02, 03, 04; Observations: 01, 03, 05)
	Revising and editing Discussion of writing as an iterative process (Interviews: 02, 05; Observations: 03, 05) Encouragement to write things down before going back to edit (Interviews: 03, 04, 06; Observation: 06)

	Writing continued over more than one lesson (Interviews: 03, 04, 05, 06; Observations: 01, 04)

There were two main ways identified that the programme model was modified (see *Intervention* section for a detailed description of the programme). In some case study schools, an emphasis was retained on including specific aspects of spelling and grammar in pupils' writing, while the CoW approach encourages writing initially without a focus on spelling and grammar, and then going back to edit and refine. Examples of this were found in observations, where teachers wrote vocabulary on the board that they expected to be included in their class's writing, or must not be misspelt. It was also observed that a teaching assistant, who would not have attended CoW training themselves, reminded children to check their spelling and grammar several times, and corrected their work during the time allocated for open-ended writing.

In addition, a key part of the approach is giving children opportunities to write more freely, with greater choice over topic and how they spend their writing time. It was noted in observations, however, that lessons tended to be broken into a number of short sections of five to twenty minutes in length, with children expected to complete a specific writing task in this time; for instance, continuing a story having been given a sentence starter. Interviewed teachers also explained why they tended to give more direction and had given their class fewer opportunities for more extensive free writing. This included giving children more specific parameters in which to write, which, it was suggested, gave children the confidence to get started with their writing, as well making it easier for the teacher to support the class and provide feedback on their work. One case study teacher described how some pupils needed additional support to start them off, as some pupils hadn't always had the experiences relevant to the activity:

> *'So the free write… my children always need a little bit of something, like the story with the description setting or something, because they don't always have that life experience. Some of them do, but like, some of them haven't even been in a forest before.' (Teacher 03)*

In one case study school, the way that CoW was delivered seemed in particular contrast to the approach laid out in the intervention model. At this school, the framework was primarily used as a tool for analysing texts to determine why authors had used certain techniques in their writing, with the class spending minimal time on the actual writing process '...*maybe 10 minutes at the end they are going to get something down in their books'* (Teacher 02). This was also found to be the case during the observation, where it was noted that, unlike in any other case study school, the lesson involved analysing a fictional text with no creative writing component. It was at this school that the class teacher described having initial concerns about taking part in the intervention, as they were recently qualified and just getting used to other aspects of English planning. Compared to other case study teachers implementing the CoW intervention, this teacher described particular difficulties with understanding how to use the Framework in practice. For instance, they described finding free writing unmanageable, because the children all wrote different pieces and the teacher felt that they had to provide tailored feedback on the children's work, which related to the different sub-components of the CoW framework.

> *'I think they found that quite difficult and I found it difficult. Every sentence in every piece, why have you done this, why have you done that and related to the Framework, and because it was so new there was a lot of building blocks that we needed to put in.' (Teacher 02)*

The teacher also felt under considerable pressure to dedicate time to CoW, while also ensuring that other curriculum requirements, particularly grammar teaching, were met. Over time, and with support from a member of the school's SLT, the teacher moved towards using the Framework primarily as a tool for examining how authors write, with the idea that pupils would then begin to apply similar techniques in their own writing.

Factors affecting implementation

Building on some of the reasons that the intervention was not always delivered with fidelity described above, this section discusses four overarching factors that were identified as affecting implementation: attendance at training; SLT 'buy-in'; the national curriculum; and pupils' response to the intervention.

Attendance at training

As there is no in-school support associated with CoW, attendance at training sessions was essential in order for teachers to be able to implement the intervention in class. Attending all of the residential and CPD training sessions was, however, a substantial challenge for schools and, as noted above, compliance with training attendance was low. The administrative data showed that a third (34%) of teachers at intervention schools were not able to attend at least one of the residentials, and over half (57%) missed at least one CPD session. According to the survey, reasons for not attending residentials included: personal commitments (including childcare); lack of school cover; the involved teacher leaving or changing during the school year; and not being aware of the training. The reasons given for not attending CPD sessions were the same, but with the following additional reasons: school commitments (including national curriculum assessments); sickness or hospital appointments; and difficulties with travelling to the CPD session.

Teachers at case study schools also described how attending the training sessions had been quite a commitment: attending weekend residentials meant giving up time on the weekend, while CPD sessions required teachers to leave their class with cover. It was clear, however, that teachers felt it was worth attending the training sessions because of what they got out of them. Teachers valued the fact that CoW gave them the opportunity to focus on improving their own writing skills, in contrast to other training courses which focus purely on teaching skills. Teachers explained that by having the space to write, try out creative writing exercises and receive feedback from tutors on residentials, they understood more about the process of writing and it helped them to understand how it feels as a pupil when you are asked to begin writing. Through the residentials, teachers also picked up teaching techniques to take back to the classroom, which were further reinforced by the CPD sessions that focused more directly on how to put the CoW intervention into practice in the classroom.

> *'...your writing improved in just sitting with the tutors and having chats with them. You knew how to improve yourself and even then delivering as a teacher by just watching the way that they delivered it. It was inspiring. It was really good.' (Teacher 03)*

SLT 'buy-in' was limited

SLT support was crucial in enabling teachers to attend residential and CPD training sessions. Within the case study schools, teachers reported that their SLT had been supportive of them attending, arranging cover and, in some cases, giving them time back after residentials to make up for the fact they had worked over the weekend. Teachers did report, however, that at other schools, teachers were not always released to attend training by their SLT, particularly because of upcoming Ofsted inspections, and because the timing of one of the CPD sessions clashed with national assessments (commonly known as SATs).

> *'I know how schools work and they will turn around and say, "what are you actually doing that day, do you need to be there?"… So, I think it's just about SLT making sure that they make time for it here and I've had no problem with that…. But that's probably where the problems are, I think where people haven't been there. It will be all SATs.' (Teacher 06)*

Another way that a school's SLT influenced implementation was through the level of freedom they gave teachers to implement the CoW intervention in the classroom. Members of the SLT could facilitate implementation through providing teachers with the time to plan CoW lessons and to deliver them within busy timetables, by embracing the teachers' use of new techniques and being responsive to teacher's requests for advice or guidance. It was, however, clear from both teacher and SLT interviews that SLT members had not provided substantial input and it was instead key that the teacher was given authority to use the new approach. Indeed, as discussed in the *Fidelity* section, in the school where the SLT member had been more closely involved in developing lesson plans, the intervention was eventually delivered less in keeping with the CoW model, and more in line with the approaches used in the rest of the school.

> *'I've not had to be [that involved]. [The teacher] takes things and runs with it...Obviously, he's talked to me about what he's doing and some of the things that he's had to do in his class and had to change from what had been planned and the expectations' (SLT member 01)*

> *'I had to do a CPD form when I got back anyway, to give feedback and on what the next steps are and what I need for the school in order to complete my task. Then with [the SLT member], I would go and have a chat with her about it and what I needed from her.' (Teacher 03)*

Finally, as indicated by case study and administrative data, in some schools CoW was not implemented, despite teachers attending training sessions, because their school had begun implementing an alternative writing scheme such as *Introduction, Point, Explain, Ending, Links and Language (IPEELL)* or *Talk for Writing*, which were seen to have elements that did not align with CoW. This highlights the importance of the school's SLT 'buying-in' to using the CoW approach, if it is to be successfully implemented.

Adherence to national curriculum limited available time for the intervention

The national curriculum and the requirements of Ofsted were raised as major challenges to implementing CoW in the classroom. Teachers described how there was little time for CoW lessons when other aspects of the English curriculum needed to be covered, such as spelling, grammar, handwriting and reading. This could be a particular problem when time for teaching English had been reduced in order to balance the time spent on teaching different subjects, as part of Ofsted's push, at the time of data collection, for a broad and balanced curriculum. As a result, some teachers reported that they had not spent as long on CoW activities, particularly techniques such as free writing, as they would have liked.

> *'I know that he's found the free writing, he really likes that side of it, but he's found it really tricky to fit it. He's not done as much of that as I know he would have loved to, due to the time constraints again with the curriculum.' (SLT member 01)*

Part of the difficulty with spending time on CoW teaching was that it was seen by some teachers to require them to deviate substantially from standard teaching. For instance, while the curriculum was seen to promote writing with neat presentation and correct grammar, CoW was seen as promoting 'rough', 'messy' work, with a focus on creativity over grammar. This led to concerns being expressed about how teachers could demonstrate to assessors that they were meeting curriculum requirements, if they dedicated time to CoW.

> *'...there's this juxtaposition where you need to get the children certain things into the writing for our assessment purposes. Then you want them to be more creative and I've felt it's been a bit of a balancing act.'(Teacher 05)*

Concerns around meeting curriculum objectives, and especially around finding time for CoW, were expressed particularly strongly by teachers who had taught CoW separately to other English teaching. Where teachers did not express such concerns as strongly, they had typically either integrated CoW teaching with their usual English teaching, for instance, by focusing writing on the cross-curricular topic, or they felt confident that CoW would enable their classes to demonstrate the skills that Ofsted are looking for. Teachers in the latter group felt that, by teaching through CoW, children produced writing that was more engaging to the reader, had greater flow, and naturally included the grammar structures that need to be covered as part of the national curriculum.

> *'We've still got expectations to meet for the local authority and Ofsted. But there seems to be a great emphasis on, "is it a great piece of writing and are you hooking the reader?" The Craft of Writing lends itself perfectly to that.' (Teacher 04)*

Additionally, within the context of the trial, pressure on teachers could arise because their school wanted consistency in lesson planning across classes, to ensure that pupils doing CoW did not miss out on key curriculum content covered by non-intervention classes. This was described as particularly difficult to navigate by a teacher who had to co-plan with teachers of combined Year 5 and Year 6 classes who did not want to change their teaching approach when some of their pupils had national assessments (SATs) coming up, but was also evident in other schools purely in relation to teaching across Year 5.

> *'...we've got [several] Year 5 classes, so it was making sure that the class that we are doing Craft of Writing was still getting the same curriculum entitlement of the others… because we didn't want either [the Craft of Writing class] to miss out on vital bits of other curriculum, equally, we didn't want the other classes to miss out [on Craft of Writing].' (SLT member 01)*

Pupils' response to the intervention necessitated an adjustment period

The fourth influence on implementation was how pupils responded to their teacher using a different approach. Though there were case study interviewees who felt that the CoW intervention was similar to how they typically taught, it was generally described as being challenging for their class to adjust to a less structured approach to writing. This influenced

implementation because it meant that there had to be a period of adjustment with teachers and pupils getting used to the intervention. For one teacher, it was felt that a slower implementation period would have helped.

> *'...because they have had years of, I've got to get this in, I've got to get this in. Then all of a sudden move from that... it just blows their mind a little bit. Maybe if it was integrated [in the] school slowly, and that was an expectation of their writing then, [then] I think that might work a lot better.' (Teacher 05)*

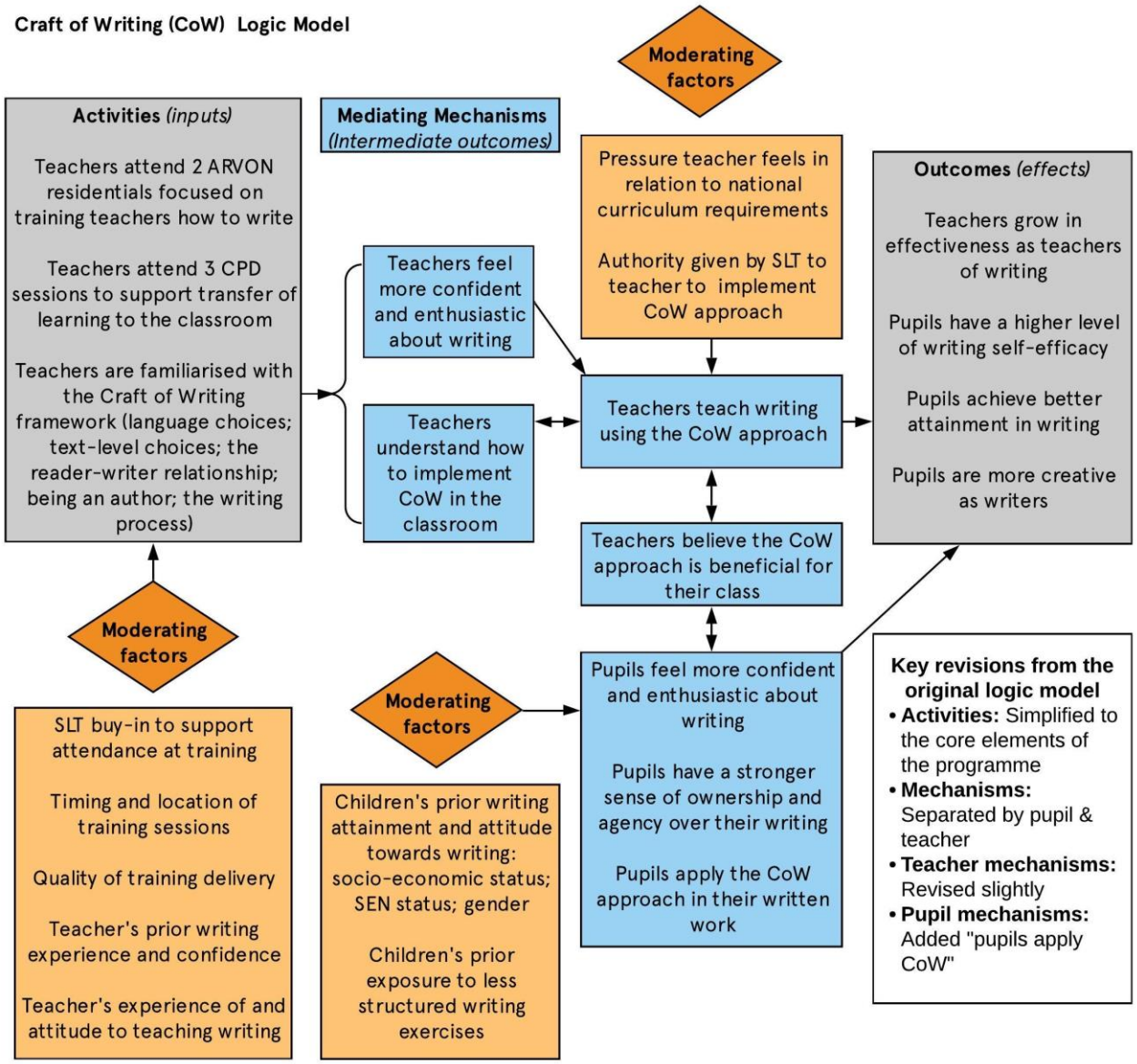
There were three groups identified who might need additional support with CoW. These were pupils with SEN who might find it difficult to adjust to the skills required by CoW, or struggle with the physical aspects or writing; children who prefer structured writing exercises; and pupils from disadvantaged areas who, some teachers felt, had limited life experiences on which to base their writing. Teachers described how support could be put in place to support children who struggle, especially when implementing CoW in areas of higher deprivation, or in classes with high numbers of pupils with SEN. Suggested ways to provide support included breaking down open writing tasks into smaller chunks for process-driven children, allowing children with limited motor skills to use computer equipment or audio recorders for writing exercises, and teachers using their own writing to model new skills to their class. Particularly in more disadvantaged areas, teachers described the importance of providing children with sensory material or new experiences to inspire their writing.

> *'...the trouble they have got with this school particularly is... a huge lack of life experiences... So I've tried to give them experiences like that. We did some writing in the year and took them to a forest and had them lying on the ground and hugging trees just to get the textures.' (Teacher 01)*

Mechanisms and perceived impact

This section considers the evidence in support of the mediating mechanisms and outcomes proposed in the original logic model (Figure 1) and provides information on where these have been revised. The first subsection looks at the mechanisms and outcomes related to teachers, while the second subsection considers these in relation to children. The logic model has been revised to include the updated activities and moderating factors (Figure 8).

Figure 8: Revised logic model for the Craft of Writing (CoW)



Teacher mechanisms and impact

In the original logic model, there were three teacher-focused mechanisms through which the intervention was anticipated to have an effect:

1 teachers growing in confidence as writers;

2 teachers having a stronger subject knowledge of the craft of writing;

3 changing teaching so that it is more focused on using craft knowledge and fostering a community of creative writers.

To reflect evidence from the IPE, these have been revised to:

1 teachers feel more confident and enthusiastic about writing;

2 teachers understand how to implement the approach;

3 teachers believe the approach is beneficial for their class;

4 teachers routinely teach writing using the CoW approach.

The mechanisms chain was envisaged to lead to one teacher-focused outcome: teachers being more effective at teaching writing. This outcome remains unchanged from the original logic model.

In the original logic model (Figure 1), the first mechanism focused on the intervention increasing teachers' writing confidence. This was found to be the case, with nearly all surveyed teachers (94%, 34 of 36 surveyed) saying that the intervention increased their confidence as a writer and two thirds (67%, 24 of 36 surveyed) saying that it had a 'very positive' impact. Increased confidence was also apparent in interviews, with teachers recognising that their writing skills had improved, for instance, because they now used more concrete language or focused on creating impact for the reader through their writing. Teachers also described feeling more confident sharing their writing, including producing written work to share with their class. For some teachers, the mechanism went beyond confidence, however, to encompass broader feelings that teachers had towards writing. Teachers talked about being more enthusiastic about writing following the intervention, with some beginning to write in their spare time. In the most striking example, a teacher who had previously not considered themselves to be a good writer, now wrote regularly and saw being a writer as part of their identity.

> *'[The tutor] said, you have a lovely turn of phrase and I was so chuffed because she is very, very straight about, you know, what is good in your writing and what needs to change. She doesn't pull her punches, but yeah, [they] paid me a great compliment.' (Teacher 04)*

> *'...I didn't think I could write. I didn't think I had it within me to be a writer and I didn't view myself as a writer. My relationship has completely changed with writing as a result of this… [I've started] to see myself as a writer. Actually, some of what I've produced has been quite good, which is weird, and I never thought I would say that.' (Teacher 06)*

Analysis of the case study data showed that the initial confidence and enthusiasm that teachers develop about writing, particularly through the residentials, was the first step towards teachers considering how to change their practice. Following the initial residential weekend, teachers began to try out the techniques and creative writing exercises they had learnt with their class. There were then a number of factors identified from the case study data that appeared to dynamically interact, which influenced how inclined teachers were to continue using and embedding the approach in their classroom. These factors included the teacher's understanding of how to implement the intervention, their impression of the fit between this approach and their usual school practices (i.e., pressure felt in relation to the curriculum and support from the SLT to use this approach), and their beliefs about whether CoW was beneficial for their class. Where teachers became more committed to using the approach, they articulated more strongly in interviews why CoW was beneficial for their pupils (elaborated on in the *Pupil mechanisms and impact* subsection below), compared to standard teaching practice. In particular, such teachers highlighted the importance of giving their class less tightly controlled writing tasks, focusing less on spelling and grammar, and making use of drafting and editing, rather than expecting writing to be right the first time, as key markers of their changing teaching practice. The quotes below illustrate cases where teachers were more and less willing to embed the intervention in their class.

> *'...it's changed the way I think about English. I used to be very much, give them a big tick sheet of national curriculum stuff and say this is what I want to see in your work, and I used to wonder why they didn't use it. I used to read their work looking for things like you know, certain grammatical things. Now I read their work and I'm looking for it to have an impact on me. Then, I look at the grammar as an afterthought – not saying it's not important, but I look at it as an afterthought.' (Teacher 01)*

> *'... I felt the pressure a little bit... I was worried about getting my grammar focuses in. Now, it's do we break off from [Craft of Writing] a little bit or spend a little bit less time on it or keep it. Do we only keep certain aspects of it, or do we plan more of it, plan it for an extra week in our planning to include the Craft of Writing framework? Where do we find that it works most? You know, there is all sorts of discussion that we need to have.' (Teacher 02)*

Pupil mechanisms and impact

The original logic model suggested that there were two pupil-focused mechanisms through which the intervention would have an effect:

1. pupils have a stronger sense of ownership and agency of writing;
2. pupils grow in confidence as writers.

These mechanisms were hypothesised to lead to three outcomes:

1. pupils have a higher level of self-efficacy;
2. pupils achieve better attainment in writing;
3. pupils are more creative as writers.

The IPE found evidence from teacher interviews and survey responses supporting the majority of elements in the hypothesised pathway. The two changes made to the model are including enthusiasm about writing as part of the second mechanism, and adding an additional mechanism: pupils apply elements of the CoW framework in their written work and writing processes.

In terms of the first two mechanisms from the original logic model, survey evidence showed that 35 of 36 teacher respondents felt that the intervention had a positive impact on pupils' writing confidence. Ninety-two percent (33 of 36 surveyed) felt it increased ownership over written work, with nearly two-thirds (61%, 22 of 36 surveyed) reporting that the impact in relation to ownership was 'very positive', the highest proportion across all perceived impact questions on the survey. Of course there may be bias in these results, since teachers who were very positive about the intervention may be more likely to respond to surveys, and we acknowledge that a self-selected sample of teacher perceptions may not accurately reflect pupils' reality. There was also striking consistency among case study interviewees in terms of reporting that CoW had increased pupils' ownership over writing, often expressed in terms of pupils seeing themselves as writers or believing in themselves as writers. This was explained as stemming from the fact that CoW puts fewer restrictions on what their writing needs to look like and gives pupils more choice, for instance, over the topic that they write about.

> *'My class has developed a culture of freedom with their writing. So you might have seen today that they are quite comfortable with just sitting there and just writing what comes into their head. They don't worry about it, and they view themselves as authors in that way.' (Teacher 06)*

> *'...it is having a huge effect on the children's self-confidence and view of themselves in terms of writers. I think that is the greatest thing, that they see themselves as writers now.' (Teacher 04)*

As these quotes illustrate, increases in ownership were seen to occur alongside increased confidence around writing. Though confidence came through less strongly in interviewees' accounts than ownership, there were mentions of pupils being calmer and more composed while writing, writing more independently, and being able to get started with writing more easily, all of which are markers of writing confidence. As with the teacher mechanisms, it was also reported by the teachers that the intervention made children more enthusiastic about writing, as well as more confident. In the survey, 97% of respondents felt the intervention had a positive impact on engagement with writing activities (53% reporting the impact had been 'very positive'). This was echoed in some case study schools, where teaching staff reported that CoW had made children much more engaged in writing, demonstrated by pupils writing outside of school, and the fact that pupils who were previously uninterested in writing and wrote the bare minimum, now enjoyed writing exercises. However, particularly at schools with high rates of FSM, teachers reported that there were pockets of pupils, purportedly more boys, who continued to be reluctant writers.

> *'It's mixed. There is some great enthusiasm there. The free writing they particularly enjoy. With some of the more structured writing they struggle and where they will do their best to do as little work as possible. I have generally got quite a low class, and I've got a big cohort of boys who just do not like writing.' (Teacher 01)*

> *'I would say having taught that year group… [some of them] didn't want to write, or only wanted to write about the war. I think they have genuinely developed the love of writing and realise that they are writers.' (SLT member 04)*

The logic model has been updated to include the additional mechanism of pupils applying elements of the CoW framework in their written work and writing processes, based on case study teachers' reported perceptions of impact. This has been done to recognise that increased ownership and confidence around writing needs to be supported by skill development in order for outcomes such as writing attainment to be improved. There were a range of skills described across case study interviewees, including pupils using 'show not tell', incorporating different senses and emotions into their writing, and knowing how to receive feedback and use it to improve their work, rather than assuming their work will be right the first time. Whereas some case study interviewees described these skills as developing across the class, there were teachers who highlighted that doing things such as thinking about language choices and helping the reader to feel certain emotions was difficult for pupils with lower levels of attainment.

> *'We've got quite a high percentage of SEN in the class… So I think that some aspects of the framework is very much beyond them and they need support... thinking about language level choice and things like that. It's too far beyond them.' (Teacher 05)*

In terms of the overall outcomes of improved writing attainment and writing being more creative, 92% (33 of 36 surveyed) of survey respondents said that CoW had a positive impact on writing skill, while 97% (35 of 36 surveyed) said that it improved pupils' ability to write creatively. Notably, a smaller proportion of respondents felt that CoW led to a 'very positive' impact in these areas (22%, 8 of 36 surveyed, and 33%, 12 of 36 surveyed respectively), compared to writing ownership and engagement, where the proportions were over 50%. The case study data revealed there were two main perspectives in terms of overall outcomes. The first perspective was that impact was concentrated on children who had already achieved higher levels of writing attainment, and that children who had not either continued to struggle, or sometimes displayed small steps of progress but continued to produce writing of a lower quality. This viewpoint tended to be expressed by teachers who described their classes as having a high number of children with SEN and were also based in more deprived areas.

> *'I feel like it's had most impact, if I'm being honest, in our higher level writers... Whereas, this class this year academically aren't as strong as the previous class I had last year. We've got quite a high percentage of SEN in the class like [name of pupil]... that's a really good piece of writing [T], but compared to a lot you know, it is very low-level writing.' (Teacher 05)*

The other main perspective was that CoW was perceived to have led to improvements in writing quality across the year group, with interviewees stating that the writing standard was markedly better than in previous years, or that children were now demonstrating 'greater depth' writing, as a result of the intervention. Interviewees described how pupils in CoW classes were writing at greater length, writing more creatively and with greater flow, and using better language and phrasing. This perspective tended to be expressed by teaching staff at schools with lower rates of FSM, but was also expressed in one school with a high rate of FSM. Though it was acknowledged that there were still some children with lower writing attainment at this school who struggled with the approach, interviewees were of the opinion that the intervention was responsible for general improvements in writing quality.

> *'...that sense of flow is a really difficult thing to put your finger, but it tends to be what separates children who are working at a greater depth [from those at expected standard]. That sense of flow is apparent in almost all of the writing that they produce for this project, because the focus is on them writing naturally... So what I found is that there is a lot of children basically who are writing at a greater depth as a result of this.' (Teacher 06)*

> *'I think when we've looked at books at moderation... I've seen a marked improvement in the standard of writing in Year 5 compared to what we have seen in previous years... there are some really high-quality pieces that are absolutely amazing.' (SLT member 03)*

The IPE did not find evidence regarding self-efficacy beyond what has already been discussed in terms of confidence.

Cost

Delivery of the CoW intervention cost approximately £669 per school for the year it was delivered, and the majority of costs are realised in the first year. Materials costs included notebooks for pupils and CoW CPD training books. Expenses included travel and subsistence costs related to attending training days. Programme fees were paid to the developer.

To calculate the total cost per pupil over three years, we assumed the number of pupils would cumulatively increase from 25 pupils in Year 1, to 50 in Year 2 and 75 in Year 3. We assumed 25 pupils, as this was the number of treated pupils per classroom. Based on these assumptions, the total cost per pupil per year over three years is £9 (set out in Table 12). The cumulative cost breakdown is set out in Table 13 (all figures are rounded to the nearest pound).

Table 12: Cost of delivering the Craft of Writing (CoW)

Item	Type of cost	Average cost (£) (minimum, maximum)	Total cost over 3 years (£)	Total cost per pupil per year over 3 years (£)
Materials and printing	Ongoing	9 (0, 36)	27	<1
Expenses	First year	159 (27, 500)	159	2
Programme fees	First year	500	500	7
Total		668	687	9

Source: Cost case study interviews with teachers (*n* = 4)

Table 13: Cumulative costs of the Craft of Writing (assuming delivery over three years)

	Year 1	Year 2	Year 3
The Craft of Writing (CoW)	£669	£678	£687

Source: Cost case study interviews with teachers (*n* = 4)

Training

To support the delivery of the intervention, schools allocated a teacher to attend two residential and three CPD sessions. Of the four cost evaluation case schools, three used internal supply cover at no additional cost to cover the days teachers were absent; only one school reported that they used a supply teacher to cover the days. On average, schools reported spending one hour arranging supply cover.

In terms of staff time, training (including residential days) was estimated at 36 hours (or 4.5 days) on average per school. Assuming the same teacher is involved in supporting the intervention for three years, the teacher or TA would only have to attend the training once.

Preparation

There was some additional administrative time associated with the residential and CPD days. Cost survey respondents reported spending, on average, four hours over the course of the intervention preparing for the CoW training.

Conclusion

Impact evaluation and IPE integration

The goal of the impact evaluation was to answer three key questions:

1. What is the effect of participating in the CoW over the course of one school year on pupils' writing skills?
2. What is the effect of participating in the CoW over the course of one school year on pupils' writing self-efficacy?
3. Does participating in the CoW over the course of one school year have an impact on pupils' perception of their own capacity to generate ideas?

The IPE focused on addressing RQs related to the implementation, delivery and perceived impact of the intervention; the IPE was designed to explore questions that were overarching across the five Learning about Culture interventions, as well as questions specific to CoW.

Evidence to support the logic model

CoW is an intervention that supports the development of teachers as writers with the goal of this knowledge being applied to help pupils in their development as writers. The original logic model (Figure 1) suggested that there were two pupil-focused mechanisms through which the CoW intervention would have an effect:

1. a stronger sense of ownership and agency of writing;
2. increased confidence as writers.

These two mechanisms were hypothesised to lead to three outcomes for pupils:

1. a higher level of self-efficacy;
2. better attainment in writing;
3. more creative writing.

The impact evaluation results for this trial found no substantial effect of the CoW intervention on pupils' writing skills, the primary outcome measure. The overall effect on the WAM (the primary outcome) was small and slightly negative. It is not statistically significant. The impact evaluation results on the secondary outcome measures: the WSEM and the ideation score, were small and positive, but not statistically significant. These results were all small in magnitude, equivalent to zero months' progress. Overall, there was no significant impact of the CoW on any outcomes for the full sample.

Nevertheless, teachers enjoyed participating in this intervention and were positive in their perceptions of it. Survey evidence from the IPE showed that teachers perceived the intervention as highly impactful in terms of pupils' writing confidence and ownership over writing. A high degree of enthusiasm and enjoyment of writing on behalf of the pupils was also expressed by the participating teachers. These results may be biased, however, by the more engaged teachers responding to the survey at higher rates. Again, the impact evaluation results for pupils' writing self-efficacy and ideation did not support these IPE findings, as the results showed no statistically significant impact of participation in the CoW on these outcomes.

The IPE found evidence supporting the majority of elements in the hypothesised pathway. The logic model was updated to include pupils' enthusiasm about writing as part of the second mechanism. An additional mechanism was also added: pupils applying the CoW framework in their written work. The WSEM captures some elements of enthusiasm and again there was no statistically significant impact of participation in the CoW on this measure.

In the original logic model, there were three teacher-focused mechanisms through which the intervention was anticipated to have an effect: (i) increased confidence as writers; (ii) stronger subject knowledge of the craft of writing; and (iii) focusing teaching on using craft knowledge and fostering a community of creative writers.

To reflect evidence from the IPE, these have been revised to:

1 teachers feel more confident and enthusiastic about writing;

2 teachers understand how to implement the approach;

3 teachers believe the approach is beneficial for their class;

4 teachers routinely teach writing using the CoW approach.

The mechanisms chain was envisaged to lead to one teacher-focused outcome: teachers being more effective at teaching writing. This outcome remains unchanged from the original logic model. The impact evaluation results did not support this teacher-focused outcome of teachers being more effective at teaching writing, as there was no statistically significant impact of the CoW programme on pupils' writing attainment.

Due to the nature of the outcome data collected for this trial, it was not possible to explore how the intervention affected teacher outcomes in the impact evaluation. For example, it was not possible to evaluate how teachers changed their teaching and own writing practices as a result of the intervention. It was also not clear whether teachers had stronger subject knowledge or if they actually fostered a community of writers as a result of the trial. These outcomes would be an interesting area for further research.

Interpretation

This intervention was challenging for participating teachers, although they reported greatly enjoying it and perceived the intervention as being beneficial both to themselves and to their pupils. Teachers found it difficult to reconcile using the new materials obtained during their residential weekends and CPD days with following the national curriculum. This made the intervention challenging to incorporate into their daily practice, and there were some examples of teachers not implementing the intervention as intended. There was also evidence from the surveys that some elements of the CoW framework were used similarly frequently by teachers in both treatment and control schools. It is possible that the intervention did not vary enough from business-as-usual to have an observable impact on pupil outcomes. The results from the complier analysis showed no difference between the overall impact evaluation results and the CACE results. There may be issues with the compliance measure used in this intervention, however, since it did not place any requirements on how or whether teachers implemented the CoW intervention. It is possible that teachers attended the residential sessions and CPD days as required for compliance, but never implemented any of the CoW in their classroom.

Those teachers who responded to the survey were very positive in their assessment of the CoW intervention and its effect on pupils as writers and the strength and quality of their writing. The teachers who participated in this intervention were self-nominated, displaying a high degree of motivation for the intervention, and it is possible that their perceptions were biased by this, relative to how pupils were actually affected by the intervention.

The existing evidence on whether teachers' writing practice impacts pupil outcomes showed mixed results. A systematic review by Cremin & Oliver (2017) found that the evidence base regarding the impact of teachers' writing on pupils' outcomes is small and does not show a clear impact. They only identify 22 studies from the period 1990–2015 that meet their standard of inclusion. The results from this trial support this existing evidence, since the effect sizes obtained are small in magnitude and are unclear in direction of the effect (primary outcome's effect size is slightly negative, secondary outcome effect sizes are slightly positive). The findings from this trial weakly confirm some of the other existing evidence on teachers' skills, knowledge and confidence in facilitating creative writing (Redmond, 2010), and on pupils' attitudes and engagement as writers (Wilson, 2010). The results of the FSM sub-group analysis showed three months' progress on ideation or creativity as a result of participating in the CoW intervention, providing some support to the existing evidence base.

It is also possible that this intervention will have a longer term impact on participating teachers, since it is an entire approach to teaching writing that has the potential to change their teaching practice in the longer term. However, we do not have evidence to support this assertion and it is not clear from the evidence we have generated how these longer term changes might come about unless teachers continue to embed the CoW framework into their teaching practice over the next several years.

Limitations and lessons learnt

The CoW intervention was found to be challenging for teachers to implement because it is an entire approach and not just a series of tasks. It requires teachers to reflect on and update their entire approach to teaching writing and their own writing process. In addition, some schools flagged that there was a perceived conflict between CoW and implementation of an alternative writing scheme such as *Introduction, Point, Explain, Ending, Links and Language (IPEELL)* or *Talk for Writing*. This highlights the importance of the school's SLT 'buying-in' to using the CoW approach, if it is to be successfully implemented.

Baseline balance checks of the sample revealed some imbalance across trial arms, with, on average, higher KS1 scores at the school level in the intervention group in comparison with the control group (this is only present in the analysis sample, not the randomised sample, implying this could be induced by systematic attrition, discussed further below). The intervention schools also had higher proportions of FSM-eligible pupils and EAL pupils. We acknowledge the potential for this to lead to bias in our point estimates (although to the extent that it is due to chance in the randomisation, this is part of the uncertainty in estimates captured by the CI). Other things equal, higher KS1 scores in the intervention group would imply our estimate to be an overestimate of the true impact (we have carried out robustness checks on this point) while, conversely, the imbalance in terms of FSM pupils might be expected to bias in the other direction. Furthermore, the imbalance in FSM and EAL are both likely to be mitigated by the inclusion of these factors as covariates in our primary analysis model.

Unfortunately, due to issues in accessing the NPD, planned pre-test measures (planned to be phonics screening check scores) were not available at point of analysis. This reduced the explanatory power of the outcome measure from covariates included in the primary analysis model compared to our expectations at design stage (although, of course, these were assumptions and so could still have turned out to be too optimistic, even if we did have phonics screening check scores available). This, combined with greater attrition than assumed and a higher ICC of the outcome measure than was anticipated (based on our assumptions and previous EEF guidance), led to a substantial increase in the MDES between the design stage and the analysis stage. This increases the uncertainty around our impact estimates, and hence decreases the likelihood of finding a statistically significant impact of the intervention for a given effect size. The higher than anticipated ICC of the outcome measure (which we judge unlikely to be attributable to the composition of our sample of schools) suggests the need for further preparatory work in understanding this aspect of non-NPD outcome measures ahead of trial design – previous literature on the outcome measure did not cover school-level clustering and neither could our small-scale piloting.

Robust assessment of writing is challenging, particularly during primary schooling. There are few measures available, and none have been used in a similar context. The measure used in this trial, the WAM (Dunsmuir et al., 2015), was a pragmatic choice, which comes with some limitations (e.g., in terms of how relatively new it is); however, existing evidence suggests that it is a valid, consistent and reliable measure. Nevertheless, despite the many dimensions of writing it aims to capture, it may not have been able to capture those proximately affected by this intervention. This does present a limitation of this evaluation, albeit one we feel we have done all we could to address given the measures available.

It was also more difficult to conduct this outcome data collection than had been anticipated, resulting in substantial differences in when schools were tested. In theory, this meant that some schools received more time being taught using the CoW intervention since they were tested later. However, a robustness check including the date of testing did not reveal any substantial differences in the impact evaluation results.

Following randomisation, eight intervention and 11 control schools were lost to follow-up. This led to substantial attrition for this trial. We have to acknowledge that attrition carries with it a risk of bias (Sterne et al., 2019) and one that cannot

typically be remedied analytically. In short, those schools and pupils with missing outcomes data could have affected the result from the trial in different ways – meaning that the result is biased away from the 'true' estimate of impact.

Additionally, 21 of the 47 schools allocated to the treatment group did not meet the minimum compliance threshold (i.e., the participating teacher did not attend both residential weekends and at least two of the three CDP days). The compliance analysis for this trial did not change the overall impact evaluation results, indicating that non-compliance did not drive the main results. Of course, the aforementioned limitations of the compliance measure may also mask key issues with fidelity of delivery.

Although participating teachers were generally positive in their assessment of the intervention, there were concerns raised about the level of buy-in from the SLT which, coupled with pressure not to deviate from the national curriculum, may have affected the implementation of the intervention. Again, these concerns were not confirmed by the compliance analysis, but our compliance measures did not capture how and when teachers actually delivered the intervention.

As a result, it would have been useful and interesting to collect data on how teachers actually implemented the CoW in their teaching practice, as well as outcome data on the teachers, in order to better understand how this intervention might have a longer term impact. Future interventions which clearly target the longer term skill development of teachers could consider this.

Generally, the data collected from intervention schools as part of the IPE (either via survey or fieldwork visits) only represent the views and experiences of a subset of the larger treatment sample. The qualitative findings are not necessarily representative and should be considered within the context of these limitations. Additionally, there may be some recall errors in survey responses.

Similarly, the case study approach to the cost evaluation represents the range and diversity of costs encountered in implementation among highly engaged schools. Sampling was done to capture variation of spend among schools with high/low proportions of FSM pupils and per pupil spending, as these were hypothesised to correlate with costs. It is possible that the sampled schools were not representative of typical costs of full implementation of the programme. However, the programme fee was the largest cost to schools and did not vary by school. Spending ranges were provided for other direct costs, so that prospective schools could consider costs they may encounter above and beyond paying for the programme itself.

Future research and publications

The results from the IPE show that the CoW intervention challenged teachers to rethink their own writing practice and how they teach writing. Overall, however, the impact evaluation showed no effect of the CoW intervention on any of the outcome measures. Since the CoW is an entire teaching approach, it has the potential to have a longer term impact than was measured in this trial (i.e., on future cohorts of pupils taught by this teacher). A follow-up evaluation of the KS2 results of pupils taught by participating teachers could shed light on this possibility. This trial also did not robustly measure any teacher outcomes that could be used in the impact evaluation (e.g., baseline and follow-up data on teaching practices used and own writing self-efficacy), which would have been interesting to explore given the nature of the intervention.

An additional overarching report on all five Learning about Culture interventions funded by EEF and the RSA will be published in 2021. This will include the three KS2 Learning about Culture interventions (Craft of Writing, Power of Pictures and the Young Journalist Academy) and the two KS1 Learning about Culture interventions (First Thing Music and Speech Bubbles). This report will pool outcome data across the trials for a combined impact evaluation and synthesise IPE results across all interventions.

In addition, we plan to publish several academic journal articles summarising the findings from this and the other trials in the Learning about Culture project, as well as discussing their implications for future research on arts-based learning interventions and methodological issues, including the challenges of measuring children's writing skills at scale.

References

Angrist, J.D., Pischke, J.-S., 2008. *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Bloom, H.S., Richburg-Hayes, L., Black, A.R., 2007. Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis* 29, 30–59. https://doi.org/10.3102/0162373707299550

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.

Bulus, M., Dong, N., Kelcey, B., Spybrook, J., 2019. *PowerUpR: Power Analysis Tools for Multilevel Randomized Experiments. R package version 1.0.4*. https://CRAN.R-project.org/package=PowerUpR

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., 2012. *The arts and achievement in at-risk youth: Findings from four longitudinal studies*. Research Report# 55. National Endowment for the Arts.

Cremin, T., Oliver, L., 2017. Teachers as writers: A systematic review. *Research Papers in Education* 32, 269–295.

Cunningham, S., 2021. *Causal inference: The mixtape*. Yale University Press.

DfE, 2016. *Schools, pupils and their characteristics: January 2016*, SFR20/2016. Department for Education, London.

DfE, 2017. *Progress in International Reading Literacy Study (PIRLS): National Report for England*. Department for Education, London.

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.

EEF, 2015. *Intra-cluster correlation coefficients*. Education Endowment Foundation, London

EEF, 2018. *Statistical analysis guidance for EEF evaluations*. Education Endowment Foundation, London.

Gerber, A.S., Green, D.P., 2012. *Field experiments: Design, analysis, and interpretation*. WW Norton.

Gutmann, L., Schoon, I., 2013. *The impact of non-cognitive skills on outcomes for young people: Literature review*. Education Endowment Foundation, London.

Hedges, L.V., 1981. Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics* 6, 107–128. https://doi.org/10.3102/10769986006002107

Ilie, S., Sutherland, A., Vignoles, A., 2017. Revisiting free school meal eligibility as a proxy for pupil socio-economic deprivation. *British Educational Research Journal* 43, 253–274. https://doi.org/10.1002/berj.3260

Imbens, G.W., Rubin, D.B., 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

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.

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.

Pajares, F., 1996. Self-efficacy beliefs in academic settings. *Review of Educational Research* 66, 543–578.

Redmond, C., 2010. *Even better writers: Final evaluation report.* Redmond Management Limited.

Ritchie, J., Lewis, J., Nicholls, C.M., Ormston, R., 2013. *Qualitative research practice: A guide for social science students and researchers*. Sage.

Rubin, D.B., 2004. *Multiple imputation for nonresponse in surveys*. John Wiley & Sons.

Sterne, J.A., Savović, J., Page, M.J., Elbers, R.G., Blencowe, N.S., Boutron, I., Cates, C.J., Cheng, H.-Y., Corbett, M.S., Eldridge, S.M., 2019. RoB 2: A revised tool for assessing risk of bias in randomised trials. *The BMJ* 366.

Strand, S., Malmberg, L., Hall, J., 2015. *English as an additional language (EAL) and educational achievement in England: An analysis of the National Pupil Database*. London: Education Endowment Foundation.

Torgerson, C.J., Torgerson, D.J., 2013. *Randomised trials in education: An introductory handbook*. London: Education Endowment Foundation.

UK Government, 2019. Get information about schools. [online].

Wilson, A., 2010. *Teachers' conceptualisations of the intuitive and the intentional in poetry composition.* English Teaching: Practice & Critique (University of Waikato) 9.

Appendix A: EEF cost rating

Figure A1: Cost rating

Cost rating	Description
£ £ £ £ £	*Very low:* less than £80 per pupil per year.
£ £ £ £ £	*Low:* up to about £200 per pupil per year.
£ £ £ £ £	*Moderate:* up to about £700 per pupil per year.
£ £ £ £ £	*High:* up to £1,200 per pupil per year.
£ £ £ £ £	*Very high:* over £1,200 per pupil per year.

Appendix B: Security classification of trial findings

OUTCOME: Writing Assessment Measure (WAM)

Rating	Criteria for rating			Initial score	Adjust	Final score
	Design	MDES	Attrition		Adjustment for threats to internal validity [0]	
5	Randomised design	<= 0.2	0–10%			
4	Design for comparison that considers some type of selection on unobservable characteristics (e.g., RDD, Diff-in-Diffs, Matched Diff-in-Diffs)	0.21 – 0.29	11–20%			
3	Design for comparison that considers selection on all relevant observable confounders (e.g., Matching or Regression Analysis with variables descriptive of the selection mechanism)	0.30 – 0.39	21–30%			
2	Design for comparison that considers selection only on some relevant confounders	0.40 – 0.49	31–40%	2		2
1	Design for comparison that does not consider selection on any relevant confounders	0.50 – 0.59	41–50%			
0	No comparator	>=0.6	>50%			

Threats to validity	Risk rating	Comments
Threat 1: Confounding	Low	It is not entirely clear how many pupils were randomised to the intervention, with the numbers being revised by the time the SAP was written. The reasons for this are explained in the report.
Threat 2: Concurrent Interventions	Low	Although schools were frequently involved in other literacy programmes, there is no reason to expect that there were any systematic differences between intervention and control schools
Threat 3: Experimental effects	Low	IPE for control group schools did not explore any possible changes in practice.

Threat 4: Implementation fidelity	Low/ Moderate	Compliance in terms of attendance at the training sessions was low. Information about how CoW was implemented in classrooms was based mainly on a small number of case studies, which suggested considerable variation in implementation practice.
Threat 5: Missing data	Moderate	There was rather a lot of missing data, with over 30% of the randomisation sample missing. While this is a threat to validity, two padlocks have already been dropped for attrition, so no further adjustments are necessary. The lack of NPD and pre-test data also causes some concerns here, though there is little more the authors could have done under the circumstances explained in the report.
Threat 6: Measurement of outcomes	Low/ Moderate	Some concerns about the primary outcome variable, which the authors address at length in the report. Piloting, administration and marking procedures reported provide adequate reassurance.
Threat 7: Selective reporting	Low	The study differs from the original protocol, but the reasons are justified to a satisfactory extent. Selective reporting not apparent.

- **Initial padlock score:** 2 Padlocks – Randomised design with 0.22 MDES at randomisation and 34.8% attrition (or 30.6% if using the revised figures in the SAP, with uncertainty around the number of pupils randomised).
- **Reason for adjustment for threats to validity**: N/A – One moderate and two low/moderate risks, with the direction of likely biases unclear. No further adjustments required.

Final padlock score: initial score adjusted for threats to validity = 2 Padlocks

Appendix C: Effect size estimation

Table C1: Effect size estimation

			Intervention group		Control group			
Outcome	Unadjusted differences in means	Adjusted differences in means	*n* (missing)	SD of outcome	*n* (missing)	SD of outcome	Pooled SD	Hedges' correction (*J*)
WAM score (ideas double weighted)	–0.593	–0.15	873 (400)	4.24	824 (507)	4.57	4.4	0.99956
WSEM score	0.319	0.37	861 (412)	10.12	813 (518)	10.51	10.31	0.99955
Ideation score	0.09	0.08	861 (412)	3.44	813 (518)	3.52	3.48	0.99955

Note. SD refers to standard deviation.

Further appendices

Appendix D: Additional tables

Table D1: Baseline characteristics of groups as analysed (n = pupil, N = school)

School-level (categorical)	National-level percentage	Intervention group		Control group		
		N (missing)	%	*N* (missing)	%	
Setting: Urban	87.3	36 (0)	92.3	36 (0)	100.0	
Setting: Rural	12.7	3 (0)	7.7	0 (0)	0.0	
Ofsted: Outstanding	17.1	9 (0)	24.9	8 (4)	26.6	
Ofsted: Good	69.4	26 (0)	64.7	21 (4)	63.9	
Ofsted: Requires improvement / Inadequate	13.4	4 (0)	10.4	3 (4)	9.5	
School type: Academy	23.6	12 (0)	10.4	8 (0)	9.5	
School type: Community	41.2	17 (0)	30.1	17 (0)	21.5	
School type: Other	35.2	10 (0)	44.4	11 (0)	48.7	
School-level (continuous)	National-level mean	*N* (missing)	Mean (SD)	*N* (missing)	Mean (SD)	Standardised difference
KS1 average performance	15.9	35 (4)	15.76 (0.99)	34 (2)	15.58 (1.20)	0.164
Pupil-level (categorical)	National-level percentage	*n* (missing)	%	*n* (missing)	%	
FSM	30.9	288 (0)	33.0	207 (0)	25.1	
Non-FSM	69.1	585 (0)	67.0	617 (0)	74.9	
EAL	15.3	204 (0)	23.4	144 (0)	17.5	

Non-EAL	84.7	669 (0)	76.6	680 (0)	82.5	

Note. School-level imbalance is calculated applying weights for the size of the school, so that schools which are relatively more important in the pupil-level impact estimation are afforded the same importance in understanding imbalance.

Table D2: Sub-group analyses

Model	Hedges' *g* (95% CI)	*N*	*p*-value of interaction term	*p*-value of treatment variable
WAM FSM sub-group analysis	–0.04 (–0.28, 0.20)	495	0.66	0.73
WSEM FSM sub-group analysis	–0.02 (–0.11, 0.08)	483	0.07	0.73
Ideation FSM sub-group analysis	0.22 (0.01, 0.42)	483	0.02	0.04

Table D3: Full sample summary statistics

Outcome	Mean	SD	ICC	*N*
WAM Score (ideas scale double weighted)	18.09	4.41	0.33	1697
WSEM-16 Score	64.57	10.31	0.14	1674
Ideation Score	20.04	3.48	0.15	1674

Table D4: Sampling and randomisation inference p-values

Outcome	Effect size	Sampling Inference *p*-value	Randomisation Inference *p*-value
WAM Score (ideas scale double weighted)	–0.03	0.68	0.75
WSEM Score	0.04	0.55	0.61
Ideation Score	0.02	0.69	0.74

Table D5: Complier average causal effect (CACE) analysis

Model	Hedges' *g* (95% CI)	*N*	First stage *F* test	Compliance / Treatment correlation	*p*-value of treatment variable
Compliance analysis	–0.05 (–0.30, 0.20)	1697	F(12, 74) = 14	0.62	0.67

Table D6: Robustness checks

Model	Effect size (Hedges' *g*)	*N*	Treatment coefficient	*p*-value of treatment variable
Primary impact evaluation	–0.03	1697	–0.148	0.676
Test date	–0.04	1671	–0.169	0.636
Marker fixed effects	–0.01	1697	–0.062	0.865
School level KS1	–0.04	1697	–0.176	0.625

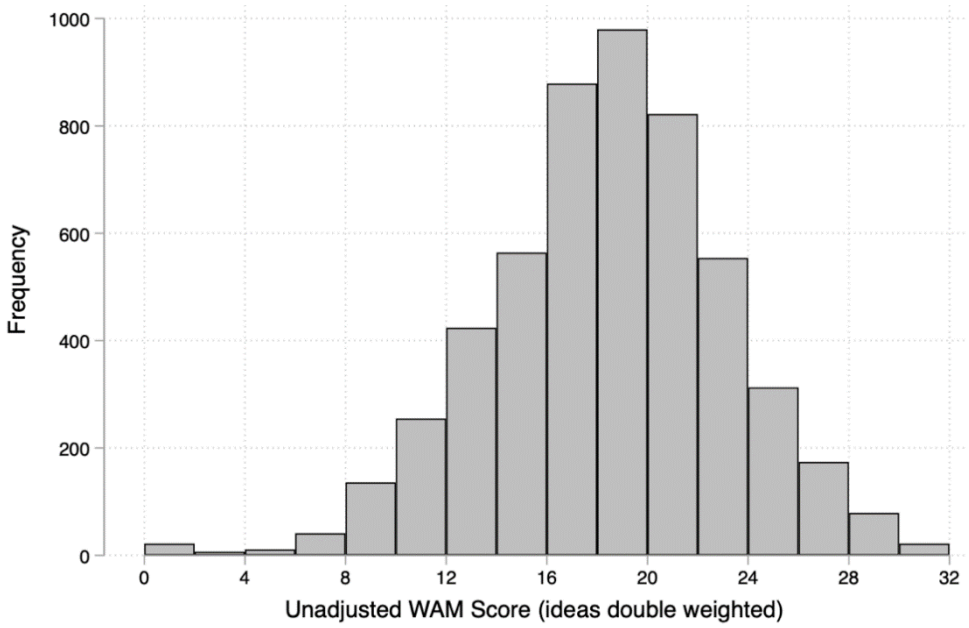
Table D7: Number of treatment and control schools (N) in each stratum

Randomisation Batch	Stratum	Treatment (*N*)	Control (*N*)
1	Low EAL/Low FSM	6	6
1	Low EAL/High FSM	4	5
1	High EAL/Low FSM	5	4
1	High EAL/High FSM	6	6
2	Low EAL/Low FSM	7	7
2	Low EAL/High FSM	6	6
2	High EAL/Low FSM	6	6
2	High EAL/High FSM	7	7

Note. Time of sign up resulted in two randomisation batches. The strata are defined by the intersection of English as an additional language (EAL) pupils (high/low split at sample median within randomisation batch) and proportion of free school meals (FSM) pupils (high/low split at sample median within randomisation batch).

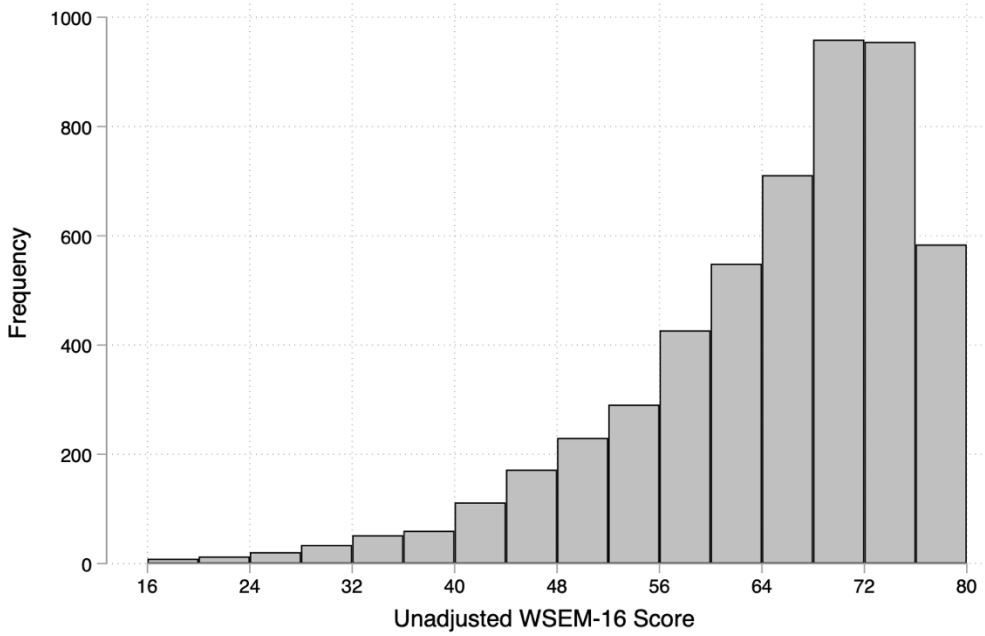
Appendix E: Additional histograms

Figure E1: Histogram of WAM scores pooled across all three KS2 Learning about Culture trials



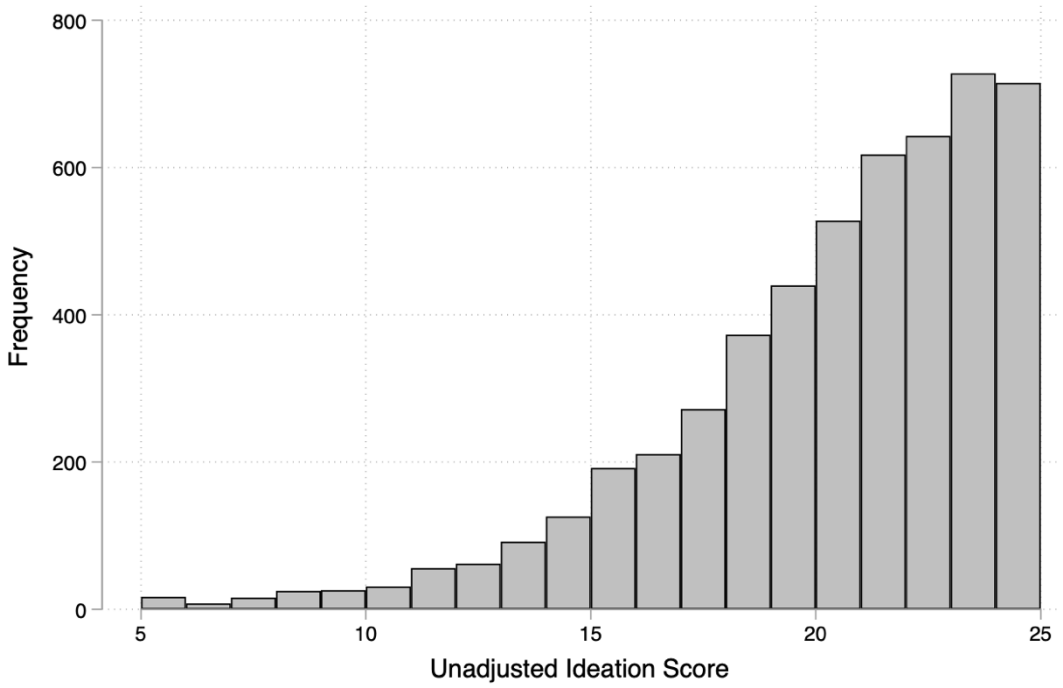
Notes. This histogram contains all WAM outcome data collected for the KS2 Learning about Culture trials (Craft of Writing, Power of Pictures and Young Journalist Academy). These data were collected as part of a single collection and marking exercise and, as such, we consider it together for the purposes of setting out details of the outcome measures.

Figure E2: Histogram of WSEM scores pooled across all three KS2 Learning about Culture trials



Notes. This histogram contains all WSEMM outcome data collected for the KS2 Learning about Culture trials (Craft of Writing, Power of Pictures and Young Journalist Academy) These data were collected as part of a single collection and marking exercise and, as such, we consider it together for the purposes of setting out details of the outcome measures.

Figure E3: Histogram of Ideation scores pooled across all three KS2 Learning about Culture trials



Notes. This histogram contains all ideation outcome data collected for the KS2 Learning about Culture trials (Craft of Writing, Power of Pictures and Young Journalist Academy). These data were collected as part of a single collection and marking exercise and, as such, we consider it together for the purposes of setting out details of the outcome measures.

Figure E4: Histogram of unadjusted ideation scores by treatment arm

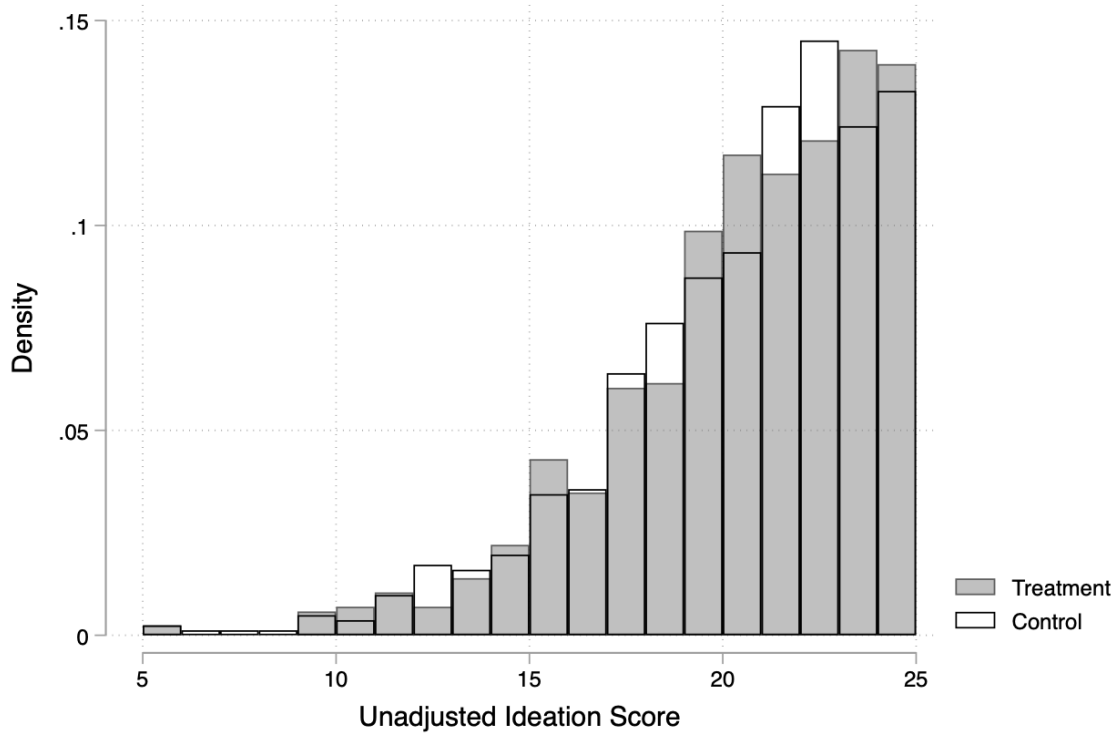
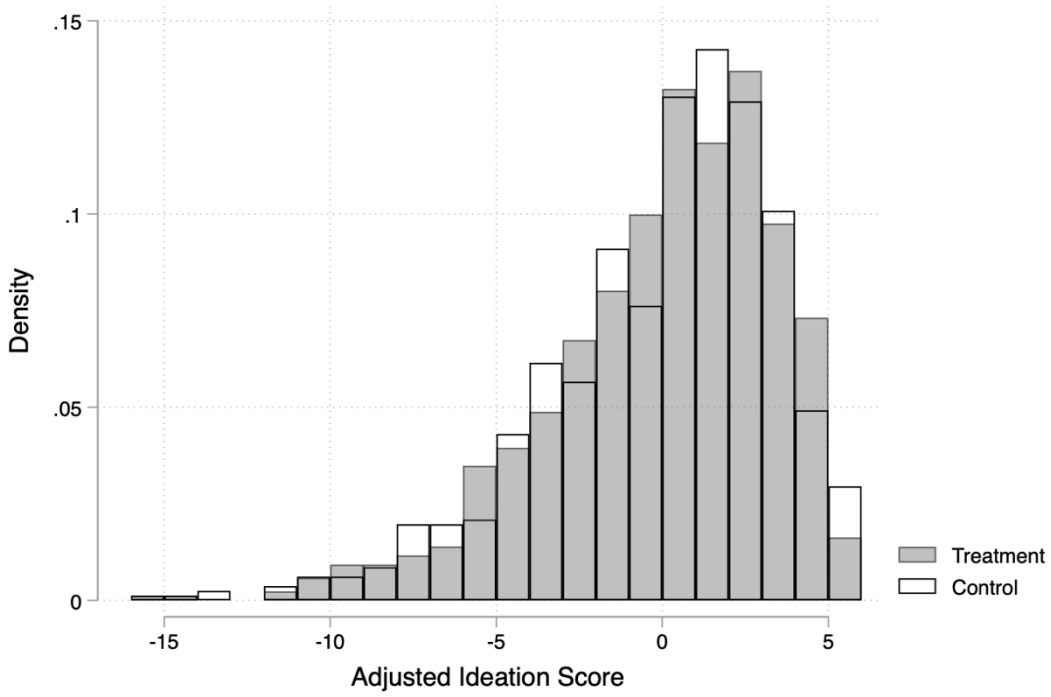


Figure E5: Histogram of adjusted ideation scores by treatment arm



Appendix F: Analysis syntax

```
ssc install blindschemes

run '[INSERT FILE PATH HERE]/lac_ks2_nationalmeans.do' // Calculates national means for the balance tabless

global workd '[INSERT FILE PATH FOR GLOBAL MACRO HERE]'

global datad '[INSERT FILE PATH FOR GLOBAL MACRO HERE]'

cd '${workd}'

clear

use '${datad}/cow_data.dta'


* Create sample variables

cap drop primarysample

gen primarysample = 0

replace primarysample = 1 if wam_score_dw<. & FSM<. & EAL<. & group<. & block<. & anonschoolid<. & treat<.


count if primarysample == 1

local primarysample_n = r(N)

count if primarysample == 1 & treat==1

local primarysample_treat_n = r(N)

count if primarysample == 1 & treat==0

local primarysample_control_n = r(N)


count if primarysample == 0

local primarysample_miss = r(N)

count if primarysample == 0 & treat==1

local primarysample_treat_miss = r(N)

count if primarysample == 0 & treat==0

local primarysample_control_miss = r(N)


cap drop secondarysample

gen secondarysample = 0
```

```
replace secondarysample = 1 if wsem16_score<. & FSM<. & EAL<. & group<. & block<. & anonschoolid<. & treat<.

count if secondarysample == 1
local secondarysample_n = r(N)
count if secondarysample == 1 & treat==1
local secondarysample_treat_n = r(N)
count if secondarysample == 1 & treat==0
local secondarysample_control_n = r(N)

count if secondarysample == 0
local secondarysample_miss = r(N)
count if secondarysample == 0 & treat==1
local secondarysample_treat_miss = r(N)
count if secondarysample == 0 & treat==0
local secondarysample_control_miss = r(N)

cap drop randomisedsample
gen randomisedsample = 0
replace randomisedsample = 1 if group<. & block<. & anonschoolid<. & treat<.
count if randomisedsample == 1
local randomisedsample_n = r(N)
count if randomisedsample == 1 & treat==1
local randomisedsample_treat_n = r(N)
count if randomisedsample == 1 & treat==0
local randomisedsample_control_n = r(N)

local primaryattrition_n = `randomisedsample_n' - `primarysample_n'
local primaryattrition_treat_n = `randomisedsample_treat_n' - `primarysample_treat_n'
local primaryattrition_control_n = `randomisedsample_control_n' - `primarysample_control_n'
local primaryattrition_percent : di %7.1fc 100-((`primarysample_n'/`randomisedsample_n')*100)
local primaryattrition_treat_p : di %7.1fc 100-((`primarysample_treat_n'/`randomisedsample_treat_n')*100)
```

```
local primaryattrition_control_p : di %7.1fc 100-((`primarysample_control_n'/`randomisedsample_control_n')*100)

count if wam_score_dw < .
local primarymeasure_n = r(N)
count if wam_score_dw < . & treat==1
local primarymeasure_treat_n = r(N)
count if wam_score_dw < . & treat==0
local primarymeasure_control_n = r(N)

count if wam_score_dw >= .
local primarymeasure_miss = r(N)
count if wam_score_dw >= . & treat==1
local primarymeasure_treat_miss = r(N)
count if wam_score_dw >= . & treat==0
local primarymeasure_control_miss = r(N)

count if wsem16_score < .
local wsem_n = r(N)
count if wsem16_score < . & treat==1
local wsem_treat_n = r(N)
count if wsem16_score < . & treat==0
local wsem_control_n = r(N)

count if wsem16_score >= .
local wsem_miss = r(N)
count if wsem16_score >= . & treat==1
local wsem_treat_miss = r(N)
count if wsem16_score >= . & treat==0
local wsem_control_miss = r(N)

count if ideation_score < .
```

```
local idea_n = r(N)

count if ideation_score < . & treat==1

local idea_treat_n = r(N)

count if ideation_score < . & treat==0

local idea_control_n = r(N)


count if ideation_score >= .

local idea_miss = r(N)

count if ideation_score >= . & treat==1

local idea_treat_miss = r(N)

count if ideation_score >= . & treat==0

local idea_control_miss = r(N)


*Number of treatment and control schools in each block by randomisation group

forval i = 1(1)2{

quietly tab anonschoolid if block == 11 & treat == 0 & group ==`i'

local block11control_schoolcount_`i' = r(r)

quietly tab anonschoolid if block == 11 & treat == 1 & group ==`i'

local block11treat_schoolcount_`i' = r(r)


quietly tab anonschoolid if block == 12 & treat == 0 & group ==`i'

local block12control_schoolcount_`i' = r(r)

quietly tab anonschoolid if block == 12 & treat == 1 & group ==`i'

local block12treat_schoolcount_`i' = r(r)


quietly tab anonschoolid if block == 21 & treat == 0 & group ==`i'

local block21control_schoolcount_`i' = r(r)

quietly tab anonschoolid if block == 21 & treat == 1 & group ==`i'

local block21treat_schoolcount_`i' = r(r)


quietly tab anonschoolid if block == 22 & treat == 0 & group ==`i'
```

```
local block22control_schoolcount_`i' = r(r)
quietly tab anonschoolid if block == 22 & treat == 1 & group ==`i'
local block22treat_schoolcount_`i' = r(r)
}

* School-level descriptives
quietly tab anonschoolid
local asrandom_schoolcount = r(r)
quietly tab anonschoolid if treat==0
local asrandom_schoolcount_control = r(r)
quietly tab anonschoolid if treat==1
local asrandom_schoolcount_treat = r(r)

quietly tab anonschoolid if primarysample==1
local wam_schoolcount = r(r)
quietly tab anonschoolid if secondarysample==1
local wsem_schoolcount = r(r)
local idea_schoolcount = r(r)

quietly tab anonschoolid if treat==0 & primarysample==1
local wam_schoolcount_control = r(r)
quietly tab anonschoolid if treat==1 & primarysample==1
local wam_schoolcount_treat = r(r)

quietly tab anonschoolid if treat==0 & secondarysample==1
local wsem_schoolcount_control = r(r)
local idea_schoolcount_control = r(r)
quietly tab anonschoolid if treat==1 & secondarysample==1
local wsem_schoolcount_treat = r(r)
local idea_schoolcount_treat = r(r)
```

```
preserve
keep if primarysample==1
gen i=1
collapse (count) count = i, by(anonschoolid)
ameans count
local wam_nhmean : di %7.0fc r(mean_h)
restore

preserve
keep if secondarysample==1
gen i=1
collapse (count) count = i, by(anonschoolid)
ameans count
local wsem_nhmean : di %7.0fc r(mean_h)
local idea_nhmean : di %7.0fc r(mean_h)
restore

sum treat if primarysample==1
local wam_treatprop : di %7.2fc r(mean)
sum treat if secondarysample==1
local wsem_treatprop : di %7.2fc r(mean)
local idea_treatprop : di %7.2fc r(mean)

* School-level descriptives specific to FSM sub-group
preserve
keep if primarysample==1 & FSM==1
gen i=1
collapse (count) count = i, by(anonschoolid)
ameans count
local wam_fsm_nhmean : di %7.0fc r(mean_h)
restore
```

```
preserve
keep if secondarysample==1 & FSM==1
gen i=1
collapse (count) count = i, by(anonschoolid)
ameans count
local wsem_fsm_nhmean : di %7.0fc r(mean_h)
local idea_fsm_nhmean : di %7.0fc r(mean_h)
restore

sum treat if primarysample==1 & FSM==1
local wam_fsm_treatprop : di %7.2fc r(mean)
sum treat if secondarysample==1 & FSM==1
local wsem_fsm_treatprop : di %7.2fc r(mean)
local idea_fsm_treatprop : di %7.2fc r(mean)

* ICC Calculation
xtset anonschoolid
xtreg wam_score_dw
local wam_icc : di %7.2fc e(sigma_u) / (e(sigma_u) + e(sigma_e))
xtreg wsem16_score
local wsem_icc : di %7.2fc e(sigma_u) / (e(sigma_u) + e(sigma_e))
xtreg ideation_score
local idea_icc : di %7.2fc e(sigma_u) / (e(sigma_u) + e(sigma_e))

* ICC Calculation - FSM sub-group
xtset anonschoolid
xtreg wam_score_dw if FSM==1
local wam_fsm_icc : di %7.2fc e(sigma_u) / (e(sigma_u) + e(sigma_e))
xtreg wsem16_score if FSM==1
local wsem_fsm_icc : di %7.2fc e(sigma_u) / (e(sigma_u) + e(sigma_e))
```

```
xtreg ideation_score if FSM==1
local idea_fsm_icc : di %7.2fc e(sigma_u) / (e(sigma_u) + e(sigma_e))

* Group variables
cap drop FSMgroup
egen FSMgroup = mean(FSM), by(anonschoolid)
cap drop EALgroup
egen EALgroup = mean(EAL), by(anonschoolid)

* Estimate pre-post test correlations
xtset anonschoolid
xtreg wam_score_dw FSM EAL FSMgroup EALgroup
local wam_r2_pupil : di %7.2fc e(r2_w)
local wam_r2_school : di %7.2fc e(r2_b)
local wam_r2_overall : di %7.2fc e(r2_o)

xtreg wsem16_score FSM EAL FSMgroup EALgroup
local wsem_r2_pupil : di %7.2fc e(r2_w)
local wsem_r2_school : di %7.2fc e(r2_b)
local wsem_r2_overall : di %7.2fc e(r2_o)

xtreg ideation_score FSM EAL FSMgroup EALgroup
local idea_r2_pupil : di %7.2fc e(r2_w)
local idea_r2_school : di %7.2fc e(r2_b)
local idea_r2_overall : di %7.2fc e(r2_o)

* Estimate pre-post test correlations - FSM sub-group
xtset anonschoolid
xtreg wam_score_dw EAL FSMgroup EALgroup if FSM==1
local wam_fsm_r2_pupil : di %7.2fc e(r2_w)
local wam_fsm_r2_school : di %7.2fc e(r2_b)
```

```
local wam_fsm_r2_overall : di %7.2fc e(r2_o)

xtreg wsem16_score EAL FSMgroup EALgroup if FSM==1
local wsem_fsm_r2_pupil : di %7.2fc e(r2_w)
local wsem_fsm_r2_school : di %7.2fc e(r2_b)
local wsem_fsm_r2_overall : di %7.2fc e(r2_o)

xtreg ideation_score EAL FSMgroup EALgroup if FSM==1
local idea_fsm_r2_pupil : di %7.2fc e(r2_w)
local idea_fsm_r2_school : di %7.2fc e(r2_b)
local idea_fsm_r2_overall : di %7.2fc e(r2_o)

** Imbalance at baseline - as randomised
* FSM
sum FSM if treat==0 & randomisedsample==1
local fsm_controlpercent_in_rsample : di %7.1fc r(mean)*100
local fsm_controlpercent_out_rsample : di %7.1fc (1-r(mean))*100
local fsm_controlmean_rsample : di %7.2fc r(mean)
local fsm_controlsd_rsample : di %7.2fc r(sd)
local fsm_controln_rsample = r(N)
count if FSM==0 & treat==0 & randomisedsample==1
local fsm_controln_out_rsample = r(N)
count if FSM==1 & treat==0 & randomisedsample==1
local fsm_controln_in_rsample = r(N)
count if FSM>=. & treat==0 & randomisedsample==1
local fsm_controlmiss_rsample = r(N)
sum FSM if treat==1 & randomisedsample==1
local fsm_treatpercent_in_rsample : di %7.1fc r(mean)*100
local fsm_treatpercent_out_rsample : di %7.1fc (1-r(mean))*100
local fsm_treatmean_rsample : di %7.2fc r(mean)
local fsm_treatsd_rsample : di %7.2fc r(sd)
```

```
local fsm_treatn_rsample = r(N)
count if FSM==0 & treat==1 & randomisedsample==1
local fsm_treatn_out_rsample = r(N)
count if FSM==1 & treat==1 & randomisedsample==1
local fsm_treatn_in_rsample = r(N)
count if FSM>=. & treat==1 & randomisedsample==1
local fsm_treatmiss_rsample = r(N)
esizei `fsm_controln_rsample' `fsm_controlmean_rsample' `fsm_controlsd_rsample' ///
        `fsm_treatn_rsample' `fsm_treatmean_rsample' `fsm_treatsd_rsample', cohensd
local fsm_in_stddiff_rsample : di %7.3fc r(d)
local fsm_out_stddiff_rsample : di %7.3fc -r(d)

* EAL
sum EAL if treat==0 & randomisedsample==1
local eal_controlpercent_in_rsample : di %7.1fc r(mean)*100
local eal_controlpercent_out_rsample : di %7.1fc (1-r(mean))*100
local eal_controlmean_rsample : di %7.2fc r(mean)
local eal_controlsd_rsample : di %7.2fc r(sd)
local eal_controln_rsample = r(N)
count if EAL==0 & treat==0 & randomisedsample==1
local eal_controln_out_rsample = r(N)
count if EAL==1 & treat==0 & randomisedsample==1
local eal_controln_in_rsample = r(N)
count if EAL>=. & treat==0 & randomisedsample==1
local eal_controlmiss_rsample = r(N)
sum EAL if treat==1 & randomisedsample==1
local eal_treatpercent_in_rsample : di %7.1fc r(mean)*100
local eal_treatpercent_out_rsample : di %7.1fc (1-r(mean))*100
local eal_treatmean_rsample : di %7.2fc r(mean)
local eal_treatsd_rsample : di %7.2fc r(sd)
local eal_treatn_rsample = r(N)
```

```
count if EAL==0 & treat==1 & randomisedsample==1
local eal_treatn_out_rsample = r(N)
count if EAL==1 & treat==1 & randomisedsample==1
local eal_treatn_in_rsample = r(N)
count if EAL>=. & treat==1 & randomisedsample==1
local eal_treatmiss_rsample = r(N)
esizei `eal_controln_rsample' `eal_controlmean_rsample' `eal_controlsd_rsample' ///
        `eal_treatn_rsample' `eal_treatmean_rsample' `eal_treatsd_rsample', cohensd
local eal_in_stddiff_rsample : di %7.3fc r(d)
local eal_out_stddiff_rsample : di %7.3fc -r(d)

** Imbalance at baseline - as analysed
* FSM
sum FSM if treat==0 & primarysample==1
local fsm_controlpercent_in_psample : di %7.1fc r(mean)*100
local fsm_controlpercent_out_psample : di %7.1fc (1-r(mean))*100
local fsm_controlmean_psample : di %7.2fc r(mean)
local fsm_controlsd_psample : di %7.2fc r(sd)
local fsm_controln_psample = r(N)
count if FSM==0 & treat==0 & primarysample==1
local fsm_controln_out_psample = r(N)
count if FSM==1 & treat==0 & primarysample==1
local fsm_controln_in_psample = r(N)
count if FSM>=. & treat==0 & primarysample==1
local fsm_controlmiss_psample = r(N)
sum FSM if treat==1 & primarysample==1
local fsm_treatpercent_in_psample : di %7.1fc r(mean)*100
local fsm_treatpercent_out_psample : di %7.1fc (1-r(mean))*100
local fsm_treatmean_psample : di %7.2fc r(mean)
local fsm_treatsd_psample : di %7.2fc r(sd)
local fsm_treatn_psample = r(N)
```

```
count if FSM==0 & treat==1 & primarysample==1
local fsm_treatn_out_psample = r(N)
count if FSM==1 & treat==1 & primarysample==1
local fsm_treatn_in_psample = r(N)
count if FSM>=. & treat==0 & primarysample==1
local fsm_treatmiss_psample = r(N)
esizei `fsm_controln_psample' `fsm_controlmean_psample' `fsm_controlsd_psample' ///
        `fsm_treatn_psample' `fsm_treatmean_psample' `fsm_treatsd_psample', cohensd
local fsm_in_stddiff_psample : di %7.3fc r(d)
local fsm_out_stddiff_psample : di %7.3fc -r(d)

* EAL
sum EAL if treat==0 & primarysample==1
local eal_controlpercent_in_psample : di %7.1fc r(mean)*100
local eal_controlpercent_out_psample : di %7.1fc (1-r(mean))*100
local eal_controlmean_psample : di %7.2fc r(mean)
local eal_controlsd_psample : di %7.2fc r(sd)
local eal_controln_psample = r(N)
count if EAL==0 & treat==0 & primarysample==1
local eal_controln_out_psample = r(N)
count if EAL==1 & treat==0 & primarysample==1
local eal_controln_in_psample = r(N)
count if EAL>=. & treat==0 & primarysample==1
local eal_controlmiss_psample = r(N)
sum EAL if treat==1 & primarysample==1
local eal_treatpercent_in_psample : di %7.1fc r(mean)*100
local eal_treatpercent_out_psample : di %7.1fc (1-r(mean))*100
local eal_treatmean_psample : di %7.2fc r(mean)
local eal_treatsd_psample : di %7.2fc r(sd)
local eal_treatn_psample = r(N)
count if EAL==0 & treat==1 & primarysample==1
```

```
local eal_treatn_out_psample = r(N)

count if EAL==1 & treat==1 & primarysample==1

local eal_treatn_in_psample = r(N)

count if EAL>=. & treat==0 & primarysample==1

local eal_treatmiss_psample = r(N)

esizei `eal_controln_psample' `eal_controlmean_psample' `eal_controlsd_psample' ///

        `eal_treatn_psample' `eal_treatmean_psample' `eal_treatsd_psample', cohensd

local eal_in_stddiff_psample : di %7.3fc r(d)

local eal_out_stddiff_psample : di %7.3fc -r(d)


* Visualisation of outcome variables

graph twoway (kdensity wam_score_dw), scheme(plotplain) xtitle('Unadjusted WAM Score (ideas double weighted)') ytitle('Density')

graph export output/wam_kdensity.png, replace

graph twoway (kdensity wsem16_score), scheme(plotplain) xtitle('Unadjusted WSEM-16 Score') ytitle('Density')

graph export output/wsem_kdensity.png, replace

graph twoway (kdensity ideation_score), scheme(plotplain) xtitle('Unadjusted Ideation Score') ytitle('Density')

graph export output/ideation_kdensity.png, replace


* Descriptive statistics for outcome variables

sum wam_score_dw

local wam_mean : di %7.2fc r(mean)

local wam_sd : di %7.2fc r(sd)

local wam_n = r(N)

sum wsem16_score

local wsem_mean : di %7.2fc r(mean)

local wsem_sd : di %7.2fc r(sd)

local wsem_n = r(N)

sum ideation_score

local idea_mean : di %7.2fc r(mean)

local idea_sd : di %7.2fc r(sd)
```

```
local idea_n = r(N)

* Primary analysis
regress wam_score_dw i.treat FSM EAL FSMgroup EALgroup group#block, vce(cluster anonschoolid)
estimates store primary
local wam_treatdiff : di %7.2fc _b[1.treat]
local wam_treatdiff_upperci : di %7.2fc _b[1.treat] + 1.96*_se[1.treat]
local wam_treatdiff_lowerci : di %7.2fc _b[1.treat] - 1.96*_se[1.treat]
local wam_treatdiff_p : di %7.2fc ttail(e(df_r), abs(_b[1.treat]/_se[1.treat]))*2
margins treat, post
local wam_controlmean : di %7.2fc _b[0.treat]
local wam_controlmean_upperci : di %7.2fc _b[0.treat] + 1.96*_se[0.treat]
local wam_controlmean_lowerci : di %7.2fc _b[0.treat] - 1.96*_se[0.treat]
local wam_treatmean : di %7.2fc _b[1.treat]
local wam_treatmean_upperci : di %7.2fc _b[1.treat] + 1.96*_se[1.treat]
local wam_treatmean_lowerci : di %7.2fc _b[1.treat] - 1.96*_se[1.treat]

* Unadjusted version
regress wam_score_dw i.treat, vce(cluster anonschoolid)
local wam_treatdiff_unadj : di %7.2fc _b[1.treat]
local wam_treatdiff_unadj_upperci : di %7.2fc _b[1.treat] + 1.96*_se[1.treat]
local wam_treatdiff_unadj_lowerci : di %7.2fc _b[1.treat] - 1.96*_se[1.treat]
local wam_treatdiff_unadj_p : di %7.2fc ttail(e(df_r), abs(_b[1.treat]/_se[1.treat]))*2
margins treat, post
local wam_controlmean_unadj : di %7.2fc _b[0.treat]
local wam_controlmean_unadj_upperci : di %7.2fc _b[0.treat] + 1.96*_se[0.treat]
local wam_controlmean_unadj_lowerci : di %7.2fc _b[0.treat] - 1.96*_se[0.treat]
local wam_treatmean_unadj : di %7.2fc _b[1.treat]
local wam_treatmean_unadj_upperci : di %7.2fc _b[1.treat] + 1.96*_se[1.treat]
local wam_treatmean_unadj_lowerci : di %7.2fc _b[1.treat] - 1.96*_se[1.treat]
```

```
* Graphical version of primary analysis
regress wam_score_dw FSM EAL FSMgroup EALgroup group#block, vce(cluster anonschoolid)
cap drop wam_analysis_resid
predict wam_analysis_resid, resid
graph twoway ///
	(kdensity wam_score_dw if treat==1) ///
	(kdensity wam_score_dw if treat==0) ///
	, scheme(plotplain) legend(order(1 'Treatment' 2 'Control')) xtitle('Unadjusted WAM Score') ytitle('Density')
graph export output/wam_kdensity_treat_uncond.png, replace
graph twoway ///
	(kdensity wam_analysis_resid if treat==1) ///
	(kdensity wam_analysis_resid if treat==0) ///
	, scheme(plotplain) legend(order(1 'Treatment' 2 'Control')) xtitle('Adjusted WAM Score') ytitle('Density')
graph export output/wam_kdensity_treat_cond.png, replace

* Secondary analysis - WSEM
regress wsem16_score i.treat FSM EAL FSMgroup EALgroup group#block, vce(cluster anonschoolid)
local wsem_treatdiff : di %7.2fc _b[1.treat]
local wsem_treatdiff_upperci : di %7.2fc _b[1.treat] + 1.96*_se[1.treat]
local wsem_treatdiff_lowerci : di %7.2fc _b[1.treat] - 1.96*_se[1.treat]
local wsem_treatdiff_p : di %7.2fc ttail(e(df_r), abs(_b[1.treat]/_se[1.treat]))*2
margins treat, post
local wsem_controlmean : di %7.2fc _b[0.treat]
local wsem_controlmean_upperci : di %7.2fc _b[0.treat] + 1.96*_se[0.treat]
local wsem_controlmean_lowerci : di %7.2fc _b[0.treat] - 1.96*_se[0.treat]
local wsem_treatmean : di %7.2fc _b[1.treat]
local wsem_treatmean_upperci : di %7.2fc _b[1.treat] + 1.96*_se[1.treat]
local wsem_treatmean_lowerci : di %7.2fc _b[1.treat] - 1.96*_se[1.treat]

* Unadjusted version
regress wsem16_score i.treat, vce(cluster anonschoolid)
```

```
local wsem_treatdiff_unadj : di %7.2fc _b[1.treat]
local wsem_treatdiff_unadj_upperci : di %7.2fc _b[1.treat] + 1.96*_se[1.treat]
local wsem_treatdiff_unadj_lowerci : di %7.2fc _b[1.treat] - 1.96*_se[1.treat]
local wsem_treatdiff_unadj_p : di %7.2fc ttail(e(df_r), abs(_b[1.treat]/_se[1.treat]))*2
margins treat, post
local wsem_controlmean_unadj : di %7.2fc _b[0.treat]
local wsem_controlmean_unadj_upperci : di %7.2fc _b[0.treat] + 1.96*_se[0.treat]
local wsem_controlmean_unadj_lowerci : di %7.2fc _b[0.treat] - 1.96*_se[0.treat]
local wsem_treatmean_unadj : di %7.2fc _b[1.treat]
local wsem_treatmean_unadj_upperci : di %7.2fc _b[1.treat] + 1.96*_se[1.treat]
local wsem_treatmean_unadj_lowerci : di %7.2fc _b[1.treat] - 1.96*_se[1.treat]

* Graphical version of secondary analysis - WSEM
regress wsem16_score FSM EAL FSMgroup EALgroup group#block, vce(cluster anonschoolid)
cap drop wsem_analysis_resid
predict wsem_analysis_resid, resid
graph twoway ///
        (kdensity wsem16_score if treat==1) ///
        (kdensity wsem16_score if treat==0) ///
        , scheme(plotplain) legend(order(1 'Treatment' 2 'Control')) xtitle('Unadjusted WSEM Score') ytitle('Density')
graph export output/wsem_kdensity_treat_uncond.png, replace
graph twoway ///
        (kdensity wsem_analysis_resid if treat==1) ///
        (kdensity wsem_analysis_resid if treat==0) ///
        , scheme(plotplain) legend(order(1 'Treatment' 2 'Control')) xtitle('Adjusted WSEM Score') ytitle('Density')
graph export output/wsem_kdensity_treat_cond.png, replace

* Secondary analysis - Ideation
regress ideation_score i.treat FSM EAL FSMgroup EALgroup group#block, vce(cluster anonschoolid)
local idea_treatdiff : di %7.2fc _b[1.treat]
local idea_treatdiff_upperci : di %7.2fc _b[1.treat] + 1.96*_se[1.treat]
```

```
local idea_treatdiff_lowerci : di %7.2fc _b[1.treat] - 1.96*_se[1.treat]
local idea_treatdiff_p : di %7.2fc ttail(e(df_r), abs(_b[1.treat]/_se[1.treat]))*2
margins treat, post
local idea_controlmean : di %7.2fc _b[0.treat]
local idea_controlmean_upperci : di %7.2fc _b[0.treat] + 1.96*_se[0.treat]
local idea_controlmean_lowerci : di %7.2fc _b[0.treat] - 1.96*_se[0.treat]
local idea_treatmean : di %7.2fc _b[1.treat]
local idea_treatmean_upperci : di %7.2fc _b[1.treat] + 1.96*_se[1.treat]
local idea_treatmean_lowerci : di %7.2fc _b[1.treat] - 1.96*_se[1.treat]

regress ideation_score i.treat, vce(cluster anonschoolid)
local idea_treatdiff_unadj : di %7.2fc _b[1.treat]
local idea_treatdiff_unadj_upperci : di %7.2fc _b[1.treat] + 1.96*_se[1.treat]
local idea_treatdiff_unadj_lowerci : di %7.2fc _b[1.treat] - 1.96*_se[1.treat]
local idea_treatdiff_unadj_p : di %7.2fc ttail(e(df_r), abs(_b[1.treat]/_se[1.treat]))*2
margins treat, post
local idea_controlmean_unadj : di %7.2fc _b[0.treat]
local idea_controlmean_unadj_upperci : di %7.2fc _b[0.treat] + 1.96*_se[0.treat]
local idea_controlmean_unadj_lowerci : di %7.2fc _b[0.treat] - 1.96*_se[0.treat]
local idea_treatmean_unadj : di %7.2fc _b[1.treat]
local idea_treatmean_unadj_upperci : di %7.2fc _b[1.treat] + 1.96*_se[1.treat]
local idea_treatmean_unadj_lowerci : di %7.2fc _b[1.treat] - 1.96*_se[1.treat]

* Graphical version of secondary analysis - ideation
regress ideation_score FSM EAL FSMgroup EALgroup group#block, vce(cluster anonschoolid)
cap drop ideation_analysis_resid
predict ideation_analysis_resid, resid
graph twoway ///
        (kdensity ideation_score if treat==1) ///
        (kdensity ideation_score if treat==0) ///
        , scheme(plotplain) legend(order(1 'Treatment' 2 'Control')) xtitle('Unadjusted Ideation Score') ytitle('Density')
```

```
graph export output/ideation_kdensity_treat_uncond.png, replace

graph twoway ///

	(kdensity ideation_analysis_resid if treat==1) ///

	(kdensity ideation_analysis_resid if treat==0) ///

	, scheme(plotplain) legend(order(1 ‘Treatment’ 2 ‘Control’)) xtitle(‘Adjusted Ideation Score’) ytitle(‘Density’)

graph export output/ideation_kdensity_treat_cond.png, replace


* Effect size calculations - primary analysis

count if primarysample==1 & treat==0

local wam_controln = r(N)

count if primarysample==1 & treat==1

local wam_treatn = r(N)

quietly summarize wam_score_dw if primarysample==1 & treat==0

local wam_controlsd : di %7.2fc r(sd)

quietly summarize wam_score_dw if primarysample==1 & treat==1

local wam_treatsd : di %7.2fc r(sd)

local wam_sdpooled = sqrt(((`wam_controln'-1)*(`wam_controlsd'^2) + (`wam_treatn'-1)*(`wam_treatsd'^2))/(`wam_controln'+`wam_treatn'-2))

local wam_sdpooled : di %7.2fc `wam_sdpooled'


local wam_j = (exp(lngamma((`wam_treatn'+`wam_controln'+2)/2))) / (((`wam_treatn'+`wam_controln'+2)/2)^.5 * exp(lngamma((`wam_treatn'+`wam_controln'+2-1)/2)))

if `wam_j' == . {

	local wam_j = 1-(3/((4*(`wam_treatn'+`wam_controln'))-9))

	}

local wam_g : di %7.2fc `wam_j'*(`wam_treatdiff'/`wam_sdpooled')

local wam_g_upperci : di %7.2fc `wam_j'*(`wam_treatdiff_upperci'/`wam_sdpooled')

local wam_g_lowerci : di %7.2fc `wam_j'*(`wam_treatdiff_lowerci'/`wam_sdpooled')

local wam_j : di %7.5fc `wam_j'


* Effect size calculations - secondary analysis - WSEM
```

```
count if wsem16_score<. & secondarysample==1 & treat==0

local wsem_controln = r(N)

count if wsem16_score<. & secondarysample==1 & treat==1

local wsem_treatn = r(N)

quietly summarize wsem16_score if secondarysample==1 & treat==0

local wsem_controlsd : di %7.2fc r(sd)

quietly summarize wsem16_score if secondarysample==1 & treat==1

local wsem_treatsd : di %7.2fc r(sd)

local wsem_sdpooled = sqrt(((`wsem_controln'-1)*(`wsem_controlsd'^2) + (`wsem_treatn'-1)*(`wsem_treatsd'^2))/(`wsem_controln'+`wsem_treatn'-2))

local wsem_sdpooled : di %7.2fc `wsem_sdpooled'


local wsem_j = (exp(lngamma((`wsem_treatn'+`wsem_controln'+2)/2))) / (((`wsem_treatn'+`wsem_controln'+2)/2)^.5 * exp(lngamma((`wsem_treatn'+`wsem_controln'+2-1)/2)))

if `wsem_j' == . {

        local wsem_j = 1-(3/((4*(`wsem_treatn'+`wsem_controln'))-9))

        }

local wsem_g : di %7.2fc `wsem_j'*(`wsem_treatdiff'/`wsem_sdpooled')

local wsem_g_upperci : di %7.2fc `wsem_j'*(`wsem_treatdiff_upperci'/`wsem_sdpooled')

local wsem_g_lowerci : di %7.2fc `wsem_j'*(`wsem_treatdiff_lowerci'/`wsem_sdpooled')

local wsem_j : di %7.5fc `wsem_j'


* Effect size calculations - secondary analysis - Ideation

count if ideation_score<. & secondarysample==1 & treat==0

local idea_controln = r(N)

count if ideation_score<. & secondarysample==1 & treat==1

local idea_treatn = r(N)

quietly summarize ideation_score if secondarysample==1 & treat==0

local idea_controlsd : di %7.2fc r(sd)

quietly summarize ideation_score if secondarysample==1 & treat==1

local idea_treatsd : di %7.2fc r(sd)
```

```
local idea_sdpooled = sqrt(((`idea_controln'-1)*(`idea_controlsd'^2) + (`idea_treatn'-1)*(`idea_treatsd'^2))/(`idea_controln'+`idea_treatn'-2))

local idea_sdpooled : di %7.2fc `idea_sdpooled'


local idea_j = (exp(lngamma((`idea_treatn'+`idea_controln'+2)/2))) / (((`idea_treatn'+`idea_controln'+2)/2)^.5 * exp(lngamma((`idea_treatn'+`idea_controln'+2-1)/2)))

if `idea_j' == . {

	local idea_j = 1-(3/((4*(`idea_treatn'+`idea_controln'))-9))

	}

local idea_g : di %7.2fc `idea_j'*(`idea_treatdiff'/`idea_sdpooled')

local idea_g_upperci : di %7.2fc `idea_j'*(`idea_treatdiff_upperci'/`idea_sdpooled')

local idea_g_lowerci : di %7.2fc `idea_j'*(`idea_treatdiff_lowerci'/`idea_sdpooled')

local idea_j : di %7.5fc `idea_j'


* Complier analysis

corr treat comply

local complier_r : di %7.2fc r(rho)

regress comply FSM EAL FSMgroup EALgroup group#block treat if primarysample==1, vce(cluster anonschoolid)

local complier_F : di %7.0f e(F)

local complier_df_m = e(df_m)

local complier_df_r = e(df_r)

ivregress 2sls wam_score_dw FSM EAL FSMgroup EALgroup group#block (i.comply = treat), vce(cluster anonschoolid)

local complier_treatdiff : di %7.2fc _b[1.comply]

local complier_treatdiff_upperci : di %7.2fc _b[1.comply] + 1.96*_se[1.comply]

local complier_treatdiff_lowerci : di %7.2fc _b[1.comply] - 1.96*_se[1.comply]

local complier_treatdiff_p : di %7.2fc normal(-abs(_b[1.comply]/_se[1.comply]))*2

margins comply, post

local complier_controlmean : di %7.2fc _b[0.comply]

local complier_controlmean_upperci : di %7.2fc _b[0.comply] + 1.96*_se[0.comply]

local complier_controlmean_lowerci : di %7.2fc _b[0.comply] - 1.96*_se[0.comply]

local complier_treatmean : di %7.2fc _b[1.comply]
```

```
local complier_treatmean_upperci : di %7.2fc _b[1.comply] + 1.96*_se[1.comply]

local complier_treatmean_lowerci : di %7.2fc _b[1.comply] - 1.96*_se[1.comply]


* Effect size calculations - complier analysis

count if primarysample==1 & comply<. & treat==0

local complier_controln = r(N)

count if primarysample==1 & comply<. & treat==1

local complier_treatn = r(N)

count if primarysample==1 & comply<.

local complier_n = r(N)

quietly summarize wam_score_dw if primarysample==1 & comply<. & treat==0

local complier_controlsd : di %7.2fc r(sd)

quietly summarize wam_score_dw if primarysample==1 & comply<. & treat==1

local complier_treatsd : di %7.2fc r(sd)

local complier_sdpooled = sqrt(((`complier_controln'-1)*(`complier_controlsd'^2) + (`complier_treatn'-1)*(`complier_treatsd'^2))/(`complier_controln'+`complier_treatn'-2))

local complier_sdpooled : di %7.2fc `complier_sdpooled'


local complier_j = (exp(lngamma((`complier_treatn'+`complier_controln'+2)/2))) / (((`complier_treatn'+`complier_controln'+2)/2)^.5 * exp(lngamma((`complier_treatn'+`complier_controln'+2-1)/2)))

if `complier_j' == . {

        local complier_j = 1-(3/((4*(`complier_treatn'+`complier_controln'))-9))

        }

local complier_g : di %7.2fc `complier_j'*(`complier_treatdiff'/`complier_sdpooled')

local complier_g_upperci : di %7.2fc `complier_j'*(`complier_treatdiff_upperci'/`complier_sdpooled')

local complier_g_lowerci : di %7.2fc `complier_j'*(`complier_treatdiff_lowerci'/`complier_sdpooled')


* FSM check for interaction then run for EVERFSM_6_P sub-group - primary analysis

regress wam_score_dw i.treat i.FSM treat#FSM EAL FSMgroup EALgroup group#block, vce(cluster anonschoolid)

local fsmwam_interaction_p : di %7.2fc ttail(e(df_r), abs(_b[1.treat#1.FSM]/_se[1.treat#1.FSM]))*2
```

```
regress wam_score_dw i.treat EAL FSMgroup EALgroup group#block if FSM==1, vce(cluster anonschoolid)

local fsmwam_treatdiff : di %7.2fc _b[1.treat]
local fsmwam_treatdiff_upperci : di %7.2fc _b[1.treat] + 1.96*_se[1.treat]
local fsmwam_treatdiff_lowerci : di %7.2fc _b[1.treat] - 1.96*_se[1.treat]
local fsmwam_treatdiff_p : di %7.2fc ttail(e(df_r), abs(_b[1.treat]/_se[1.treat]))*2
margins treat, post
local fsmwam_controlmean : di %7.2fc _b[0.treat]
local fsmwam_controlmean_upperci : di %7.2fc _b[0.treat] + 1.96*_se[0.treat]
local fsmwam_controlmean_lowerci : di %7.2fc _b[0.treat] - 1.96*_se[0.treat]
local fsmwam_treatmean : di %7.2fc _b[1.treat]
local fsmwam_treatmean_upperci : di %7.2fc _b[1.treat] + 1.96*_se[1.treat]
local fsmwam_treatmean_lowerci : di %7.2fc _b[1.treat] - 1.96*_se[1.treat]

* FSM check for interaction then run for EVERFSM_6_P sub-group - secondary analysis (WSEM)
regress wsem16_score i.treat i.FSM treat#FSM EAL FSMgroup EALgroup group#block, vce(cluster anonschoolid)
local fsmwsem_interaction_p : di %7.2fc ttail(e(df_r), abs(_b[1.treat#1.FSM]/_se[1.treat#1.FSM]))*2

regress wam_score_dw i.treat EAL FSMgroup EALgroup group#block if FSM==1, vce(cluster anonschoolid)

local fsmwsem_treatdiff : di %7.2fc _b[1.treat]
local fsmwsem_treatdiff_upperci : di %7.2fc _b[1.treat] + 1.96*_se[1.treat]
local fsmwsem_treatdiff_lowerci : di %7.2fc _b[1.treat] - 1.96*_se[1.treat]
local fsmwsem_treatdiff_p : di %7.2fc ttail(e(df_r), abs(_b[1.treat]/_se[1.treat]))*2
margins treat, post
local fsmwsem_controlmean : di %7.2fc _b[0.treat]
local fsmwsem_controlmean_upperci : di %7.2fc _b[0.treat] + 1.96*_se[0.treat]
local fsmwsem_controlmean_lowerci : di %7.2fc _b[0.treat] - 1.96*_se[0.treat]
local fsmwsem_treatmean : di %7.2fc _b[1.treat]
local fsmwsem_treatmean_upperci : di %7.2fc _b[1.treat] + 1.96*_se[1.treat]
local fsmwsem_treatmean_lowerci : di %7.2fc _b[1.treat] - 1.96*_se[1.treat]
```

```
* FSM check for interaction then run for EVERFSM_6_P sub-group - secondary analysis (Ideation)
regress ideation_score i.treat i.FSM treat#FSM EAL FSMgroup EALgroup group#block, vce(cluster anonschoolid)
local fsmidea_interaction_p : di %7.2fc ttail(e(df_r), abs(_b[1.treat#1.FSM]/_se[1.treat#1.FSM]))*2

regress ideation_score i.treat EAL FSMgroup EALgroup group#block if FSM==1, vce(cluster anonschoolid)

local fsmidea_treatdiff : di %7.2fc _b[1.treat]
local fsmidea_treatdiff_upperci : di %7.2fc _b[1.treat] + 1.96*_se[1.treat]
local fsmidea_treatdiff_lowerci : di %7.2fc _b[1.treat] - 1.96*_se[1.treat]
local fsmidea_treatdiff_p : di %7.2fc ttail(e(df_r), abs(_b[1.treat]/_se[1.treat]))*2
margins treat, post
local fsmidea_controlmean : di %7.2fc _b[0.treat]
local fsmidea_controlmean_upperci : di %7.2fc _b[0.treat] + 1.96*_se[0.treat]
local fsmidea_controlmean_lowerci : di %7.2fc _b[0.treat] - 1.96*_se[0.treat]
local fsmidea_treatmean : di %7.2fc _b[1.treat]
local fsmidea_treatmean_upperci : di %7.2fc _b[1.treat] + 1.96*_se[1.treat]
local fsmidea_treatmean_lowerci : di %7.2fc _b[1.treat] - 1.96*_se[1.treat]

* Effect size calculations - FSM sub-group analysis
count if primarysample==1 & FSM==1 & treat==0
local primarysample_fsm_control_n = r(N)
count if primarysample==1 & FSM==1 & treat==1
local primarysample_fsm_treat_n = r(N)
count if primarysample==1 & FSM==1
local primarysample_fsm_n = r(N)

count if secondarysample==1 & FSM==1 & treat==0
local secondarysample_fsm_control_n = r(N)
count if secondarysample==1 & FSM==1 & treat==1
local secondarysample_fsm_treat_n = r(N)
```

```
count if secondarysample==1 & FSM==1

local secondarysample_fsm_n = r(N)


* Primary

quietly summarize wam_score_dw if primarysample==1 & FSM==1 & treat==0

local fsmwam_controlsd : di %7.2fc r(sd)

quietly summarize wam_score_dw if primarysample==1 & FSM==1 & treat==1

local fsmwam_treatsd : di %7.2fc r(sd)

local fsmwam_sdpooled = sqrt(((`primarysample_fsm_control_n'-1)*(`fsmwam_controlsd'^2) + (`primarysample_fsm_treat_n'-
1)*(`fsmwam_treatsd'^2))/(`primarysample_fsm_control_n'+`primarysample_fsm_treat_n'-2))

local fsmwam_sdpooled : di %7.2fc `fsmwam_sdpooled'


local fsmwam_j = (exp(lngamma((`primarysample_fsm_treat_n'+`primarysample_fsm_control_n'+2)/2))) /
(((`primarysample_fsm_treat_n'+`primarysample_fsm_control_n'+2)/2)^.5 *
exp(lngamma((`primarysample_fsm_treat_n'+`primarysample_fsm_control_n'+2-1)/2)))

if `fsmwam_j' == . {

        local fsmwam_j = 1-(3/((4*(`primarysample_fsm_treat_n'+`primarysample_fsm_control_n'))-9))

        }

local fsmwam_g : di %7.2fc `fsmwam_j'*(`fsmwam_treatdiff'/`fsmwam_sdpooled')

local fsmwam_g_upperci : di %7.2fc `fsmwam_j'*(`fsmwam_treatdiff_upperci'/`fsmwam_sdpooled')

local fsmwam_g_lowerci : di %7.2fc `fsmwam_j'*(`fsmwam_treatdiff_lowerci'/`fsmwam_sdpooled')


* Secondary (WSEM)

quietly summarize wsem16_score if secondarysample==1 & FSM==1 & treat==0

local fsmwsem_controlsd : di %7.2fc r(sd)

quietly summarize wsem16_score if secondarysample==1 & FSM==1 & treat==1

local fsmwsem_treatsd : di %7.2fc r(sd)

local fsmwsem_sdpooled = sqrt(((`secondarysample_fsm_control_n'-1)*(`fsmwsem_controlsd'^2) + (`secondarysample_fsm_treat_n'-
1)*(`fsmwsem_treatsd'^2))/(`secondarysample_fsm_control_n'+`secondarysample_fsm_treat_n'-2))

local fsmwsem_sdpooled : di %7.2fc `fsmwsem_sdpooled'
```

```
local fsmwsem_j = (exp(lngamma((`secondarysample_fsm_treat_n'+`secondarysample_fsm_control_n'+2)/2))) / (((`secondarysample_fsm_treat_n'+`secondarysample_fsm_control_n'+2)/2)^.5 * exp(lngamma((`secondarysample_fsm_treat_n'+`secondarysample_fsm_control_n'+2-1)/2)))

if `fsmwsem_j' == . {

        local fsmwsem_j = 1-(3/((4*(`secondarysample_fsm_treat_n'+`secondarysample_fsm_control_n'))-9))

        }

local fsmwsem_g : di %7.2fc `fsmwsem_j'*(`fsmwsem_treatdiff'/`fsmwsem_sdpooled')

local fsmwsem_g_upperci : di %7.2fc `fsmwsem_j'*(`fsmwsem_treatdiff_upperci'/`fsmwsem_sdpooled')

local fsmwsem_g_lowerci : di %7.2fc `fsmwsem_j'*(`fsmwsem_treatdiff_lowerci'/`fsmwsem_sdpooled')


* Secondary (Ideation)

quietly summarize ideation_score if secondarysample==1 & FSM==1 & treat==0

local fsmidea_controlsd : di %7.2fc r(sd)

quietly summarize ideation_score if secondarysample==1 & FSM==1 & treat==1

local fsmidea_treatsd : di %7.2fc r(sd)

local fsmidea_sdpooled = sqrt(((`secondarysample_fsm_control_n'-1)*(`fsmidea_controlsd'^2) + (`secondarysample_fsm_treat_n'-1)*(`fsmidea_treatsd'^2))/(`secondarysample_fsm_control_n'+`secondarysample_fsm_treat_n'-2))

local fsmidea_sdpooled : di %7.2fc `fsmidea_sdpooled'


local fsmidea_j = (exp(lngamma((`secondarysample_fsm_treat_n'+`secondarysample_fsm_control_n'+2)/2))) / (((`secondarysample_fsm_treat_n'+`secondarysample_fsm_control_n'+2)/2)^.5 * exp(lngamma((`secondarysample_fsm_treat_n'+`secondarysample_fsm_control_n'+2-1)/2)))

if `fsmidea_j' == . {

        local fsmidea_j = 1-(3/((4*(`secondarysample_fsm_treat_n'+`secondarysample_fsm_control_n'))-9))

        }

local fsmidea_g : di %7.2fc `fsmidea_j'*(`fsmidea_treatdiff'/`fsmidea_sdpooled')

local fsmidea_g_upperci : di %7.2fc `fsmidea_j'*(`fsmidea_treatdiff_upperci'/`fsmidea_sdpooled')

local fsmidea_g_lowerci : di %7.2fc `fsmidea_j'*(`fsmidea_treatdiff_lowerci'/`fsmidea_sdpooled')


* Missing data analysis - create missing indicator variables

cap drop wam_score_dw_miss

gen wam_score_dw_miss = 0
```

```
replace wam_score_dw_miss = 1 if wam_score_dw>=.

logit wam_score_dw_miss i.treat FSM FSMgroup EAL EALgroup group#block, vce(cluster anonschoolid)

cap drop wsem16_score_miss

gen wsem16_score_miss = 0

replace wsem16_score_miss = 1 if wsem16_score>=.

logit wsem16_score_miss i.treat FSM FSMgroup EAL EALgroup group#block, vce(cluster anonschoolid)


* Missing data analysis - only outcome variable missing

regress wam_score_dw i.treat FSM FSMgroup EAL EALgroup group#block, vce(cluster anonschoolid)

* [IF THERE WERE ANY OTHER STATISTICALLY SIGNIFICANT PREDICTORS OF wam_score_dw_miss NOT ALREADY HERE WE'D ADD THEM AND REPORT THIS AS A ROBUSTNESS CHECK MODEL]


* Missing data analysis - predictor variables

* [NO MISSING PREDICTOR VARIABLE DATA IN ANALYSIS SAMPLE]


* CREATE TABLES


* Balance table - Randomisation sample

cap file close baltab

file open baltab using output/balance_rsample.csv, write replace

file write baltab ‘, , Control, , Intervention’ _n

file write baltab ‘Pupil-level (categorical), National-level mean, n/N (missing), Count (%), n/N (missing), Count (%), Standardised difference’ _n

file write baltab ‘Ever FSM, ${natmean_fsm}, `fsm_controln_rsample' / `asrandom_schoolcount_control' (`fsm_controlmiss_rsample'), `fsm_controln_in_rsample' (`fsm_controlpercent_in_rsample'), ‘

file write baltab ‘`fsm_treatn_rsample' / `asrandom_schoolcount_treat' (`fsm_treatmiss_rsample'), `fsm_treatn_in_rsample' (`fsm_treatpercent_in_rsample'), `fsm_in_stddiff_rsample'‘ _n

file write baltab ‘Non-Ever FSM, ${natmean_nonfsm}, `fsm_controln_rsample' / `asrandom_schoolcount_control' (`fsm_controlmiss_rsample'), `fsm_controln_out_rsample' (`fsm_controlpercent_out_rsample'), ‘

file write baltab ‘`fsm_treatn_rsample' / `asrandom_schoolcount_treat' (`fsm_treatmiss_rsample'), `fsm_treatn_out_rsample' (`fsm_treatpercent_out_rsample'), `fsm_out_stddiff_rsample'‘ _n

file write baltab ‘EAL, ${natmean_eal}, `eal_controln_rsample' / `asrandom_schoolcount_control' (`eal_controlmiss_rsample'), `eal_controln_in_rsample' (`eal_controlpercent_in_rsample'), ‘
```

```
file write baltab ‘`eal_treatn_rsample' / `asrandom_schoolcount_treat' (`eal_treatmiss_rsample'), `eal_treatn_in_rsample' (`eal_treatpercent_in_rsample'), `eal_in_stddiff_rsample'‘ _n

file write baltab ‘Non-EAL, ${natmean_noneal}, `eal_controln_rsample' / `asrandom_schoolcount_control' (`eal_controlmiss_rsample'), `eal_controln_out_rsample' (`eal_controlpercent_out_rsample'), ‘

file write baltab ‘`eal_treatn_rsample' / `asrandom_schoolcount_treat' (`eal_treatmiss_rsample'), `eal_treatn_out_rsample' (`eal_treatpercent_out_rsample'), `eal_out_stddiff_rsample'‘ _n

file close baltab


* Balance table - Primary analysis sample

cap file close baltab

file open baltab using output/balance_psample.csv, write replace

file write baltab ‘, , Control, , Intervention’ _n

file write baltab ‘Pupil-level (categorical), National-level mean, n/N (missing), Count (%), n/N (missing), Count (%), Standardised difference’ _n

file write baltab ‘Ever FSM, ${natmean_fsm}, `fsm_controln_psample' / `asrandom_schoolcount_control' (`fsm_controlmiss_psample'), `fsm_controln_in_psample' (`fsm_controlpercent_in_psample'), ‘

file write baltab ‘`fsm_treatn_psample' / `asrandom_schoolcount_treat' (`fsm_treatmiss_psample'), `fsm_treatn_in_psample' (`fsm_treatpercent_in_psample'), `fsm_in_stddiff_psample'‘ _n

file write baltab ‘Non-Ever FSM, ${natmean_nonfsm}, `fsm_controln_psample' / `asrandom_schoolcount_control' (`fsm_controlmiss_psample'), `fsm_controln_out_psample' (`fsm_controlpercent_out_psample'), ‘

file write baltab ‘`fsm_treatn_psample' / `asrandom_schoolcount_treat' (`fsm_treatmiss_psample'), `fsm_treatn_out_psample' (`fsm_treatpercent_out_psample'), `fsm_out_stddiff_psample'‘ _n

file write baltab ‘EAL, ${natmean_eal}, `eal_controln_psample' / `asrandom_schoolcount_control' (`eal_controlmiss_psample'), `eal_controln_in_psample' (`eal_controlpercent_in_psample'), ‘

file write baltab ‘`eal_treatn_psample' / `asrandom_schoolcount_treat' (`eal_treatmiss_psample'), `eal_treatn_in_psample' (`eal_treatpercent_in_psample'), `eal_in_stddiff_psample'‘ _n

file write baltab ‘Non-EAL, ${natmean_noneal}, `eal_controln_psample' / `asrandom_schoolcount_control' (`eal_controlmiss_psample'), `eal_controln_out_psample' (`eal_controlpercent_out_psample'), ‘

file write baltab ‘`eal_treatn_psample' / `asrandom_schoolcount_treat' (`eal_treatmiss_psample'), `eal_treatn_out_psample' (`eal_treatpercent_out_psample'), `eal_out_stddiff_psample'‘ _n

file close baltab


* Outcome measure descriptive statistics

cap file close outcomes

file open outcomes using output/outcomes.csv, write replace

file write outcomes ‘Outcome, Mean, SD, ICC, N’ _n
```

```
file write outcomes ‘WAM Score (ideas scale double weighted), `wam_mean', `wam_sd', `wam_icc', `wam_n'‘ _n

file write outcomes ‘WSEM-16 Score, `wsem_mean', `wsem_sd', `wsem_icc', `wsem_n'‘ _n

file write outcomes ‘Ideation Score, `idea_mean', `idea_sd', `idea_icc', `idea_n'‘ _n

file close outcomes


* Primary analysis

cap file close primary

file open primary using output/primary.csv, write replace

file write primary ‘Unadjusted means, Full sample , Control group, , Intervention group, , Effect size calculation, , ‘ _n

file write primary ‘Outcome, n (missing), Mean (95% CI), n (missing), Mean (95% CI), n (missing), Total n (intervention; control), Hedges' g (95% CI), p-value’ _n

file write primary ‘WAM Score (ideas scale double weighted), `primarymeasure_n' (`primarymeasure_miss'), ‘

file write primary ‘`wam_controlmean_unadj' (`wam_controlmean_unadj_lowerci'; `wam_controlmean_unadj_upperci'), `primarymeasure_control_n' (`primarymeasure_control_miss'), ‘

file write primary ‘`wam_treatmean_unadj' (`wam_treatmean_unadj_lowerci'; `wam_treatmean_unadj_upperci'), `primarymeasure_treat_n' (`primarymeasure_treat_miss'), ‘

file write primary ‘`primarysample_n' (`primarysample_treat_n'; `primarysample_control_n'), `wam_g' (`wam_g_lowerci'; `wam_g_upperci'), `wam_treatdiff_p'‘ _n

file close primary


* Secondary analysis

cap file close secondary

file open secondary using output/secondary.csv, write replace

file write secondary ‘Unadjusted means, Full sample , Control group, , Intervention group, , Effect size, , ‘ _n

file write secondary ‘Outcome, n (missing), Mean (95% CI), n (missing), Mean (95% CI), n (missing), Total n (intervention; control), Hedges' g (95% CI), p-value’ _n

file write secondary ‘WSEM Score, `wsem_n' (`wsem_miss'), ‘

file write secondary ‘`wsem_controlmean_unadj' (`wsem_controlmean_unadj_lowerci'; `wsem_controlmean_unadj_upperci'), `wsem_control_n' (`wsem_control_miss'), ‘

file write secondary ‘`wsem_treatmean_unadj' (`wsem_treatmean_unadj_lowerci'; `wsem_treatmean_unadj_upperci'), `wsem_treat_n' (`wsem_treat_miss'), ‘

file write secondary ‘`secondarysample_n' (`secondarysample_treat_n'; `secondarysample_control_n'), `wsem_g' (`wsem_g_lowerci'; `wsem_g_upperci'), `wsem_treatdiff_p'‘ _n

file write secondary ‘Ideation Score, `idea_n' (`idea_miss'), ‘
```

file write secondary ‘`idea_controlmean_unadj' (`idea_controlmean_unadj_lowerci'; `idea_controlmean_unadj_upperci'), `idea_control_n' (`idea_control_miss'), ‘

file write secondary ‘`idea_treatmean_unadj' (`idea_treatmean_unadj_lowerci'; `idea_treatmean_unadj_upperci'), `idea_treat_n' (`idea_treat_miss'), ‘

file write secondary ‘`secondarysample_n' (`secondarysample_treat_n'; `secondarysample_control_n'), `idea_g' (`idea_g_lowerci'; `idea_g_upperci'), `idea_treatdiff_p'‘ _n

file close secondary

* Sub-group analyses

cap file close subgroup

file open subgroup using output/subgroup.csv, write replace

file write subgroup ‘Model, Hedges' g (95% CI), N, *p*-value of interaction term, *p*-value of treatment variable ‘ _n

file write subgroup ‘WAM FSM sub-group analysis, `fsmwam_g' (`fsmwam_g_lowerci'; `fsmwam_g_upperci'), `primarysample_fsm_n', `fsmwam_interaction_p', `fsmwam_treatdiff_p'‘ _n

file write subgroup ‘WSEM FSM sub-group analysis, `fsmwsem_g' (`fsmwsem_g_lowerci'; `fsmwsem_g_upperci'), `secondarysample_fsm_n', `fsmwsem_interaction_p', `fsmwsem_treatdiff_p'‘ _n

file write subgroup ‘Ideation FSM sub-group analysis, `fsmidea_g' (`fsmidea_g_lowerci'; `fsmidea_g_upperci'), `secondarysample_fsm_n', `fsmidea_interaction_p', `fsmidea_treatdiff_p'‘ _n

file close subgroup

* Compliance analysis

cap file close subgroup

file open subgroup using output/compliance.csv, write replace

file write subgroup ‘Model, Hedges' g (95% CI), N, First stage F test, Compliance/treatment correlation, *p*-value of treatment variable ‘ _n

file write subgroup `’Compliance analysis, `complier_g' ‘(`complier_g_lowerci', `complier_g_upperci')’, `complier_n', ‘F(`complier_df_m', `complier_df_r') = `complier_F'‘, `complier_r', `complier_treatdiff_p'‘' _n

file close subgroup

* Attrition table

cap file close attrition

file open attrition using output/attrition.csv, write replace

file write attrition ‘ , , Intervention, Control, Total’ _n

```
file write attrition ‘Number of pupils, Randomised, `randomisedsample_treat_n', `randomisedsample_control_n', `randomisedsample_n'‘ _n

file write attrition ‘ , Analysed, `primarysample_treat_n', `primarysample_control_n', `primarysample_n'‘ _n

file write attrition ‘Pupil attrition, Number, `primaryattrition_treat_n', `primaryattrition_control_n', `primaryattrition_n''‘ _n

file write attrition ‘(from randomisation to analysis) , Percentage, `primaryattrition_treat_p', `primaryattrition_control_p', `primaryattrition_percent''‘ _n

file close attrition


* Effect size calculation appendix table

cap file close esizecalc

file open esizecalc using output/esizecalc.csv, write replace

file write esizecalc ‘ , , , Control group, , Intervention group’ _n

file write esizecalc ‘Outcome, Unadjusted difference in means, Adjusted difference in means, n (missing), Outcome SD, n (missing), Outcome SD, Pooled SD, Hedges' correction (J)’ _n

file write esizecalc ‘WAM Score (ideas double weighted), `wam_treatdiff_unadj', `wam_treatdiff', ‘

file write esizecalc ‘`primarymeasure_control_n' (`primarymeasure_control_miss'), `wam_controlsd', ‘

file write esizecalc ‘`primarymeasure_treat_n' (`primarymeasure_treat_miss'), `wam_treatsd', ‘

file write esizecalc ‘`wam_sdpooled', `wam_j'‘ _n

file write esizecalc ‘WSEM Score, `wsem_treatdiff_unadj', `wsem_treatdiff', ‘

file write esizecalc ‘`wsem_control_n' (`wsem_control_miss'), `wsem_controlsd', ‘

file write esizecalc ‘`wsem_treat_n' (`wsem_treat_miss'), `wsem_treatsd', ‘

file write esizecalc ‘`wsem_sdpooled', `wsem_j'‘ _n

file write esizecalc ‘Ideation Score, `idea_treatdiff_unadj', `idea_treatdiff', ‘

file write esizecalc ‘`idea_control_n' (`idea_control_miss'), `idea_controlsd', ‘

file write esizecalc ‘`idea_treat_n' (`idea_treat_miss'), `idea_treatsd', ‘

file write esizecalc ‘`idea_sdpooled', `idea_j'‘ _n

file close esizecalc


* Power Calculation

powercalc, cluster n(`wam_nhmean') g(`wam_schoolcount') icc(`wam_icc') r2(`wam_r2_pupil') r2_g(`wam_r2_school') treated(`wam_treatprop') blocks(8) regressors(2)

local wam_mdes : di %7.2fc r(mdes)
```

powercalc, cluster n(`wsem_nhmean') g(`wsem_schoolcount') icc(`wsem_icc') r2(`wsem_r2_pupil') r2_g(`wsem_r2_school') treated(`wsem_treatprop') blocks(8) regressors(2)

local wsem_mdes : di %7.2fc r(mdes)

powercalc, cluster n(`idea_nhmean') g(`idea_schoolcount') icc(`idea_icc') r2(`idea_r2_pupil') r2_g(`idea_r2_school') treated(`idea_treatprop') blocks(8) regressors(2)

local idea_mdes : di %7.2fc r(mdes)

* Power Calculation - FSM sub-group

powercalc, cluster n(`wam_fsm_nhmean') g(`wam_schoolcount') icc(`wam_fsm_icc') r2(`wam_fsm_r2_pupil') r2_g(`wam_fsm_r2_school') treated(`wam_fsm_treatprop') blocks(8) regressors(2)

local wam_fsm_mdes : di %7.2fc r(mdes)

powercalc, cluster n(`wsem_fsm_nhmean') g(`wsem_schoolcount') icc(`wsem_fsm_icc') r2(`wsem_fsm_r2_pupil') r2_g(`wsem_fsm_r2_school') treated(`wsem_fsm_treatprop') blocks(8) regressors(2)

local wsem_fsm_mdes : di %7.2fc r(mdes)

powercalc, cluster n(`idea_fsm_nhmean') g(`idea_schoolcount') icc(`idea_fsm_icc') r2(`idea_fsm_r2_pupil') r2_g(`idea_fsm_r2_school') treated(`idea_fsm_treatprop') blocks(8) regressors(2)

local idea_fsm_mdes : di %7.2fc r(mdes)

* Table with inputs for power calculation tables

cap file close powercalc

file open powercalc using output/powercalcinputs.csv, write replace

file write powercalc ‘Outcome, MDES, Pupil-level R2, School-level R2, Overall R2, ICC, Average Cluster Size, Control Schools, Intervention Schools, Total Schools, Control Pupils, Intervention Pupils, Total Pupils’ _n

file write powercalc ‘WAM Score, `wam_mdes', `wam_r2_pupil', `wam_r2_school', `wam_r2_overall', `wam_icc', `wam_nhmean',’

file write powercalc ‘`wam_schoolcount_control', `wam_schoolcount_treat', `wam_schoolcount', `primarysample_control_n', `primarysample_treat_n', `primarysample_n‘‘ _n

file write powercalc ‘WSEM Score, `wsem_mdes', `wsem_r2_pupil', `wsem_r2_school', `wsem_r2_overall', `wsem_icc', `wsem_nhmean',’

file write powercalc ‘`wsem_schoolcount_control', `wsem_schoolcount_treat', `wsem_schoolcount', `secondarysample_control_n', `secondarysample_treat_n', `secondarysample_n‘‘ _n

file write powercalc ‘Ideation Score, `idea_mdes', `idea_r2_pupil', `idea_r2_school', `idea_r2_overall', `idea_icc', `idea_nhmean', ‘

file write powercalc ‘`idea_schoolcount_control', `idea_schoolcount_treat', `idea_schoolcount', `secondarysample_control_n', `secondarysample_treat_n', `secondarysample_n‘‘ _n

```
file write powercalc ‘WAM Score (FSM), `wam_fsm_mdes', `wam_fsm_r2_pupil', `wam_fsm_r2_school', `wam_fsm_r2_overall', `wam_fsm_icc', `wam_fsm_nhmean',’

        file write powercalc ‘`wam_schoolcount_control', `wam_schoolcount_treat', `wam_schoolcount', `primarysample_fsm_control_n', `primarysample_fsm_treat_n', `primarysample_fsm_n'‘ _n

file close powercalc


* Robustness check including additional controls (time to test, marker fixed effects, KS1)

regress wam_score_dw i.treat FSM EAL FSMgroup EALgroup group#block if primarysample==1 & TestingDate<., vce(cluster anonschoolid)

estimates store notestdate

estadd scalar es = `wam_j'*(_b[1.treat]/`wam_sdpooled')

regress wam_score_dw i.treat FSM EAL FSMgroup EALgroup group#block TestingDate if primarysample==1, vce(cluster anonschoolid)

estimates store testdate

estadd scalar es = `wam_j'*(_b[1.treat]/`wam_sdpooled')


recode wam_marker_num .=0

regress wam_score_dw i.treat FSM EAL FSMgroup EALgroup group#block if primarysample==1 & wam_marker_num<., vce(cluster anonschoolid)

estimates store nomarker

estadd scalar es = `wam_j'*(_b[1.treat]/`wam_sdpooled')

regress wam_score_dw i.treat FSM EAL FSMgroup EALgroup group#block i.wam_marker_num if primarysample==1, vce(cluster anonschoolid)

estimates store marker

estadd scalar es = `wam_j'*(_b[1.treat]/`wam_sdpooled')

recode wam_marker_num 0=.


mi set flong

mi register imputed tks1ave_round

mi impute regress tks1ave_round FSM EAL FSMgroup EALgroup group#block if primarysample==1, add(20) rseed(11849)

mi estimate, post: regress wam_score_dw i.treat FSM EAL FSMgroup EALgroup group#block if primarysample==1 & tks1ave_round<., vce(cluster anonschoolid)

estimates store noks1
```

```
estadd scalar es = `wam_j'*(_b[1.treat]/`wam_sdpooled')

mi estimate, post: regress wam_score_dw i.treat FSM EAL FSMgroup EALgroup group#block tks1ave_round if primarysample==1, vce(cluster anonschoolid)

estimates store ks1

estadd scalar es = `wam_j'*(_b[1.treat]/`wam_sdpooled')

mi extract 0, clear


estimates restore primary

estimates store primary

estadd scalar es = `wam_g'


estout primary notestdate testdate nomarker marker noks1 ks1 using output/robustness.csv, replace ///

        mlabels('Primary' 'No Test Date' 'Test Date' 'No Marker FE' 'Marker FE' 'No KS1' 'KS1', nonumbers nodepvars) ///

        keep(1.treat) varlabels(1.treat 'Treatment') ///

        indicate('Test Date = TestingDate' 'Marker FE = 4.wam_marker_num' 'KS1 = tks1ave_round') ///

        cells(b(fmt(3) star) p(fmt(3) par)) delimiter(', ') ///

        stats(es N, fmt(2 0) labels('Effect Size' 'N'))


* Randomisation inference as a robustness check

cap drop groupblock

egen groupblock = concat(group block)

encode groupblock, gen(groupblock_enc)

drop groupblock

rename groupblock_enc groupblock

ritest treat _b[1.treat], r(2000) strata(groupblock) cluster(anonschoolid) seed(987234) : ///

   regress wam_score_dw i.treat FSM EAL FSMgroup EALgroup group#block, vce(cluster anonschoolid)

local wam_ri_p : di %7.2fc r(p)[1,1]

ritest treat _b[1.treat], r(2000) strata(groupblock) cluster(anonschoolid) seed(987234) : ///

   regress wsem16_score i.treat FSM EAL FSMgroup EALgroup group#block, vce(cluster anonschoolid)

local wsem_ri_p : di %7.2fc r(p)[1,1]
```

```
ritest treat _b[1.treat], r(2000) strata(groupblock) cluster(anonschoolid) seed(987234) : ///
    regress ideation_score i.treat FSM EAL FSMgroup EALgroup group#block, vce(cluster anonschoolid)
local idea_ri_p : di %7.2fc r(p)[1,1]

* Table to compare p values from randomisation inference and sampling inference
cap file close ricomp
file open ricomp using output/ricomp.csv, write replace
file write ricomp ‘Outcome, Effect size, Sampling Inference p-value, Randomisation Inference p-value’ _n
file write ricomp ‘WAM Score (ideas scale double weighted), `wam_g', `wam_treatdiff_p', `wam_ri_p'‘ _n
file write ricomp ‘WSEM Score, `wsem_g', `wsem_treatdiff_p', `wsem_ri_p'‘ _n
file write ricomp ‘Ideation Score, `idea_g', `idea_treatdiff_p', `idea_ri_p'‘ _n
file close ricomp

* Table with number of treatment and control schools in each block
cap file close strata
file open strata using output/strata.csv, write replace
file write strata ‘Randomisation Batch, Stratum, Treatment (N), Control (N)’ _n
file write strata ‘1, Low EAL/Low FSM, `block11treat_schoolcount_1', `block11control_schoolcount_1'‘ _n
file write strata ‘1, Low EAL/High FSM, `block12treat_schoolcount_1', `block12control_schoolcount_1'‘ _n
file write strata ‘1, High EAL/Low FSM, `block21treat_schoolcount_1', `block21control_schoolcount_1'‘ _n
file write strata ‘1, High EAL/High FSM, `block22treat_schoolcount_1', `block22control_schoolcount_1'‘ _n
file write strata ‘2, Low EAL/Low FSM, `block11treat_schoolcount_2', `block11control_schoolcount_2'‘ _n
file write strata ‘2, Low EAL/High FSM, `block12treat_schoolcount_2', `block12control_schoolcount_2'‘ _n
file write strata ‘2, High EAL/Low FSM, `block21treat_schoolcount_2', `block21control_schoolcount_2'‘ _n
file write strata ‘2, High EAL/High FSM, `block22treat_schoolcount_2', `block22control_schoolcount_2'‘ _n
file close strata
```

Appendix G: Teacher interview guide

Craft of Writing: Interviews with Classroom Teacher

The interviews should last around 30 minutes. The timings given for each section are a guide - you may spend longer or shorter on each section. Lead questions are presented in bold, with potential follow-up questions presented in a non-bold typeface. As the interviews are semi-structured, not all questions need to be asked and they do not need to be asked in order. The interviewer should be responsive to what the interviewee, following the direction of the conversation and following-up with additional questions as needed.

Main objective	Purpose of section	Guide timings
1. Introduction	Explains the purpose and 'ground rules' of the interview.	3 mins
2. Background context	Allows the participant to settle into the interview, as well as providing some background to the school, so that we understand more about the school context that the programme is intended to benefit.	5 mins
3. Experience of Craft of Writing sessions	This section will focus on understanding the perceived quality of the intervention, as well as experiences of the programme's delivery, including barriers and facilitators to delivery.	5 mins
4. Programme mechanisms	To explore whether the hypothesised intermediate mechanisms from the programme's theory of change are taking place, particularly focusing on whether the teacher's have engaged in writing outside the training sessions and whether they have used the techniques from the training in the classroom.	5 mins
5. Programme impact	To understand the longer-term changes of the programme, both positive and negative, for the teacher, pupils and the wider school.	5 mins
6. Close	Thank you and close	2 mins
7. Reflection following observation of a lesson involving writing	Wherever possible, this will be an opportunity to discuss areas of interest arising from the lesson observation.	5 mins

1. Introduction	
Introduction: • Introduce yourself • Introduce BIT and IOE – explain that we are independently evaluating the Craft of Music programme, which is one of five programmes that are part of the Cultural Learning programme that is jointly funded by the Education Endowment Foundation and Royal Society of Arts. Aims of this interview: We are here to learn more about your involvement in the Craft of Writing programme, what has helped the programme to work, and what the challenges have been. We'd also like to understand any impact the programme has had on your school, particularly for the teacher attending the Craft of Writing sessions and the pupils in his/her class. This interview: • Should last around 30 minutes • Stress that you want to understand the intervention from their point of view. No answers are right or wrong – and we are not here to judge the decisions made or views held by the interviewee. Anonymity and privacy: • All information gathered will be in strict confidence, unless there are concerns about safeguarding. When we write up the research we will ensure that no one is identifiable from any reporting, • Explain that if at any point they feel uncomfortable or prefer not to answer a specific question they can just say so. • Explain that it is their choice whether they take part in the interview and they can end the interview at any point, without giving a reason. Recording: 1. Explain that recording enables us to have an accurate record of what was said, which can be typed up for analysis alongside other interviews. We may also use quotes from this interview, but these will be included in a way that means no individual or school is identifiable. 2. Check if they have any questions about the interview. If they are happy to go ahead, obtain verbal permission to digitally	Orientates respondent and gets them prepared to take part in the discussion. Outlines the 'rules' of the interview.

record and take notes (written permission should already have been obtained). 3. Once you have consent, start the voice recorder. 4. State interview number/participant ID	
2. Background context	5 mins
How many years have you been teaching? How long have you been working at this school? Could you tell me a little bit about your role at the school? ● What would you say are the school's main strengths and challenges? How did you become involved with the Craft of Writing programme? What were your thoughts about the Craft of Writing programme when you first heard of it? ● How did you feel about attending the Craft of Writing training sessions? How confident did you feel about your ability to write prior to beginning the Craft of Writing programme? ● How regularly did you engage in longer writing exercises? Probe for more detail about anything they did previously. ● Had you had any training on writing skills previously? If yes, ask them to describe what this involved.	Allows the participant to settle into the interview, as well as providing some background to the school, so that we understand more about the school context that the programme is intended to benefit.
3. Experience of the Craft of Writing training sessions	5 mins
How many residential weekends and CPD sessions did you attend? ● How did you find making the time to attend these sessions? Probe for detail on any challenges raised and how these were overcome. ● How did members of the Senior Leadership Team feel about your attending the sessions? Probe for detail about SLT buy-in. Can you tell me a bit about what the CPD sessions that you attended? ● What did you cover in the CPD sessions?	This section will focus on understanding the experience of the Craft of Writing training, including perceived quality of the CPD and residential weekends and the barriers and facilitators to engaging in the training sessions.

• Can you describe any writing skills or techniques that you learnt? • Can you describe any ways you used what you learnt in the classroom? Can you tell me a bit about the residential weekends that you attended? • What did you cover during the residential weekends? • What skills or techniques did you learn at the residential weekends? • Can you describe any ways you used what you learnt in the classroom? • How did the residential weekends compare to the CDP sessions that you did? Overall, what did you think of the different training sessions you did for Craft of Writing? • What did you think of the way the training was delivered? ○ To what extent was the training pitched at the right level? • How did you find the other teachers in your group? • What (else) did you like about the training? • What (else) did you find more challenging/less useful about the training sessions? • What recommendations do you have about how the programme could be improved?	
4. Programme Mechanisms	5 mins

As I understand, teachers who attend the Craft of Writing programme are asked to write outside the CPD sessions and residential weekends. Could you tell me about any writing you have done?
• If they have:
  ○ What kinds of things have you written about?
  ○ How did you find that process?
  ○ When did you find the time to do this?
• If they have not:	

○ Why do you think you didn't do any writing outside the sessions? ○ Could anything have helped you to do so? ○ Do you think this affected what you got out of the programme in any way? If yes, probe for detail. Can you describe any ways that taking part in the Craft of Writing programme has affected your teaching? ● Can you describe any skills or techniques you learnt at the Craft of Writing training sessions that you used with your class? ● How did your pupils respond to this? ● Can you describe any children for whom this worked particularly well? Why do you think that is? ● Can you describe any children for whom it worked less well with? Why do you think that is? ● How does the Craft of Writing approach compare to what you typically do in terms of writing lessons?	
5. Programme Impact	5 mins
What effect, if any, do you think taking part in the Craft of Writing programme has had for you? ● How do you think the programme has affected your writing? ○ Can you describe any effect it has had on your confidence as a writer? ● Can you describe any less positive impacts on the programme? ● For any changes specified, probe: What do you think it was about the programme that led to [change specified]? ● How are you intending to use what you've learnt through the Craft of Writing programme going forwards? What effect, if any, do you think you attending the Craft of Writing programme has had on your pupils? ● What effect do you think you attending the programme has had on your pupils writing skills? ● What do you think it is about the programme that has led to [any changes described]?	

• Can you tell me about any negative effects for your pupils of you attending the Craft of Writing programme? Can you describe any ways that you have shared what you learnt at the Craft of Writing training with other teachers or members of staff in this school? • Do you think this has impacted their teaching in any way? If yes, probe for detail. • How do you think you will use the Craft of Writing programme going forwards, if at all?	
6. Close	2 mins
Overall, would you recommend the Craft of Writing programme to other schools? • Why/Why not? Was there anything else that you were hoping to discuss that we haven't yet had a chance to talk about? Thank the interviewee for their time and reassure them of the confidentiality of their responses, as explained at the beginning of the interview.	Thank you and close.
7. Reflection following observation of CoW session	5 mins
Below are some exemplar questions, however, the interviewer may want to add some additional questions based on their observations. Probe for their perception of student engagement; implementer support; and mechanisms. How did you find that session went? • What do you think went well? • What do you think the challenges were? • Which students engaged particularly well? How/why? • Which students engaged less well? How/why? If you could go back and do that session again, what might you do differently?	Opportunity to follow-up on any areas of interest arising from the observation.

What skills or techniques that you learnt at the Craft of Writing training did you use in that lesson? How do you think those worked? I noticed this [describe significant moment in the lesson]: what was your view about what was happening?	

Appendix H: Senior leadership team (SLT) interview guide

Craft of Writing: Interviews with Senior Leadership Team

The interviews should last around 20 minutes. The timings given for each section are a guide - you may spend longer or shorter on each section. Lead questions are presented in bold, with potential follow-up questions presented in a non-bold typeface. As the interviews are semi-structured, not all questions need to be asked and they do not need to be asked in order. The interviewer should be responsive to what the interviewee, following the direction of the conversation and following-up with additional questions as needed.

Main objective	Purpose of section	Guide timings
1. Introduction	Explains the purpose and ground rules for the interview.	3 mins
2. Background context	Allows the participant to settle into the interview, as well as providing some background to the school, so that we understand more about the school context that the programme is intended to benefit.	2 mins
2. Engagement	To understand the extent to which the school, and particularly the SLT, has engaged with and supported the programme.	5 mins
3. Quality	This section will focus on understanding the perceived quality of the intervention, as well as the SLT's perception of the barriers and facilitators to engaging in the programme. We will also try to determine what 'business-as-usual' looks like within the school in terms of creative writing.	4 mins
4. Mechanisms	To explore the SLT member's perception of the impact of the programme for the teachers themselves and pupils, and the mechanisms that brought about any change identified, including the ways in which the training translated into teaching practice.	4 mins
5. Close	Thank you and close	2 mins

1. Introduction	3 mins
Introduction: • Introduce yourself • Introduce BIT and IOE – explain that we are independently evaluating the Craft of Music programme, which is one of five programmes that are part	Orientates respondent and gets them prepared to take part in the discussion. Outlines the 'rules' of the interview.

of the Cultural Learning programme that is jointly funded by the Education Endowment Foundation and Royal Society of Arts. Aims of this interview: We are here to learn more about your involvement in the Craft of Writing programme, what has helped the programme to work, and what the challenges have been. We'd also like to understand any impact the programme has had on your school, particularly for the teacher attending the Craft of Writing sessions and the pupils in his/her class. This interview: ● Should take no more than 25 minutes ● Stress that you want to understand the intervention from their point of view. No answers are right or wrong – and we are not here to judge the decisions made or views held by the interviewee. Anonymity and privacy: ● All information gathered will be in strict confidence, unless there are concerns about safeguarding. When we write up the research we will ensure that no one is identifiable from any reporting, ● Explain that if at any point they feel uncomfortable or prefer not to answer a specific question they can just say so. ● Explain that it is their choice whether they take part in the interview and they can end the interview at any point, without giving a reason. Recording: 1. Explain that recording enables us to have an accurate record of what was said, which can be typed up for analysis alongside other interviews. We may also use quotes from this interview, but these will be included in a way that means no individual or school is identifiable. 2. Check if they have any questions about the interview. If they are happy to go ahead, obtain verbal permission to digitally record and take notes (written permission should already have been obtained). 3. Once you have consent, start the voice recorder. 4. State interview number/participant ID	
2. Background Context	2 mins
How many years have you been teaching? How long have you been working at the school? Could you tell me about your role at the school?	Allows the participant to settle into the interview, as well as providing some background to the school, so that we understand

Could you tell me a little more about what the school is like? ● What would you say are the schools' main strengths? ● What would you say are some of the school's biggest challenges?	more about the school context in which the programme is intended to work.
2. Engagement	**5 mins**
Can you describe what you understand about what the Craft of Writing programme involves? What do you understand about what the programme is intended to achieve? How did your school became involved with the Craft of Writing programme? ● When did you first hear about the Craft of Writing programme? ● Why did your school decide to get involved in the programme? What initial expectations did you have for the programme? Can you describe what has your involvement with the Craft of Writing programme has been? ● Can you describe anything that you, or other members of SLT, have done to enable the programme to run? ● Can you describe any resources that you, or the wider SLT, have made available to enable the programme to run? ● How have you found making time for [teacher's name] to attend the Craft of Writing training? Follow-up on any challenges raised and how these were overcome.	To understand the extent to which the school, and particularly the SLT, has engaged with and supported the programme.
3. Quality	**4 mins**
Overall, what has your experience of the Craft of Writing programme been? ● What do you think is good about the programme? ● What do you think is less good about it? ● Can you describe any challenges you've had with the programme? If yes, how have these been overcome? ● Can you tell me any ways that you would improve the programme? What support or training, if any, do you typically provide to teachers to enhance their writing skills?	

• If they do, what do you think of the Craft of Writing approach in comparison? How are writing skills typically taught to pupils in Year 5 at this school? • What do you think the Craft of Writing programme offers in comparison to this?	
4. Mechanisms	
From what you understand, how has [the teacher attending Craft of Writing training] found the programme? • What effect, if any, do you think the programme has had for [name of teacher]? • What skills/techniques, if any, do you think [name of teacher] has developed as a result of attending the training sessions? • Can you describe any ways that the Craft of Writing has influenced the teacher's practice? ○ What effect do you think this has had for the pupils? • Probe for what it was about the programme that they think has led to the effects described. Can you describe any ways that [name of teacher] has shared learning from the Craft of Writing within the school? • If yes: ○ How was this received? ○ What impact do you think it had on the practice of other teachers within the school? • If no: ○ Why do you think the reasons are for [name of teacher] not sharing any of the learning?	
5. Close	2 mins
Overall, would you recommend the programme to other schools? • Why/Why not? Was there anything else that you were hoping to discuss that we haven't yet had a chance to talk about?	Thank you and close

Thank the interviewee for their time and reassure them of the confidentiality of their responses, as explained at the beginning of the interview.	

Appendix I: Observation proforma

Craft of Writing: Observation Proforma

Date and time: Name of school and class: Number of pupils present: Name of observer: Other adults present: Notes	
What is covered in the introduction to the session, e.g., are the aims and expectations covered? Is an expectation created that this lesson is a continuation of writing from a previous lesson?	
What activities are carried out as part of the session? Are they individual, small group or whole class activities? Is the writing activity a continuation of writing started in a previous lesson?	
Are the children given time to write freely (without being rushed)? To what extent are children given genuine choices (topic, form, collaboration, time spent e.g., more or less than one lesson) as part of their writing?	
Is there any discussion about language choices and how they affect writing? If yes, what is discussed?	
Is writing presented as an iterative process that involves stages of drafting and revising? If yes, describe how this is explained.	
What evidence is there of discussions about text-level features (e.g., form/structure/genre) including attention to readers/likely readers of the writing?	
How confident is the teacher delivering the session? How well does the teacher's knowledge appear to relate to the aims and expectations of CoW?	
To what extent are children engaged in the session, e.g., are they focused on writing, or are they distracted and talking to their peers about unrelated things?	
What is the behaviour of the pupils like, e.g., mainly on-task; mainly disruptive; individual children not on-task but the majority engaged?	

To what extent do children seem able to understand and follow the session? What difficulties do children who struggle seem to have? How are they supported?	
What strategies are used to account for the needs and abilities of different pupils?	
Any other observations	

Appendix J: Raw treatment condition survey data

Please note: some percentages may add up to slightly above or below 100% due to rounding.

1.1 What is your role at the school?	
Response	Number (%)
Class teacher	34 (94%)
Other members of SLT (e.g., Head of Key Stage 2)	4 (11%)
Deputy headteacher	1 (3%)
Other (please specify):	
English lead	1 (3%)
Oversee Key Stage 2 (KS2)	1 (3%)
Headteacher	0 (0%)
Teaching assistant	0 (0%)

N.B. Total percentage is greater than 100% because respondents could select more than one role.

2.1 To what extent do you agree or disagree with the following statement: I enjoy writing	
Response	Number (%)
Strongly agree	16 (44%)
Agree	18 (50%)
Neither agree nor disagree	1 (3%)
Disagree	1 (3%)
Strongly disagree	0 (0%)

2.2 To what extent do you agree or disagree with the following statement: I am a good writer	
Response	Number (%)

Strongly agree	4 (11%)
Agree	23 (64%)
Neither agree nor disagree	8 (22%)
Disagree	1 (3%)
Strongly disagree	0 (0%)

2.3 To what extent do you agree or disagree with the following statement: I enjoy teaching writing skills to my class

Response	Number (%)
Strongly agree	16 (44%)
Agree	19 (53%)
Neither agree nor disagree	0 (0%)
Disagree	1 (3%)
Strongly disagree	0 (0%)

3.1 Did you attend residential visits as part of the Craft of Writing programme?

Response	Number (%)
Yes, I attended both the residential visits	26 (72%)
Yes, I attended one of the residential visits	7 (19%)
No, I did not attend any visits	3 (8%)

N.B. Question 3.2 was not asked to participants who said they had attended both of the residentials.

3.2 Which of the following, if any, are reasons for not attending both of the Craft of Writing residentials?

Response	Number (%)

I had personal commitments	4 (11%)
The teacher taking part changed during the school year	3 (8%)
The teacher taking part left during the school year	1 (3%)
I could not find cover	1 (3%)
I did not know about the training	1 (3%)
Other (please specify): did not have childcare	1 (2.78%)
Cost of travel and/or subsistence	0 (0%)
I could not travel the distance required	0 (0%)
I did not feel my my attendance would be useful	0 (0%)
I did not receive time off in lieu	0 (0%)
I had competing school commitments	0 (0%)
I was unwell	0 (0%)
SLT were not supportive of me attending	0 (0%)

N.B. Respondents could select more than one answer.

3.3 How many continuing professional development (CPD) sessions did you attend as part of the Craft of Writing programme?	
Response	Number (%)
Three	23 (64%)
Two	10 (28%)
None	2 (6%)
One	1 (3%)

N.B. Question 3.4 was not asked to participants who said they had attended ‘3’ CPD sessions.

3.4 Which of the following, if any, are reasons for not attending all of the CPD days?	
Response	Number (%)
I had competing school commitments	2 (6%)
I had personal commitments	2 (6%)
I could not find cover	2 (6%)
I was unwell	2 (6%)
I could not travel the distance required	1 (3%)
The teacher taking part changed during the school year	1 (3%)
The teacher taking part left during the school year	1 (3%)
I did not know about the training	1 (3%)
Other (please specify):	
car breakdown	1 (3%)
during SATs week	1 (3%)
hospital appointment	1 (3%)
Cost of travel and/or subsistence	0 (0%)
I did not feel my attendance would be useful	0 (0%)
I did not receive time off in lieu	0 (0%)
SLT were not supportive of me attending	0 (0%)

N.B. Respondents could select more than one answer.

4.1 Which of the following techniques or strategies, if any, have you tried to incorporate into the classroom?	
Response	Number (%)
Teaching about the effects on readers of choosing specific words, phrases and other writing devices	35 (97%)
Free writing (i.e., giving children the opportunity to write continuously for a set period, without concern for spelling, grammar or topic)	34 (94%)
Teaching about writing for different audiences	32 (89%)

Providing opportunities for children to write in a way that involved different phases of writing beyond ‘rough’ copy ‘neat’ copy	31 (86%)
Encouraging children to choose their own topic to write about	27 (75%)
None of the above	0 (0%)
Other (please specify)	0 (0%)

N.B. Respondents could select more than one answer.

4.2 How did you find adapting the strategies and techniques you learnt from the residentials and/or CPD to the classroom?	
Response	Number (%)
Very easy	2 (6%)
Easy	13 (36%)
Neither easy nor hard	14 (39%)
Hard	5 (14%)
Very hard	0 (0%)
Don’t know	2 (6%)

5.1 What impact, if any, do you think the intervention had on your confidence as a writer?	
Response	Number (%)
Very positive impact	24 (67%)
Somewhat positive impact	10 (28%)
No impact	2 (6%)
Somewhat negative impact	0 (0%)
Very negative impact	0 (0%)
Don’t know	0 (0%)

5.2 How confident, if at all, do you feel about the following: I understand the purpose of Craft of Writing.

Response	Number (%)
I feel very confident	21 (58%)
I feel somewhat confident	12 (33%)
I feel neither confident nor underconfident	2 (6%)
I feel somewhat unconfident	0 (0%)
I do not feel confident at all	0 (0%)
Don’t know	1 (3%)

5.3 How confident, if at all, do you feel about the following: I can equip children with the skills to write creatively.

Response	Number (%)
I feel very confident	19 (53%)
I feel somewhat confident	14 (39%)
I feel neither confident nor underconfident	3 (8%)
I feel somewhat unconfident	0 (0%)
I do not feel confident at all	0 (0%)
Don’t know	0 (0%)

5.4 How confident, if at all, do you feel about the following: I can keep children engaged in creative writing.

Response	Number (%)
I feel very confident	21 (58%)
I feel somewhat confident	13 (36%)
I feel neither confident nor underconfident	1 (3%)
I feel somewhat unconfident	1 (3%)
I do not feel confident at all	0 (0%)
Don’t know	0 (0%)

5.5 How confident, if at all, do you feel about the following: I can structure creative writing sessions	
Response	Number (%)
I feel very confident	18 (50%)
I feel somewhat confident	14 (39%)
I feel neither confident nor underconfident	3 (8%)
I feel somewhat unconfident	1 (3%)
I do not feel confident at all	0 (0%)
Don’t know	0 (0%)

5.5 How confident, if at all, do you feel about the following: I have all the materials to deliver creative writing sessions	
Response	Number (%)
I feel very confident	18 (50%)
I feel somewhat confident	14 (39%)
I feel neither confident nor underconfident	4 (11%)
I feel somewhat unconfident	0 (0%)
I do not feel confident at all	0 (0%)
Don’t know	0 (0%)

5.6 How confident, if at all, do you feel about the following: I have the space required to deliver creative writing sessions	
Response	Number (%)
I feel very confident	19 (53%)
I feel somewhat confident	11 (31%)
I feel neither confident nor unconfident	4 (11%)
I feel somewhat unconfident	2 (6%)

I do not feel confident at all	0 (0%)
Don’t know	0 (0%)

6.1 Thinking about your class as a whole, what kind of impact, if any, do you think your Craft of Writing experience has had on: pupils’ confidence about writing tasks	
Response	Number (%)
Very positive impact	17 (47%)
Somewhat positive impact	18 (50%)
No impact	1 (3%)
Somewhat negative impact	0 (0%)
Very negative impact	0 (0%)
Don’t know	0 (0%)

6.2 Thinking about your class as a whole, what kind of impact, if any, do you think your Craft of Writing experience has had on: pupils’ engagement with writing activities	
Response	Number (%)
Very positive impact	19 (53%)
Somewhat positive impact	16 (44%)
No impact	1 (3%)
Somewhat negative impact	0 (0%)
Very negative impact	0 (0%)
Don’t know	0 (0%)

6.3 Thinking about your class as a whole, what kind of impact, if any, do you think your Craft of Writing experience has had on: pupils’ ability to generate ideas	
Response	Number (%)
Very positive impact	10 (28%)

Somewhat positive impact	25 (69%)
No impact	1 (3%)
Somewhat negative impact	0 (0%)
Very negative impact	0 (0%)
Don’t know	0 (0%)

6.4 Thinking about your class as a whole, what kind of impact, if any, do you think your Craft of Writing experience has had on: pupils’ ability to write creatively	
Response	Number (%)
Very positive impact	0 (0%)
Somewhat positive impact	23 (64%)
No impact	1 (3%)
Somewhat negative impact	0 (0%)
Very negative impact	0 (0%)
Don’t know	0 (0%)

6.5 Thinking about your class as a whole, what kind of impact, if any, do you think your Craft of Writing experience has had on: pupils’ ownership of written work	
Response	Number (%)
Very positive impact	22 (61%)
Somewhat positive impact	11 (31%)
No impact	3 (8%)
Somewhat negative impact	0 (0%)
Very negative impact	0 (0%)
Don’t know	0 (0%)

6.6 Thinking about your class as a whole, what kind of impact, if any, do you think your Craft of Writing experience has had on: pupils' overall writing skills	
Response	Number (%)
Very positive impact	8 (22%)
Somewhat positive impact	25 (69%)
No impact	3 (8%)
Somewhat negative impact	0 (0%)
Very negative impact	0 (0%)
Don't know	0 (0%)

7.1 Overall, how would you rate Craft of Writing?	
Response	Number (%)
Very good	25 (69%)
Good	8 (22%)
OK	3 (8%)
Poor	0 (0%)
Very poor	0 (0%)
Don't know	0 (0%)

Appendix K: Raw control condition survey data

Please note: some percentages may add up to slightly above or below 100% due to rounding.

1.1 What is your role at the school?	
Response	Number (%)
Class teacher	34 (92%)
Head or Deputy Head	2 (5%)
Other members of SLT	1 (3%)

2.1 Thinking back over the last academic year, have you received any training or continuing professional development focused on teaching writing skills?	
Response	Number (%)
Yes	25 (68%)
No	12 (32%)

2.2 Thinking back over the last academic year, have you received any training or CPD focused on developing your own writing skills?	
Response	Number (%)
No	32 (86%)
Yes	4 (11%)
Don't know	1 (3%)

3.1 How confident do you feel teaching writing, where 1 is not confident at all and 5 is very confident.	
Response	Number (%)
5	9 (24%)

4	21 (57%)
3	7 (19%)
2	0 (0%)
1	0 (0%)

4.1 Over the last academic year, on average, how often have your pupils engaged in writing activities?

Response	Number (%)
Daily	25 (68%)
Several times a week	12 (32%)

4.2 Which of the following things have you done when teaching writing to your class this academic year?

Response	Number (%)
Teaching about the effects on readers of choosing specific words, phrases and other writing devices	34 (92%)
Free writing (i.e., giving children the opportunity to write continuously for a set period, without concern for spelling, grammar or topic)	15 (41%)
Teaching about writing for different audiences	36 (97%)
Providing opportunities for children to write in a way that involved different phases of writing beyond ‘rough’ copy ‘neat’ copy	33 (89%)
Allowed children to choose their own topic to write about	13 (35%)
None of the above	0 (0%)
Other (please specify)	0 (0%)

N.B. Respondents could select more than one answer.

Appendix L: The Craft of Writing Memorandum of Understanding (MoU) for participating schools

This project is exploring how the opportunity for teachers to work with professional writers might change their understanding of being a writer, how they teach writing, and improve outcomes in writing for the children they teach. Its impact will be evaluated by comparing it with the 'teaching as usual' approach using a randomised controlled trial (RCT).

During this project, you will be contacted by both the Project Team (University of Exeter, Open University and Arvon), who are responsible for developing and supporting the new teaching approach, and by the Evaluation Team (University College London [UCL] and Behavioural Insights Team [BIT]), who are carrying out an independent evaluation of its effectiveness.

This memorandum of understanding (MoU) explains what your school's participation in the study will entail.

In order to formalise your consent to participate, please choose the relevant button above.

Randomised Controlled Trial (June 2018–July 2019)

The trial will involve your school being randomly assigned either to participate in the Craft of Writing intervention (the intervention group) or to continue with your normal teaching approach (the comparison group). Teachers in the intervention group will attend two 3 day residentials and 3 CPD days across the year, and to teach writing drawing on the learning from the residentials and CPD input. Teachers in the Comparison group will teach writing as they normally do. Schools in the intervention group will be asked to pay £500 to participate as a partial contribution to the costs of the residentials and CPD days, and will receive a minimum of £500 towards supply costs. Schools in the comparison group will receive a £500 payment, which can be used towards the cost of attending an Arvon Teachers as Writers residential after the project has ended, if desired.

The following information and evaluation data will be required by the Evaluation and Project teams:

Prior to randomisation

Schools will:

Provide school URN and LAESTAB number.

Provide contact details of the Year 5 Project Teacher (valid email address and telephone number) to the Project Team for use by both the Project Team and the Evaluation team.

Provide, via the means specified by the evaluation team, pupil names, DOB and Unique Pupil Numbers (UPNs) of the Year 5 Project Class, along with details of any setting or streaming by attainment, to the Evaluation Team by the end of March 2018.

Follow the secure procedures requested by Evaluation team to allow parents to opt their pupils out of the evaluation part of this research.

During the evaluation

Participating teachers will:

Update UPNs and pupil names of Year 5 Project Class by the end of September 2018.

Facilitate visits by the Project Team and/or the Evaluation Team to gather data on the implementation of the intervention or on routine teaching (for example, through observations or interviews etc).

Facilitate an-end-of project written assessment which will be administered by the Evaluation Team.

Use of Data

All pupil data will be treated with the strictest confidence and will be stored in accordance with the Data Protection Act (1998) and with the forthcoming General Data Protection Regulation. Named data will be matched with the National Pupil Database using pupils' UPNs by the Evaluation Team and shared (anonymously) with the Education Endowment Foundation. All results will be anonymised so that no schools will be identifiable in the report or dissemination of any results. For the purpose of research, the pupil data will be linked with information from the National Pupil Database held by the Department for Education, other official records, and shared with Exeter University, Open University, the Department for Education, EEF, EEF's data contractor FFT Education and in an anonymised form to the UK Data Archive and for research purposes. Confidentiality will be maintained at all times.

Requirements for Schools

The school is not participating in another research project or evaluation that would interfere with development and evaluation of the above approach in Year 5 writing.

The Year 5 Project Teacher will be working with a Year 5 class in the academic year 2018-2019.

Intervention group teachers will attend the two residentials and the three CPD days.

The school will deliver letters to parents giving them information about the study and an opportunity to opt their child out of the data gathering process. They will inform the Evaluation Team of any responses arising.

The school will provide data requested to the Project Team and the Evaluation Team as detailed above. The school will permit the publication of anonymised data collected and its use in presentations.

Teachers will, at the earliest opportunity, notify the Project Team if there are support or operational issues which could affect their participation.

Teachers will complete a survey for each child in their class at the beginning and end of the study.

If the school has to withdraw from the project for operational or other unavoidable reasons, it will notify the UCL/BIT Evaluation Team straight away and, wherever possible, still provide test data for the project.

Responsibilities of the Project Team

- Provide the Arvon residentials and 3 CPD days
- Act as the first point of contact for any questions about the project Provide on-going support to the school
- Provide information sheets for parents
- Collect Year 5 Project Teacher contact names and email details. Disseminate the research findings.

Responsibilities of the Evaluation Team

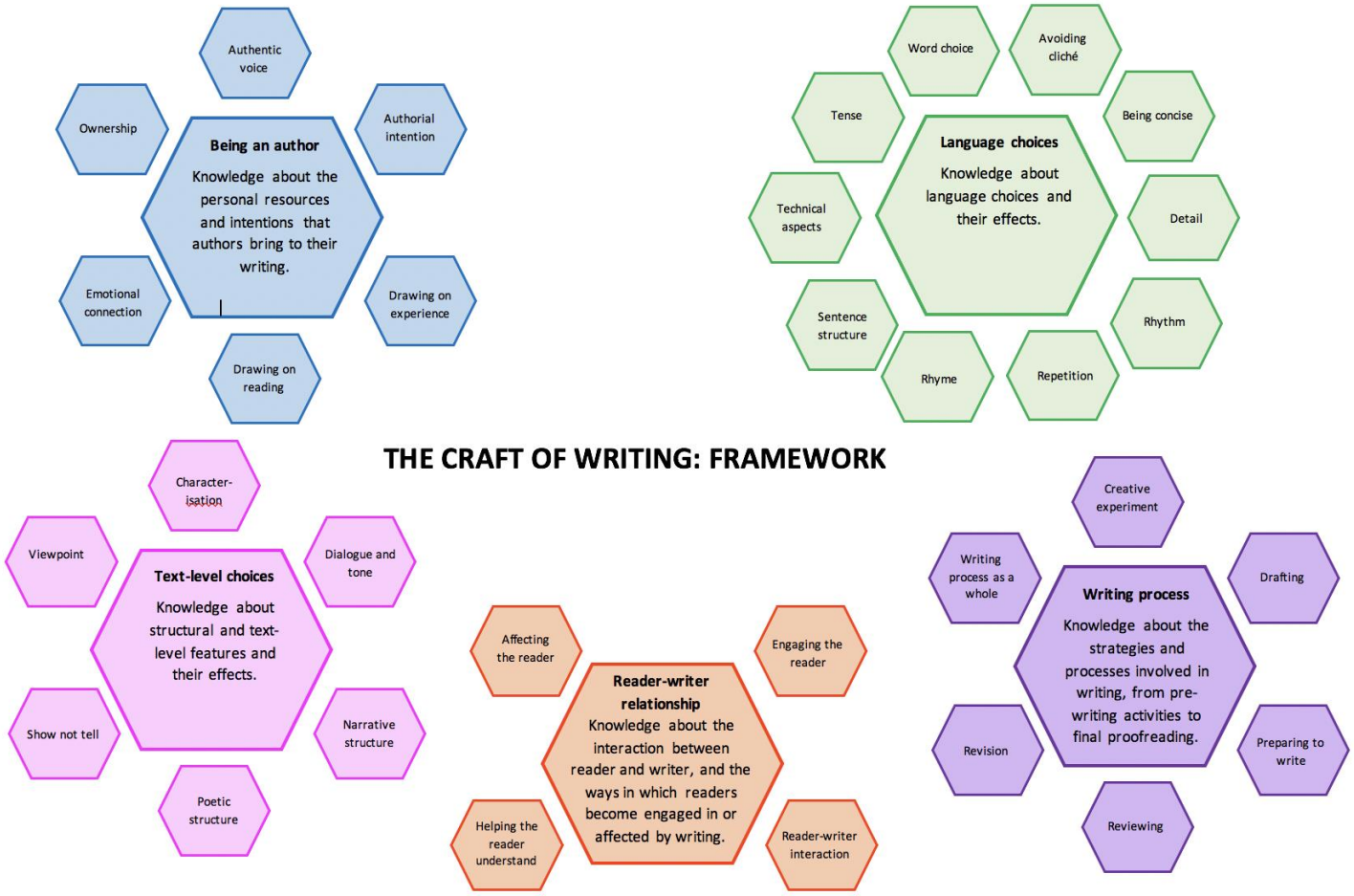
- Conduct the random allocation
- Work with the project team and schools to collect class and pupil level data (including name, date of birth, UPN)
- Collect data about how the schools are implementing the intervention, or teaching as usual (control groups) Request NPD data using pupil details
- Administer the writing assessment
- Analyse the data from the project

- Disseminate the research findings

In order to formalise your consent to participate, please click on the relevant link below (make sure you choose the right button for your role!).

- Teacher consent form
- Headteacher consent form

Appendix M: CoW Framework



Appendix N: WAM prompt

Adapted from Dunsmuir et al. (2015)

Your name: _____

Imagine that you are going to go on a school trip with your class, your teacher, and some other adults. Imagine you could go anywhere at all.

Write a short story about your chosen trip, and some of the things that you did while on the trip.

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

Appendix O: WSEM prompt

Adapted from Bruning et al. (2013)

Thinking about myself as a Writer

Name: _____

Instructions: Put a tick in **one** box for each question below to show how confident **YOU** feel about the different things a writer does. This isn't a test. We want to know how you really feel when you are writing.

	I'm sure I can't do it	I don't think I can do it	I'm not sure if I can do it or not	I can mostly do it	I'm very sure I can do it
1. I can think of lots of new ideas for my writing.					
2. I can use my ideas in my writing.					
3. I can think of the words I need to write down my ideas.					
4. I can show how I feel in my writing					
5. I know where to place my ideas in my writing.					
6. I can spell my words correctly.					
7. I can write whole sentences.					
Please turn over to the other side					
	I'm sure I can't do it	I don't think I can do it	I'm not sure if I can do it or not	I can mostly do it	I'm very sure I can do it
8. I can use punctuation marks (like full stops and					

question marks) correctly.					
9. I can begin my paragraphs in the right places.					
10. I can focus on my writing for at least half an hour.					
11. I can start writing quickly.					
12. I am always calm and in control when I write.					
13. I can think of my goals for my writing before I write.					
14. I can keep on writing even when it gets difficult.					
15. I enjoy writing.					
16. I am confident as a writer					

Thank you for helping us by thinking about writing.

Appendix P: School and parent information sheets, objection forms, and revised GDPR privacy notice

School information sheet

THE CRAFT OF WRITING

What is this about?

Arvon, the Open University, and the University of Exeter are working together on a project funded by EEF and the RSA, exploring how the opportunity for teachers to work with professional writers might change their understanding of being a writer, how they teach writing, and improve outcomes in writing for the children they teach. The effectiveness of the project will be evaluated by a team from University College London and Business Insights Team. **This is a rare opportunity for deep, sustained professional development in the teaching of writing!**

What will the project look like?

The project is investigating the effect of a residential writing experience for teachers at Arvon, where teachers work with professional writers learning about writing and being a writer. We intend to work with 96 primary school teachers, and their year 5 classes during 2018–19. Teachers in intervention schools (see below) will participate in two 3-day Arvon writing residentials at Lumb Bank, one of Arvon's beautiful residential centres. The 'teachers as writers' residentials provide space for writing workshops, for one-to-one tutorials with writers, and for time and space to write as well as reflecting on ideas to take back to the classroom. Three linked professional development days will support you in thinking through the implications of the residentials for your teaching, and explore classroom practices which will motivate children to write and strengthen their understanding of how to revise and evaluate their texts. You can find out more about the value of an Arvon residential from www.arvon.org.

What are 'intervention' and 'comparison' schools?

An important element of these EEF projects is that schools interested in participating are randomly allocated to either the intervention group, who will receive the training, or a comparison group, who contribute to the data required for comparison. It is important that all interested schools understand that they could be in either group. ***It is important that you understand you could be a comparison school.***

What commitment would this project require?

If you are an 'intervention' school, you would need to commit to allowing the Year 5 project teacher to:

- provide student profile and attainment data and UPNs;
- attend the two Arvon residentials in June 2018 and Jan 2019;
- attend three professional development days between June 2018 and May 2019;
- allow the research team access to collect data (for example through observations and interviews);
- allow the evaluation team to visit your school to administer a writing assessment at the end of the project;

complete a survey for each child in the project class at the beginning and end of the study.

If you are a 'comparison' school, you would need to commit to allowing the project teacher to:

- provide student profile and attainment data and UPNs;
- allow the evaluation team to visit your school to administer a writing assessment at the end of the project;
- complete a survey for each child in the project class at the beginning and end of the study.

Finances

- Schools in the **intervention** group will be asked to pay £500 to participate as a partial contribution to the costs of the residentials and 3 CPD days, but will also receive a minimum of £500 towards the supply cover costs;

Schools in the **comparison** group will receive a £500 payment, which can be used towards the cost of attending an Arvon residential after the project is complete.

If you would like to know more, or if you have any questions, please contact: XXX

Parent information sheet

THE CRAFT OF WRITING

What is this about?

Arvon Writing Retreats, the Open University, and the University of Exeter are working together on a project funded by Education Endowment Foundation and the Royal Society of the Arts, exploring how the opportunity for teachers to work with professional writers might change their understanding of being a writer, how they teach writing, and improve writing outcomes for the children they teach. The effectiveness of the project will be researched by a team from UCL Institute of Education and the Behavioural Insights Team (the 'evaluation team'). This research has been reviewed and approved by the research ethics committees of UCL Institute of Education, University of Exeter and the Open University.

What will the project look like?

The project investigates the effect of The Craft of Writing programme, run by a team from the University of Exeter, the Open University and Arvon Writing Retreats (the 'project team'), which will work with teachers to train them in different methods of teaching writing that we think will benefit your child's learning.

We plan to work with 96 primary school teachers, and their Year 5 classes during 2018–19. Teachers in 'intervention' schools (see below) will participate in residential writing workshops (including one-to-one tutorials with writers, with the time and space to write as well as reflect on ideas to take back to the classroom) and three professional development days specifically to link these residential workshops to your child's teaching. You can find out more about Arvon Writing Retreats from www.arvon.org.

What are 'intervention' and 'comparison' schools?

An important element of these EEF projects is that schools are randomly chosen either to be in the intervention group, who will receive the training this year, or a comparison group, who contribute to the data required for comparison (and might choose to do the training in future).

Whether your child's teacher will receive the training this year ('intervention' schools) or not ('comparison schools') will be randomly decided by evaluators from UCL Institute of Education and the Behavioural Insights Team to help them understand how effective the training has been. If your child's teacher does not receive the training this year, they will receive a payment that will allow them to go on the training next year (or towards other training, if they prefer).

What does this mean for me as a parent?

As part of measuring the success of this training programme, your child will be asked to complete a writing test and survey during their normal classes towards the end of the school year. This will take about half an hour. We are doing this test for the purposes of the research project, to help us understand if the Craft of Writing programme helps children like yours with their writing. We will also obtain your child's UPN (Unique Pupil Number) to allow longer term understanding of whether this writing programme worked. We will then save this information in a data format that will prevent anyone from identifying your child.

This data will then by linked with the National Pupil Database (held by the Department for Education, part of the UK government) and shared with the project team, the Department for Education, the Education Endowment Foundation (EEF, who are funding this research), EEF's data contractor FFT Education and kept in an anonymised form in the UK Data Archive. No information that can identify individual children will be made available to anyone outside these teams and your child's school. This data will be kept securely under password protection. We will not use your child's name or

the name of the school in any report arising from the research, and no information that could otherwise identify your child will be made public.

Although we think the project will help your child, you have the right to ask us not to use your child's data in this way. Please see the letter that came with this information sheet or contact your child's class teacher. If you have any concerns and would like to know more, or if you have any questions, please contact XXX.

Objection letter

Dear Parent / Carer,

Your child's school has applied to take part in research that aims to improve their self-confidence and performance in writing. The Craft of Writing programme, run by a team from the University of Exeter, the Open University and Arvon Writing Retreats (the 'project team'), will work with teachers to train them in different methods of teaching writing that we think will benefit your child's learning.

This process will then be researched by a team from UCL Institute of Education, the Behavioural Insights Team, along with some evaluation by the Royal Society of Arts (the 'evaluation team'). There's more information on the information sheet that came with this letter.

Not all teachers in the study will necessarily receive the training this year. Whether your child's teacher will receive the training this year will be decided by the evaluation team at random to help them understand how effective the training has been.

As part of measuring the success of this training programme, all Year 5 children will complete a classroom based writing test and survey towards the end of the year. This will take about half an hour. We are doing this test for the purposes of the research project, to help us understand if the Craft of Writing programme helps children like yours with their writing.

Your child's name and other data held by the school, alongside their writing test scores, will be collected by the evaluation and project teams. No information that can identify individual children will be made available to anyone outside these teams and your child's school. This data will be kept securely under password protection. We will not use your child's name or the name of the school in any report arising from the research, and no information that could otherwise identify your child will be made public.

We will also obtain your child's UPN (Unique Pupil Number) to allow us to link up our data with the National Pupil Database (held by the Department for Education) and other official records to understand whether being part of this project is linked with test scores when they are older. This involves us sharing data with the Department for Education (part of the UK Government), the Education Endowment Foundation (EEF, who funded the trial), EEF's data contractor FFT Education and in a form that will prevent anyone from identifying your child to the UK Data Archive.

This research has been reviewed and approved by the research ethics committees of UCL Institute of Education, University of Exeter and the Open University.

If you have any questions you would like to ask before replying, please contact the research team by XXX.

Because we are doing this research to improve understanding about what works in improving pupils' education, if you are happy for information about your child to be used in the Craft of Writing research project you do not need to do anything. Thank you for your help with this research, your support is much appreciated.

If you **DO NOT** want your information about your child to be used to understand whether the Craft of Writing programme can help children to write better, please complete the enclosed form and return it to your child's school by [INSERT DATE]. If you do this, then no information about your child will be shared with the evaluation or project teams at any point during the project.

Craft of Writing research programme

(If you are happy for your child to participate in the research on whether this programme improves writing, you DO NOT need to return this form.)

I **DO NOT** wish my data about my child to be collected as part of this research.

Child's name: ……Date of birth: ……

Child's class Teacher: ……..

School:……

Parent name (BLOCK CAPITALS) ……

Parent signature: ……

Date ……

(Please return the completed form to your child's class teacher.)

Revised GDPR privacy information

EEF/RSA EVALUATION OF THE CRAFT OF WRITING

Data Privacy Notice

Dear Parent,

We've previously been in touch because your school is taking part in a project funded by the Education Endowment Foundation (EEF) to understand the potential benefits of The Craft of Writing. As part of that information, you were given the opportunity to tell your school not to pass any data about your child to us to be used as part of this project. **Please rest assured that if you contacted the school to make this request, nothing in this letter changes that.**

We wanted to get in touch again to provide you with further details about the way will be handling pupils' data as part of this project. **It is very important to us that that we do this responsibly and providing these details are an important part of that.** They are also important in fulfilling our responsibilities under the UK's data protection laws, which we take very seriously. These require us to provide you with some specific information about **our plans** and **your rights**.

Some of this description involves rather technical terms, which we've left in so you know the official concepts we are talking about. We've tried to keep the explanations as simple as possible. If we haven't managed that well enough and you have any questions now, or at any point during this project, then you should contact a member of the team with the first point of call being XXX.

Our Plans

Using pupils' data as part of research is not something we do without thinking about it. Under data protection law, we require a 'lawful basis' for the data processing that we carry out. UCL will be using the lawful basis known as the 'public task' basis, while the lawful basis BIT are using is known as the 'legitimate interests' basis (it is different at UCL and BIT because UCL is a university). To use the 'legitimate interests' basis, we must consider why this is a legitimate interest and inform you of this. Here, it is because our work is for the purpose of promoting the education or well-being of children in England and couldn't be achieved without analysing these test scores. We balance that against your and child's rights by providing you with the right to object to our use of your child's data in this way.

Your child's name and other data held by the school, alongside the test scores we will collect, will be collected and processed by us for the purpose of understanding how participating in this project has affected their learning. No information that can identify individual children will be made available to anyone outside these teams and your child's school (with an important exception, as explained below). We will also obtain your child's UPN (Unique Pupil Number) from the school to allow us to link up our data with the National Pupil Database (held by the Department for Education) and other official records to understand whether being part of this project is linked with test scores when they are older. This involves us sharing data with the Department for Education (part of the UK Government), the Education Endowment Foundation (who funded the trial), EEF's data processor FFT Education and (in a form that will prevent anyone from identifying your child) to the UK Data Archive.

The data we hold will be kept securely at all times, transferred using secure (encrypted) methods, and kept on secure computer systems at UCL and BIT's offices under password protection. We will never use your child's name or the name of the school in any report arising from the research, and no information that could otherwise identify your child will be made public.

Pupils' personal data will be processed by us only for the purposes of this research project. Once that is complete then the data will be securely destroyed from our computer systems. Personal data will certainly not be more than 10 years, in line with UCL's policy on storing research data.

Your rights

Under data protection law, you have a right to be informed about our plans. This letter, as well as the information that you previously received from your school, are all part of this. You also need to know exactly who is involved in the data

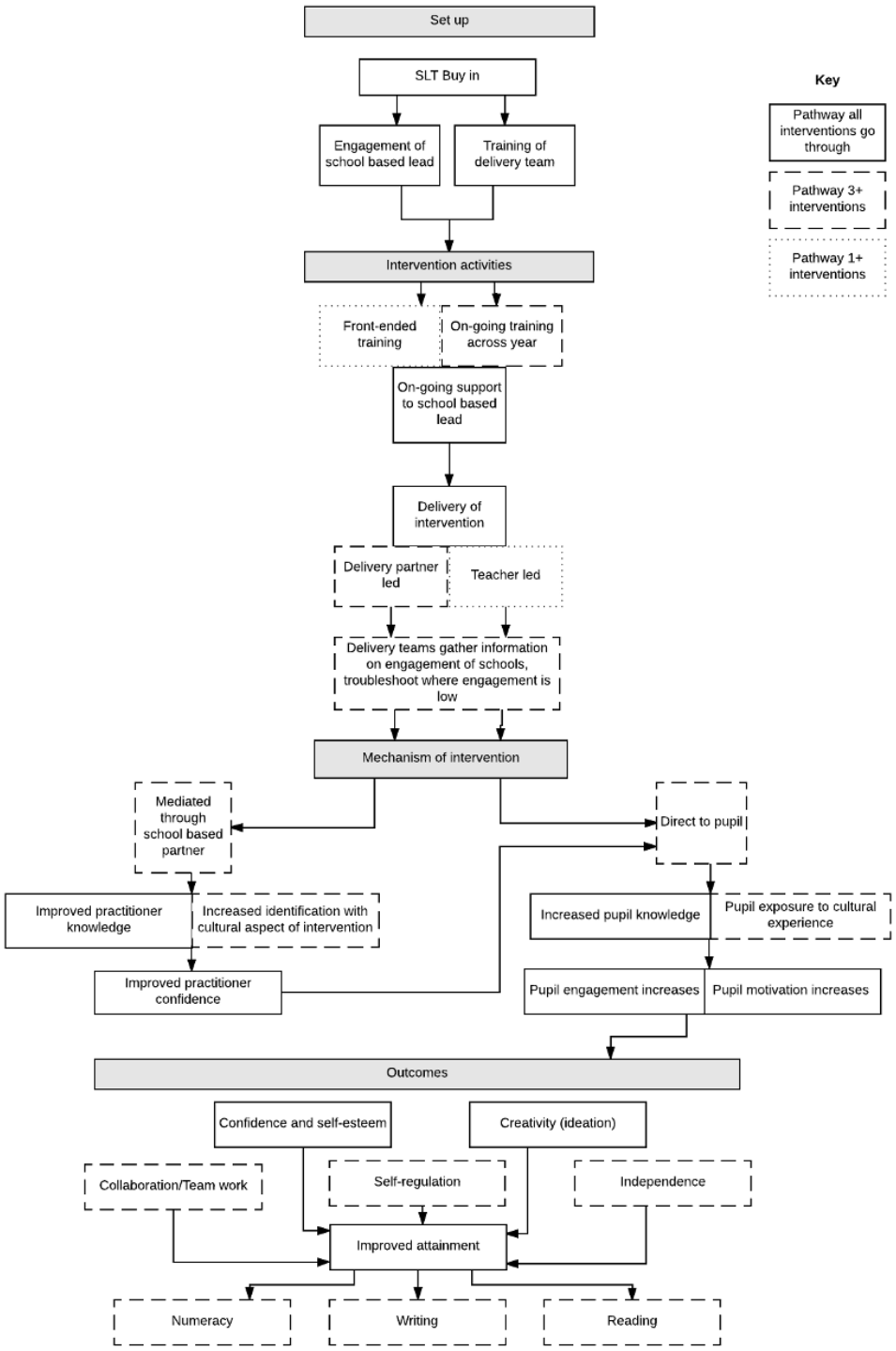
processing. In legal terms, University College London (UCL) and the Behavioural Insights Team (BIT) are considered joint data controllers for this project. The law requires our organisations to have named Data Protection Officers, who are ultimately responsible for overseeing data processing that goes on in their respective organisations. UCL's Data Protection Officer is XXX, who can be contacted via data-protection@ucl.ac.uk. BIT's Data Protection Officer is XXX, who can be contacted via XXX. You should contact these individuals if you have any complaints about how we are processing data. However, if you remain unsatisfied, you may wish to contact the Information Commissioner's Office (ICO). Contact details, and details of data subject rights, are available on the ICO website at: https://ico.org.uk/for-organisations/data-protection-reform/overview-of-the-gdpr/individuals-rights/

As noted above, we provided you with the right to object to data processing before schools handed any information over to us. You can also contact us at any point during the project to request information we hold about your child, to request rectification of any information that is incorrect, to stop using their data as part of the project or to destroy their data. If you wish to make such a request or ask any questions about it then please contact us. The best place to start is to contact XXX.

Once again, we are extremely grateful to you for supporting this project. We hope to learn a lot about the role cultural and arts education can play in supporting pupils' learning.

Yours sincerely,

Appendix Q: Amalgamated logic model for all five LAC trials



Appendix R: Qualitative analysis framework

1. Background	1.1. School/class strengths and challenges
	1.2 Teacher's experience of teaching (inc. writing/English)
	1.3 Teacher's attitude towards CoW/writing
2. School environment & support	2.1 Fit with the curriculum/other teaching requirements
	2.2 SLT attitude and support
	2.3 Comparison with how school typically teaches English/writing
3. Training	3.1 Training content
	3.2 Experience of training
	3.3 Making time to attend training/residentials
	3.4 Translating training into the classroom: free writing
	3.5 Translating training into the classroom: reader–writer relationship
	3.6 Translating training into the classroom: giving feedback
	3.5 Challenges translating into classroom: general/other
4. Children's experience of CoW	4.1 Adjustment to different style
	4.2 Struggling
	4.3 Engagement
	4.4 Extra support given to children for CoW
5. Impacts	5.1 Impacts on children
	5.2 Impact on teacher's writing
	5.3 Impact on teacher's teaching
	5.4 Impact on wider school

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