



Education
Endowment
Foundation

Grammar for Writing

Evaluation report and executive summary

February 2019

Independent evaluators:

Louise Tracey, Jan R. Boehnke, Louise Elliott, Kate Thorley, Sarah Ellison and Claudine Bowyer-Crane



UNIVERSITY
of York



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 Trust (now part of Impetus - Private Equity Foundation) 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
Research and Publications Manager

Education Endowment Foundation
5th Floor, Millbank Tower
21–24 Millbank
SW1P 4QP
p: 020 7802 1679
e: danielle.mason@eefoundation.org.uk
w: www.educationendowmentfoundation.org.uk

About the evaluator

The independent Evaluation Team was led by Dr Louise Tracey (University of York) and included Dr Jan R. Boehnke (University of Dundee), Mrs Louise Elliott (University of York) and Dr Claudine Bowyer-Crane (University of York). Kate Thorley, Sarah Ellison, Imogen Fountain, Mary Robison, Madeline Crossthwaite and Niamh Robinson provided research support at various stages of the study. Dr Pam Hanley (University of York) was involved in writing the original project proposal. The Evaluation Team was responsible for the conduct of the research, including the randomisation, data collection, analysis and reporting of the study.

e. louise.tracey@york.ac.uk

Contents

Executive summary.....	4
Introduction	6
Methods	14
Impact evaluation	31
Process evaluation.....	50
Conclusion.....	68
Appendix A: EEF cost rating.....	74
Appendix B: Security classification of trial findings.....	75
Appendix C: Information and Consent Forms	76
Appendix D: Technical Report	90

Executive summary

The project

Grammar for Writing is a way of teaching writing designed to help pupils to understand how linguistic structures convey meaning, rather than teaching grammatical rules in the abstract. The teaching of the grammar is therefore explicit, but embedded in the context of teaching about writing genres (e.g. narrative and persuasive writing). The aim is to improve pupils' "metalinguistic awareness" – their understanding of the language choices they make when they write. This study builds on a previous evaluation of Grammar for Writing funded by the EEF.

Teachers received three days of training to develop their grammar subject knowledge and to prepare them to teach two units of work using the Grammar for Writing approach. A fourth day of training covered future use of the approach beyond the two specific units. The units were delivered by the class teacher to Year 6 pupils (aged 10-11) in a whole class setting in place of existing writing lessons. The units were designed to be delivered in daily 1 hour sessions. The first unit lasted for four weeks and the second unit for two weeks. Training in the programme was provided by the Development Team, University of Exeter, supported by Babcock LDP.

The evaluation was a randomised controlled trial involving 155 schools. The primary outcome was writing skills, measured using a bespoke test based on previous Key Stage 2 (KS2) assessment papers. Alongside the impact evaluation an implementation and process evaluation was conducted which involved a teacher survey and a visit to a sample of schools to conduct lesson observations and teacher interviews. The evaluation took place between October 2016 and July 2017.

Key conclusions

1. The project found no evidence that Grammar for Writing improves writing attainment for children in Year 6, as measured by the bespoke test.
2. The project found no evidence that Grammar for Writing improves reading, writing or grammar, punctuation and spelling (GPS) as measured by KS2 SATS. Indeed, it found a small, negative effect size (equivalent to one month less progress) for the GPS outcome.
3. Pupils that have ever been eligible for free school meals made a small amount of additional progress compared to similar pupils in control schools. This result is not statistically significant. This means that the statistical evidence does not meet the threshold set by the evaluator to conclude that the true impact was not zero.
4. Grammar knowledge as measured using a teacher quiz did not improve for teachers who had done Grammar for Writing, although there was some evidence that this quiz was not a reliable measure. In contrast, more than 90% of surveyed teachers agreed that they found the programme, training and materials useful in their teaching.
5. Nearly three-quarters of intervention teachers indicated that they had adapted the programme for delivery. In addition, fidelity to two of the key programme principles, 'connections made between grammar and effect/purpose in writing' and 'discussion used to tease out thinking and choice-making' was regarded by the evaluator to be compromised in a number of the schools observed.

EEF security rating

These findings have a moderate to high security rating. The trial was an effectiveness trial, which tested whether the intervention worked under everyday conditions in a large number of schools.

The trial was a well-designed two-armed randomised controlled trial. The trial was well-powered. The pupils in Grammar for Writing schools were similar to those in the comparison schools in terms of prior

attainment. However, the security of the trial was reduced because more than 28% of the pupils who started the trial were not included in the final analysis, due to a large number of schools withdrawing from the study, and problems with the testing at the end of the trial caused in part by a change in the outcome measure, as discussed in the Methods section.

Additional findings

The main analysis for the impact evaluation found no evidence that the children in the intervention schools had improved their writing skills at the end of Year 6 as measured by the bespoke writing measure as a result of the programme, compared with children in the control condition. No statistically significant effect was found for children in receipt of FSM in the intervention schools. Similarly, no effects were found for the secondary outcomes of KS2 writing attainment and KS2 reading attainment. A small, negative effect was found for the KS2 GPS assessment.


There was some evidence that students' prior attainment influenced the impact of Grammar for Writing. Pupils who performed equal to or above the sample median in the pre-test were not observed to have benefitted from the intervention. However, for the lower performing pupils a potential small negative effect was found (ES=-0.11; equivalent to 2 months' progress), although this result was not statistically significant. Evidence from the process evaluation suggests that teachers felt that the resources were not suitably differentiated for all students

Whilst teachers in the intervention schools reported high levels of satisfaction with the programme, the process evaluation indicates that there were high levels of adaptation of the programme by teachers which may have impacted on fidelity and dosage of the programme as delivered. In particular, adherence to two central tenets of the programme, 'pupil discussion surrounding decisions and choice-making' and 'connections made between grammar and effect', was found to be low. The process evaluation suggests that teachers would welcome more guidance on acceptable adaptations to the programme and differentiation within the classroom.

Cost

Grammar for Writing as delivered in this evaluation cost approximately £28.70 per pupil in the first year, and these costs covered 4 days of training, lesson plans and resources for delivery. Additional costs to schools of delivering the programme in subsequent years were minimal. The cost per pupil per year over 3 years is estimated to be approximately £10. This figure assumes approximately 25 pupils per class.

Summary of impact on primary outcome

Outcome/ Group	Effect size (95% confidence Interval)	Estimated months' progress	EEF security rating	No. of pupils	EEF cost rating
Writing	-0.02 (-0.08, 0.03)	0		5182	££££££
Writing EverFSM pupils	0.05 (-0.03, 0.13)	1	N/A	2362	££££££

Introduction

Background evidence

The concept of improving children's grammar in parallel with their writing by using a contextual approach is a promising idea. A developer-led RCT of the Grammar for Writing programme in secondary schools was conducted with 744 Year 8 pupils in 31 schools. This study found statistically significant positive results in favour of the intervention, with pupils with higher prior attainment benefitting the most (Jones et al, 2012; Myhill et al, 2012). It also found that teachers' grammar subject knowledge was a mediating factor in influencing student outcomes (Myhill et al., 2012; Myhill et al., 2013). According to the authors 'The study represents the first large-scale study in any country of the benefits or otherwise of teaching grammar within a purposeful context in writing' (Myhill et al., 2012, p161). However, some weaknesses in the overall study design have since been highlighted, namely intention-to-treat analysis was not used, and analysis was conducted at the pupil-level rather than school-level (Wyse & Torgerson, 2017).

The above trial also focused on secondary school pupils whereas the focus on grammar within writing is currently embedded in the primary phase of schooling. The teaching of grammar as an aid to improve writing skills in primary schools is explicit in the national curriculum and since 2011 this knowledge has been formally tested through the introduction of the Grammar, Punctuation and Spelling (GPS) test at the end of KS2.¹

A version of the Grammar for Writing programme adapted for use at the end of the primary phase of schooling (Year 6) was the subject of a large-scale RCT funded by the EEF (Torgerson et al, 2014). Conducted with 2,510 pupils in 53 schools, this efficacy trial looked at whole-class and small group delivery in a 4-week version of the programme. Using a within-school design, one class within each school was allocated to receive the intervention and one to the control condition. Within each intervention class children expected to receive between a Level 3c and a Level 4b in the KS2 assessments were further randomised to receive whole-class teaching only or whole-class teaching plus small-group teaching. The study found only limited effects measured by children's performance on the GL Progress in English assessment with a small and statistically non-significant effect found for the whole-class intervention ($ES = 0.10$). The impact for those additionally taught Grammar for Writing in small groups (i.e. in addition to receiving whole class teaching of Grammar for Writing) was higher, although it was not much higher than those taught in larger groups (i.e. whole-class teaching) in either the control or intervention conditions ($ES = 0.24$) (Torgerson et al., 2014). Consequently, this difference could have been as a result of teaching children in small groups *per se* rather than as a result of the programme (Torgerson et al, 2014, p.33). In addition, this trial was focused on children during the transition phase of primary education, so although the results suggested only a small effect of Grammar for Writing instruction on writing outcomes it was felt that the timing of the delivery 'could have led to an underestimation of the teaching effectiveness' (Wyse & Torgerson, 2017, p.24), as could the short period of delivery. Consequently, the programme lacks conclusive evidence at the primary phase.

This second EEF-funded RCT of the Grammar for Writing programme is an effectiveness trial, testing a scalable model of the programme under everyday conditions. It focuses on whole-class delivery of two units of Grammar for Writing during Year 6. It partially addresses the limitations of the previous trial (which focused on the transition period after KS2 assessments in Year 6 prior to commencing secondary school in Year 7) by delivering a longer version of the programme (6 weeks as opposed to 4 weeks previously), delivered over a longer time period prior to KS2 assessments (two units to be delivered in the Spring and Summer terms. In addition, while the same number of CPD training days were offered to intervention schools, this was delivered at intervals over a whole academic year), thereby providing the opportunity for the programme, and any potential benefits, to become embedded

¹ Otherwise referred to as the Spelling, Punctuation and Grammar (SPAG) test.

within schools. However, limitations still remained. First, by also evaluating the programme with Year 6 pupils issues arose relating to the teaching focus on KS2 SATs and their administration during that year. Second, the primary outcome measure was changed during the course of the trial, from using the national KS2 SATs writing assessments to administering past-KS2 writing assessments under test conditions within schools after the current KS2 SATs writing assessment was deemed unsuitable (further details of this change are provided in the Outcomes Measures section below). This meant that the primary outcome analysis, whilst designed to be Intention to Treat, only included those schools that did not withdraw from the programme or the trial.

Intervention

The 'Grammar for Writing' programme draws on the concept of improving children's grammar in parallel with their writing by using a contextual approach. Drawing on a theorised understanding of grammar as a meaning-making resource for writing development it is a way of teaching writing that assumes that rather than teaching grammatical rules in the abstract, teachers should help pupils to understand how linguistic structures convey meaning (Jones et al., 2013). Consequently, the programme aims to improve writing by developing pupils' understanding of grammatical choices. Underpinned by key pedagogical principles, Grammar for Writing is embedded in the context of teaching about writing genres. Using authentic texts, the core elements of the programme are:

- linking grammar with creating different effects in writing;
- using examples to show choices in writing rather than lengthy explanations; and
- using high quality talk to develop discussion about grammatical choices and effects.

This approach is supported through the materials provided by the programme, with the emphasis on authentic texts and examples demonstrating choices in writing, with additional scaffolding, particularly in linking grammar with creating different effects in writing and using high quality talk, provided through the CPD. High quality talk is defined as that which is used to develop and build on pupil learning, including using pupils' prior knowledge in order to support their progress and, in this context, enable them to consider and discuss their writing choices.

The programme is designed as a series of units with each unit focusing on a different genre of writing (e.g. narrative writing, persuasive writing etc.). The units can be delivered by teachers as standalone units of work in order to focus on a particular genre or as a series of units, connected through the core principles of the Grammar for Writing teaching approach as described above. Aspects of grammar are embedded within each unit in the context of real-world texts and the programme is designed to encourage pupils to make connections between particular linguistic features and the effect it has on writing. Consequently, the aim is to encourage pupils to make choices in their writing based on an increased knowledge of the range of linguistic features available to them and therefore improve the overall quality of their writing output. For the purposes of this study two units of work were delivered within a whole class setting to Year 6 pupils. Teachers were supported in this delivery with three days of Continuing Professional Development (CPD), followed by a fourth day of CPD at the end of the programme designed to support teachers in planning for Year 6 writing using Grammar for Writing programme principles in the future. The provision of lesson plans and materials was designed to ensure that teacher planning and preparation time was kept to a minimum (and focused on adaptations to the needs of their own class).

The CPD was delivered by members of the Development Team, University of Exeter and Babcock LDP. Babcock LDP is an education support and improvement service which provides training within the school sector. They were chosen to partner with the University of Exeter due to their existing links with the Development Team which ensured that they were already aware of, and closely aligned to, the Grammar for Writing approach. The University of Exeter and Babcock LDP held two team training days with the aim of co-creating the teacher CPD and familiarising the training team with the teaching

materials. The Grammar for Writing training provided for intervention schools was co-delivered by University of Exeter and Babcock LDP with a member of each team co-delivering each CPD day. The book *No Nonsense Grammar* (Babcock LDP, 2016), produced by Babcock LDP, which is designed to align with the National Curriculum and draws on the work of the Exeter team was used as the basis for the teacher support materials.

The first unit of work focused on narrative writing: 'Merlin, King Arthur and the Knights of the Round Table: A land of myth, a time of magic'. This consisted of 16 lesson plans designed to be delivered for one hour daily, four days a week, over a four-week period. The focus was on developing pupils' awareness of the need to shape fictional narratives through the use of visual and authentic written text. It was intended that this unit of work would be delivered in the Spring Term of 2017. The second unit of work concerned persuasive writing: 'Food Waste: Can You Change the World?' This consisted of seven lesson plans designed to be delivered for one hour daily, over a two-week period (four lessons in Week 1 and three lessons in Week 2). The focus was on developing pupils' awareness of the need to shape a persuasive text using joint composition and collaborative revision. It was intended that this unit of work would be delivered in the Summer Term 2017. Each unit of work was designed to result in an individual piece of written work being produced by the pupils.

The first day of CPD was designed to introduce the teachers to the principles of the programme and address grammatical needs in Year 6. This was delivered in October 2016. The second day of CPD was delivered in November 2016 and focused on delivering the first unit of work. Training for the second unit of work was provided in the third CPD day in March 2017. The fourth day of CPD, designed to provide teachers with the skills and knowledge to apply the programme principles and techniques to their teaching of writing using authentic texts in the following school year, was offered to schools in May 2017, when programme delivery was expected to be completed. CPD delivery was provided regionally to facilitate attendance by schools. There were four regional training hubs with each CPD day delivered once in each area: Newcastle, London, Leeds and Devon. In addition to the CPD, teachers were provided with login details to a secure website which provided the training materials, lesson plans, and access to associated resources including video clips, class handout materials and lesson PowerPoint slides.

Further details of the programme as evaluated are provided in Table 1.

Table 1: Template for Intervention Description and Replication (TIDieR)²

Aspect of TIDieR	Exemplification relating to the evaluation
Brief name	Grammar for Writing.
Why: Rationale, theory and/or goal of essential elements of the intervention	A previous efficacy trial of the programme showed evidence of promise with small (but non-significant) effect sizes for whole group delivery in Year 6 (Torgerson et al., 2014). This evaluation consists of an effectiveness trial of the programme delivered on a larger-scale and over a longer time-period.
Who: Recipients of the intervention	Year 6 pupils.

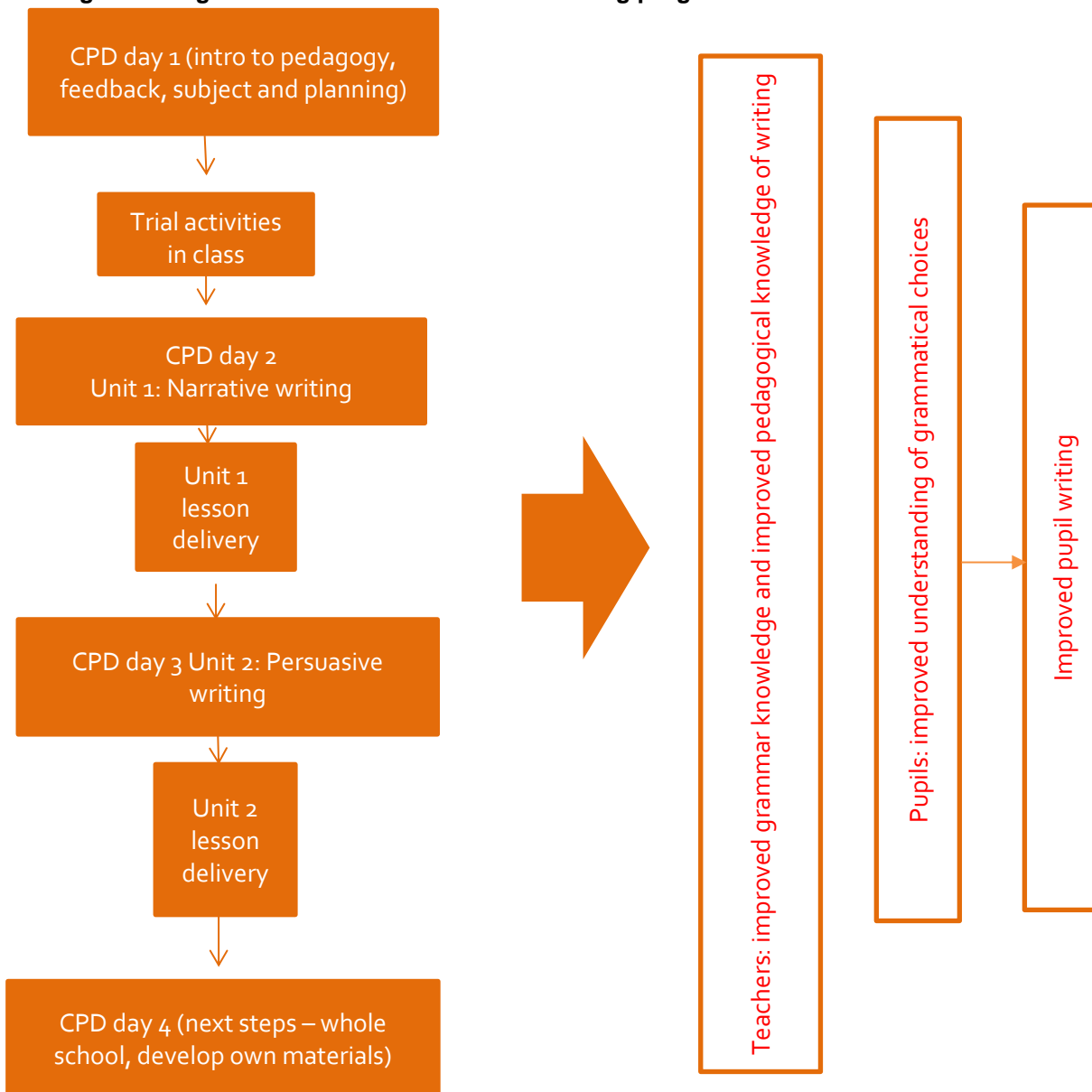
² Adapted from Hoffman et al., 2014.

What: Physical or informational materials used in the intervention	<p>The following are provided for each school:</p> <ul style="list-style-type: none"> • Four days of CPD training provided to all Year 6 teachers in intervention schools. • Detailed lesson plans. • Additional materials to support delivery including paper-based and on-line video resources <p>The lesson plans and resources were provided via a password-protected website during the evaluation year.</p>
What: Procedures, activities and/or processes used in the intervention	<p>All Year 6 teachers in intervention schools trained in the use of grammar and in programme delivery over 3 days of CPD, with a 4th day scheduled to train teachers in continuing the approach in the future. Year 6 teachers deliver lessons as provided in lesson plans in a whole class setting using additional materials provided.</p>
Who: Intervention providers/ Implementers	<p>CPD provided by University of Exeter and Babcock LDP. Programme to be delivered to Year 6 pupils by their regular classroom teachers.</p>
How: Mode of delivery	<p>Delivery of Grammar for Writing units in a whole class setting in place of regular literacy/writing lessons.</p>
Where: Location of the intervention	<p>Whole class setting.</p>
When and how much: Duration and dosage of the intervention	<p>2 units of work scheduled for delivery.</p> <p>Unit 1: Narrative Writing. 16 lesson plans to be delivered for 1 hour a week, 4 days a week over a 4 weeks period. To be delivered in Spring Term 2017.</p> <p>Unit 2: Persuasive Writing. 7 lesson plans to be delivered for 1 hour a day over a 2-week period. To be delivered in Summer Term 2017.</p> <p>For compliance all lessons in each unit must be delivered.</p>
Tailoring: Adaptation of the intervention	<p>The lesson plans provide scope for adaptation and personalisation to individual pupils. Support and challenge sections are provided in each lesson plan to allow for differentiation. However, teachers were instructed to adhere to the grammar points and terminology addressed in each unit and to ensure that the individual final writing took place. For Unit 2, the peer composition and revision elements were also to be included for compliance.</p>
How well (planned): Strategies to maximise effective implementation	<p>In order to maximise the effectiveness of the implementation the following strategies were adopted:</p> <ul style="list-style-type: none"> • Teachers able to book sessions in alternative locations if unable to attend scheduled CPD in own region. • Resources, including lesson plans available on-line to intervention schools.

The lesson plans and resources are available for free (at the time of writing) at: <http://socialsciences.exeter.ac.uk/education/research/centres/centreforresearchinwriting/grammar-teacher-resources/samplelessonplansandschemes/>. See Year 5/6 narrative writing and persuasive writing schemes of work.

A theory of change for the programme as a whole was developed by the Development Team in conjunction with the Evaluation Team and the EEF. This was subsequently adapted into a Logic Model for the evaluation of the programme as delivered in this study.³ This is presented in Figure 1 below detailing programme processes (on the left) and anticipated consequences of those processes (on the right).

³ For the difference between a Theory of Change and a Logic Model see: <http://whatworks.org.nz/frameworks-approaches/logic-model/>

Figure 1: Logic Model for the Grammar for Writing programme

Evaluation objectives

The study was composed of an impact evaluation and a process evaluation, as detailed in the evaluation protocol.⁴ The impact evaluation was designed to assess the effectiveness of the Grammar for Writing programme. The primary research question was:

- How effective is Grammar for Writing in improving the writing skills in Year 6 pupils?

A secondary research question asked whether or not Grammar for Writing impacted on other literacy outcomes (i.e. reading, grammar, punctuation and spelling) for Year 6 pupils. Finally, given that the Grammar for Writing programme aims to increase teachers' grammar knowledge and subsequently

⁴

https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Protocols/Regrant_-_Grammar_for_Writing_effectiveness.pdf.

increase pupils' literacy outcomes, a mediation hypothesis relating to the impact of Grammar for Writing on teacher grammar subject knowledge and this teacher knowledge on student outcomes was tested.

The process evaluation aimed to:

- examine more closely the relationship between the level of implementation of the intervention and its impact on pupil outcomes;
- explain variability in implementation, including understanding the context of the implementation and social processes within schools; and
- address possible barriers to implementation.

Ethics and trial registration

Ethical approval was granted by the Education Ethics Committee, University of York (Ref: 16/18) and the Social Sciences and International Studies Ethics Committee, University of Exeter (STF/16/17/11) in April 2016 prior to school recruitment.

Informed ('opt-in') consent was obtained at the school level from the headteacher and from Year 6 teachers. Information sheets were sent home to parents of Year 6 pupils, including 'opt-out' consent to enable parents to withdraw their child's data from the evaluation if they wished to do so. Consent included linking to the National Pupil Database and data archiving.

The primary outcome was changed during the evaluation period (see Methods Section below). At this point headteachers were requested to provide consent for the new measure to be administered. This was sent via email and obtained digitally. Supplementary information sheets were also sent to parents with additional opt-out from this data being used in the evaluation. Ethical approval for this was obtained from the Education Ethics Committee, University of York and the Social Sciences and International Studies Ethics Committee, University of Exeter in January 2017.

The trial is registered with the ISRCTN registry (ref: ISRCTN83236864).

Data protection

All data was stored and processed in accordance with the Data Protection Act (1998).

Schools were informed of the data requirements through the Memorandum of Understanding and all parents/carers of pupils in the trial classes received an information sheet that outlined the data schools were providing about the pupils in the trial and how it would be used. Parents/carers were given the option to withdraw their child from data sharing. All consents and information sheets are included in Appendix C.

Schools provided pupil details (name, unique pupil number (UPN) and date of birth) for all pupils in the trial class(es) at baseline to allow the Evaluation Team to request KS1/KS2 results and FSM status for these pupils from the National Pupil Database. Access to pupil details was limited to members of the Evaluation and Project Teams. The NPD data was used for statistical analysis and will be shared with the Department for Education, the EEF, FFT Education and in an anonymised form to the UK Data Archive.

All results have been anonymised so that no school or individual pupil should be identifiable in the report or any dissemination of the results.

Project team

The independent Evaluation Team was led by Dr Louise Tracey (University of York) and included Dr Jan R. Boehnke (University of Dundee), Mrs Louise Elliott (University of York) and Dr Claudine Bowyer-

Crane (University of York). Kate Thorley, Sarah Ellison, Imogen Fountain, Mary Robison, Madeline Crossthwaite and Niamh Robinson provided research support at various stages of the study. Dr Pam Hanley (University of York) was involved in writing the original project proposal. The Evaluation Team was responsible for the conduct of the research, including the randomisation, data collection, analysis and reporting of the study.

The Development Team was based at the University of Exeter and led by Professor Debra Myhill, Director of the Centre for Research in Writing, Graduate School of Education. The team included Dr Susan Jones, Dr Helen Lines, Ms Sara Venner, and Ms Marijke Shakespeare. The Project Team was responsible for school recruitment, intervention development, training and delivery of the programme.

The primary outcome measure was administered by the National Foundation for Educational Research (NFER). The team was led by Kathryn Hurd, Head of Survey Operations and included Guvi Chohan, Research Manager, and Priscilla Antwi, Researcher.

Methods

Trial design

The evaluation was a two-arm effectiveness RCT with randomisation occurring at the school-level to reduce the possibilities of diffusion, which could occur with an in-school design.

Table 2: Grammar for Writing Trial Design

Trial type and number of arms		Two-arm effectiveness
Unit of randomization		School-level
Stratification variable(s) (if applicable)		Region (North-East/Not North-East)
Primary outcome	Variable	Writing
	Measure (instrument, scale)	Total Score (excluding handwriting) (Bespoke Writing Assessment based on past KS2 tests, 0-40)
Secondary outcome(s)	Variable(s)	Reading Writing Grammar, Punctuation and Spelling
	Measure(s) (instrument, scale)	Raw Score (KS2 Reading assessment, 0-50) Level (KS Writing assessment, 1-7) Grammar, Punctuation and Spelling (KS2 Grammar, Punctuation and Spelling (GPS) assessment, 0-70)

Intervention schools received the Grammar for Writing programme and associated materials and training in exchange for paying a reduced rate of £500 per school. All Year 6 teachers in intervention schools were expected to attend the four days of CPD and to deliver the two units of work to their Year 6 classes within their usual scheduled literacy lessons.

Control schools were expected to continue their Year 6 'teaching as usual' and received £500 on completion of all requested measures, at the end of the intervention period. This could then be used towards funding Grammar for Writing training for the academic year 2017-2018 if desired. This incentive was chosen as the burden placed on schools to participate in the programme was not originally considered high (although this subsequently increased; see Outcomes section below). Providing control schools with the option of whether or not to take up the training at the end of the evaluation was also felt to avoid potential ethical issues if the intervention was not shown to be effective.

Teachers in both control and intervention conditions also received an extra payment of £20 in on-line Amazon vouchers in exchange for completing the pre- and post-intervention on-line surveys (see Process Evaluation section).

No changes to the protocol were made after acceptance of an amended protocol in the May 2017. The protocol was amended at this time to account for a change in the primary outcome measure for the evaluation (as detailed below).

Participant selection

The target population for this study was state primary schools in England. Eligible schools were those that had not (i) taken part in the previous EEF Grammar for Writing trial or (ii) implemented the programme previously. Although they did not have to be two-form entry, very small schools (fewer than 20 Year 6 pupils) were kept to a minimum by deliberately targeting larger schools for recruitment. It was also aimed to include a high proportion of disadvantaged schools in the trial (aiming for an average of 29% of pupils identified as EverFSM in the National Pupil Database across the sample as a whole). Half of the schools were recruited from the North East⁵ and the other half from across the rest of England.

Recruitment was conducted by the Development Team (University of Exeter), with support from the Evaluation Team (University of York). For pragmatic reasons (i.e. ease of organizing training sessions) specific regions were targeted in addition to the North-East: the North West, South West and London. A primarily dual approach was then adopted. First, all schools in the local authorities in those target areas were systematically identified and approached. Second, existing relationships were used and new relationships developed with key stakeholders in these areas, including literacy consultants, local authority leads, research connections and the National Association for the Teaching of English. These two approaches were supplemented by a number of untargeted, opportunistic approaches made using social media (e.g. Twitter and Facebook).

Consent was obtained initially from headteachers for participation in the study via a Memorandum of Understanding (MOU). Subsequently, teacher opt-in consent was also collected to participate in the trial. A request to complete the teacher survey was sent to teachers after consent was obtained and a request for Year 6 pupils' UPNs and associated teacher name was also made. Parental information sheets (including opt-out consent) were also sent to be distributed via the school. MOUs, consent forms and parental information sheets are included in Appendix C.

Schools were only regarded as fully consented and therefore eligible for randomisation after:

- The Head Teacher had signed a Memorandum of Understanding;
- All Year 6 teachers had consented to participate in the trial;
- Class lists with pupil UPNs had been provided; and
- Teachers had completed the pre-intervention on-line survey.

Outcome measures

The original protocol for this study stated that the national writing assessments for KS2 would be the primary outcome for the evaluation. However, in 2016 there were significant changes to how KS2 writing was assessed and the implications for the suitability of these tests was still uncertain. Consequently, provision was made in the protocol for the evaluators in conjunction with the delivery team and EEF to review the suitability of using the KS2 writing assessments as an outcome measure in November 2016 and update the protocol in the event of any changes. As a result of this review it was felt that the KS2 writing assessment results were unsuitable as a primary outcome measure. This was because the 2017 writing assessment for KS2 consisted of a portfolio of teacher-assessed work which, whilst externally moderated, was judged to be 'working toward', 'working at' or 'working above' the expected standard for the end of Key Stage 2. Although not aligned to the current national curriculum, past KS2 writing papers were agreed by all stakeholders to present an alternative that was covering relevant content, less likely to be inherent to the treatment, could be marked by independent and blinded markers, and

⁵ That is, Local Authorities in the former Government Office Region 1: Darlington, Durham, Hartlepool, Gateshead, Middleborough, Newcastle upon Tyne, North Tyneside, Northumberland, Redcar and Cleveland, South Tyneside, Stockton-on-Tees and Sunderland.
(<http://webarchive.nationalarchives.gov.uk/20080728115009/http://www.dcsf.gov.uk/rsgateway/region1.shtml>)

provided a sufficiently differentiated marking scheme. The 2017 KS2 writing assessments were kept as secondary outcomes to provide effect estimates for the current national curriculum.

Consequently, the primary outcome for the evaluation was the combined results of two tasks (writing prompts) selected from past Key Stage 2 writing assessments which were in use pre-2013. Prior to 2013 KS2 writing was assessed through a national, externally marked written assessment. Since 2013 KS2 writing has been teacher-assessed. The advantage of using sample past KS2 writing tasks was that they could be administered in controlled conditions within schools and have a set marking scheme, which is sufficiently differentiated to be able to conduct a meaningful and sufficiently robust analysis to assess the impact of the programme on KS2 writing.

The tasks were selected by the Evaluation Team to include one shorter written task and one longer written task. The shorter task was that set in the KS2 assessments in 2011 and was a prompt for a piece of persuasive writing. The longer task, covering narrative writing, was taken from the 2003 assessments.⁶

As narrative and persuasive writing are implicit within the KS2 curriculum (i.e. pupils are expected to 'write effectively for a range of purposes and audiences', Standards & Testing Agency, 2017), this primary measure was not felt to be inherent to treatment – as teachers in the control condition would also be teaching their pupils to write narratively and persuasively. However, at the same time, as these two genres are covered by the intervention, it could be expected that this primary measure would be able to detect the presence of any effects as a result of the Grammar for Writing programme.

The Development Team remained blind to the exact content of these past KS2 papers as used for the primary outcome. The assessments were administered in schools at the whole class level under assessment conditions. They were conducted independently by the National Foundation for Educational Research (NFER) in June 2017, after the KS2 assessments took place. This was to ensure controlled conditions and to reduce any burden on schools.

The assessments were marked by a team of experienced assessors, at the University of York. All assessors received training from the Evaluation Team. The marking followed the published guidelines for the assessments focusing on three assessment foci: (1) 'sentence structure and punctuation'; (2) 'text structure and organisation'; and (3) 'composition and effect'. The result of the task was scored between 0 (which meant that none of the criteria for the lowest scoring band had been met for the three assessment focuses) and 40 (which meant that all the criteria for each of the three assessment criteria had been met to a high standard). The fourth assessment focus, 'handwriting' for which 0-3 marks could be obtained was not included in the outcome scoring as this was neither a focus of the programme, nor the focus of the primary research question.

Papers were single marked by markers blind to condition. To ensure consistency a small selection of papers were marked as a group, and individual marker's work was periodically checked for consistency by the Assessment Co-Ordinator. Once all papers were marked a randomly sampled 5% check within each school was carried out by 2 of the original markers. These moderators did not moderate their own papers. The original and moderated marks were compared to ensure consistency of marking. An inter-rater reliability check was conducted and a level of agreement at less than 80% within school was deemed to be inconsistent and all the papers for that school were subsequently remarked. Consequently, assessment papers for 32 (24%) schools were remarked. The overall inter-rater reliability for the 5% check in the remaining 103 schools was 92%.

It is also important to assess whether any improvement in these aspects of writing have been at the expense of other elements of literacy, maybe as a result of reduced focus on these. For this reason, secondary outcomes included KS2 scores on each element of literacy assessed in KS2 SATSs (writing;

⁶ The prompts and the associated marking schemes were obtained from: <http://primarytools.co.uk/pages/pastpapers.html>.

reading; grammar, punctuation and spelling). The Key Stage raw scores were used for reading and grammar, punctuation and spelling. The reading assessment is scored from 0 to 50 and the grammar, punctuation and spelling assessment is scored from 0-70. These assessments are marked externally to the school. The KS2 writing results, as they are teacher assessed from a portfolio of student's written work, are graded 'working towards the expected standard for most 11-year olds', 'working at the expected standard for most 11-year olds' and 'working at greater depth at the expected standard for most 11-year olds'. Despite reservations for use as the primary outcome (see above) the KS2 SATs writing assessments were deemed sufficiently externally moderated to be included as a secondary outcome,

These KS2 results (variables WRITTAOUTCOME, KS2_READMRK and KS2_GPSMRK) were obtained from the National Pupil Database as using primarily nationally collected data minimised the cost and the burden on schools and pupils. These measures are high in contextual validity and, since they constitute the main indicators of school and student academic performance, all teachers (intervention and control) would be focused on ensuring that pupils succeeded in them. With the addition of the past KS2 writing tasks, the outcome measures aimed to provide a measure of all-round performance on literacy, and, specifically, any indirect effects of the intervention. The protocol was updated in May 2017 to reflect the change in the primary outcome measure and the inclusion of the 2017 KS2 writing assessment as a secondary outcome measure (Tracey et al., 2017). The intervention training and intended programme delivery by schools was designed so that both units of the programme would be delivered prior to the primary outcome assessments and KS2 assessments were administered in schools (see Timeline below).

The pre-test measure was the Key Stage 1 writing results (obtained from the National Pupil Database). Historically, KS1 English results were highly correlated with the previous KS2 assessments in English ($r = 0.73$) and we assumed that this remains high using the KS2 and KS1 writing measures proposed (EEF, 2013).

An intermediary measure was a teacher 'grammar quiz' included in the pre- and post-test teacher survey (see Process Evaluation below). The pre-test 'quiz' was developed by the Exeter team for use as part of the 'Grammar for Writing' training to assess teacher grammar knowledge prior to training and as a test for the evaluation of their training sessions. It is used routinely in Grammar for Writing initial training sessions as a diagnostic test for teachers and to raise issues for discussion about grammar knowledge. Consequently, it was not designed as a validated instrument and its psychometric properties have not previously been tested. The team made it available for testing during the trial after it was agreed by all stakeholders that it would be the most appropriate assessment tool for this trial. As the pre-test was taken prior to allocation, teachers' and researchers were blind to allocation. A similar quiz was developed for the post-intervention survey specifically for this trial by the Evaluation Team at the University of York to ensure that it was not 'inherent to treatment'. The initial, pre-test quiz used an extract from a real text and requested teachers answer a number of multiple choice and free response questions relating to aspects of grammar included therein. The post-test quiz mirrored the pre-test, as far as possible, using a different text. This was piloted within the Evaluation Team and with two independent primary school teachers to determine that it was accessible to teachers, provided some differentiation, and that there was no ceiling effect. The pre-test was scored from 0-30 and the post-test was scored from 0-29 with 1 point awarded for each correct answer (and 0 awarded for each incorrect or incomplete answer) in both tests.

The pre- and post-test teacher surveys were delivered online using Qualtrics (Qualtrics, Provo, UT), which was programmed to code the answers digitally as it was being completed. The surveys were then downloaded into the SPSS software program. Using SPSS, the following procedure was followed:

Where Qualtrics had not already coded answers into correct (1) or incorrect (0) scores (e.g. because a free text response was requested as opposed to a multiple-choice response) values assigned within SPSS were recoded into correct (1) or incorrect (0) following the quiz writers' marking scheme;

- These recodes were cross-tabulated with the original qualtrics codes to check for accuracy;
- The recoded answers (correct and incorrect) were then summed to create an overall score.

The pre-test quiz could be scored between 0 and 30, and the post-test quiz from 0 to 29. The data was then exported into an Excel database for merging with the overall dataset.

Finally, an implementation fidelity measure was devised for the sub-group analysis. This is discussed in further detail in the Process Evaluation section below.

Sample size

The statistical power of the proposed analyses was estimated using the formula provided as a standard by the EEF (EEF, 2013) and updated when the statistical analysis plan was finalised (the protocol mentions minimally different estimates):

$$MDES = M_{J-k} \sqrt{\frac{\rho(1-R_2^2)}{P(1-P)J} + \frac{(1-\rho)(1-R_1^2)}{P(1-P)nJ}}$$

The following assumptions were made⁷:

- Pupils per school per class: 25 (i.e. with two classes per school $n = 50$ per treatment (intervention/control) per school)
- Student-level pre-post correlation (squared): $R_1^2 = 0.53$ ($r = 0.73$)
- Intraclass correlation: $\rho = 0.15$
- Criterion for statistical significance: $p < .05$ and $\beta = 0.80$ (therefore $M_{J-k} = 2.85$)

Consequently, a sample of 150 schools would result in a $MDES = 0.18$. As we would expect stratification variables to explain some of the variance (Explained variance between schools $R_2^2 = 0.10$), the analytic model proposed below could identify a $MDES = 0.17$.

Assuming 16 pupils in receipt of Free School Meals (FSM) per school, this sample of 150 schools would enable an effect size of $MDES = 0.18$ (with stratification $MDES = 0.17$) to be detected in the FSM sub-sample (defined by NPD EVERFSM_ALL_SPR17). This assumption was considered reasonable given the school recruitment strategy (i.e. the targeting of schools for recruitment with above average proportions of pupils in receipt of FSM).

Consequently, the recruitment target was 150 schools. Given the large number of schools required for the study a larger number of schools than 150 was not specified to account for attrition. Instead, it was agreed to take the pragmatic approach that some additional schools could be included once the 150 specified schools were recruited if it was deemed feasible by the Development Team and EEF to do so. In the event, 155 schools were recruited.

Randomisation

Randomisation was conducted at the school level using minimisation. Minimisation uses algorithms to ensure balance at baseline on key observables between intervention and control conditions (i.e. to minimise differences) and permits ongoing allocation so schools know which condition they have been assigned to soon after recruitment. For this study, schools were stratified by region (to ensure balance

⁷ For further details see Education Endowment Foundation (EEF) (2013). *Pre-testing in EEF evaluations and the Statistical Analysis Plan*: (https://educationendowmentfoundation.org.uk/public/files/Projects/Grammar_for_Writing_SAP.pdf)

in the number of schools allocated to each condition within region) with the two stratification variables being North East and not-North East. Randomisation was conducted and recorded by a member of the Evaluation Team not involved in the recruitment process (Louise Elliott, University of York Evaluation Team) using MinimPy software (MinimPy, 2013; Saghaei & Saghaei, 2011; v3.0; default settings). Randomisation was conducted in 6 batches between July and October 2016 due to the extended recruitment period to allow schools to be informed of their allocation as soon as possible after completion of the pre-randomisation tasks and hence enable them to plan more effectively for the school year.⁸

In total there were 77 schools allocated to treatment, and 78 schools allocated to the control condition.

Statistical analysis

The approach to the statistical analysis largely followed that described in the Statistical Analysis Plan.⁹

A description of the approach, including some minor deviations from the planned analysis is provided below. A more detailed description of the analysis approaches taken are provided in Appendix D of this report.

Primary intention-to-treat (ITT) analysis

The impact evaluation used a mixed effects model in which pupils were nested within schools. This made it possible to separate within-school variation in the outcome from between-school variation. As stated in the protocol the analysis was planned as intention-to-treat, which means that schools were analysed according to the condition they were allocated (control or Grammar for Writing), not that which they actually received. As described in detail below, this was not possible due to the pattern of school-drop out and only the secondary outcome analyses are fully intention-to-treat.

This study was planned for a single primary outcome, the writing assessment developed by the team from previously used Key Stage 2 assessments ($KS2_{past}$). In accordance with the power analysis, pre-test data from the Key Stage 1 (KS1) writing results were used as a student-level covariate ($KS1$) without random variation across schools (assuming the same relationship between pre-test and outcome, γ_{10}). An individual student i 's $KS2_{past}$ result in a specific school was modelled as depending on school j 's average $KS2_{past}$ attainment (random school-level intercept; μ_{0j}) and a random error term (ε_{ij}). Each school's average $KS2_{past}$ performance (μ_{0j}) was predicted by an overall intercept (average performance; γ_{00}); in which recruitment region the school was located was used as a stratification variable (North East/ not-North East; REG); and the intervention to which the school was randomised (GfW). This model is summarised in formulas (1)-(3).

$$KS2_{past_{ij}} = \mu_{0j} + \mu_{1j}KS1_{ij} + \varepsilon_{ij} \quad (1)$$

$$\text{with } \varepsilon_{ij} \sim N(0, \sigma^2)$$

$$\mu_{0j} = \gamma_{00} + \gamma_{01}REG_{0j} + \gamma_{02}GfW_{0j} + u_{00} \quad (2)$$

⁸ The dates of randomisation (and number of schools) were as follows: 15 July 2016 (36 schools); 17 August 2016 (34 schools); 12 September 2016 (31 schools); 22 September 2016 (32 schools); 30 September 2016 (19 schools); and 5 October 2016 (3 schools).

⁹ See https://educationendowmentfoundation.org.uk/public/files/Projects/Grammar_for_Writing_SAP.pdf

$$\mu_{1j} = \gamma_{10} \quad (3)$$

with $u_{00} \sim N(0, \tau_1^2)$

The analysis was performed in the R environment (R Core Team, 2017; version 3.4.2); specifically the R-package `lme4` (Bates, Mächler, Bolker, & Walker, 2015) was used with the corresponding formula expression in the command `lmer()`:

$$KS2_{past} \sim KS1 + REG + GfW + (1|school)$$

For the Grammar for Writing programme to be considered to show an effect, the average bootstrapped point estimate for the coefficient of the intervention effect (γ_{02}) was expected to be positive (i.e. on average intervention schools achieve higher scores on $KS2_{past}$) and the 95%-bootstrap confidence interval of this coefficient did not include 0.

The analysis was cluster-bootstrapped (Hanley, Böhneke, Slavin, Elliott, & Croudace, 2016; Huang, 2016) as agreed in the Statistical Analysis Plan. From each school a random sample of the same size as its actual sample was drawn with replacement and across these school-wise bootstrap samples, the mixed model was then estimated.¹⁰ This process was repeated $b = 1000$ times and for a 95%-confidence interval the statistical estimates (here the γ_{02} values) were saved and their top and bottom 2.5%-quantiles identified. The average of the bootstrapped values was treated as the point estimate. As stated in the Statistical Analysis Plan (SAP), no p -values are reported for any analysis.

Missing data

The amount of missing data was documented for each variable individually as well as for the patterns of missing values which occur. The results are presented in the Baseline Comparison section below (Table 7). Further details on missing data by variable, patterns of missing data and the relative frequency of pupils with any missing data by school is reported in Appendix D. To evaluate the impact of missing data on the robustness of findings from the ITT analyses of the primary outcome, sensitivity analyses were run to evaluate the robustness of the results where either > 5% missing data for the primary outcome analysis were encountered (i.e. 5% of cases were deleted listwise for the analysis); or if at least one school in the ITT analysis had more than > 15% missing responses for the primary outcome. For the ITT analysis of the primary outcome multiple imputation via the expectation-maximisation (EM) algorithm (King, Honaker, Joseph, & Scheve, 2001) was used to impute missing values.

As with other imputation techniques, EM¹¹ uses the observed relationships between variables to predict missing values, but instead of imputing only one variable at a time, all variables entered into the algorithm are jointly imputed. The algorithm iterates through a number of cycles, each time updating the imputed values for all variables. The R package *Amelia* (Honaker, King, & Blackwell, 2011; King, Honaker, Joseph, & Scheve, 2001) was used for this analytic step and in this specific case, the following variables were entered into the algorithm:

¹⁰ E.g. if there were observations 1,2,3,4,5 in a school, one resample could be [1,2,2,5,4] and another [1,5,1,1,3].

¹¹ The protocol and SAP for the project erroneously mention multiple imputation via chained equations, but the analysis was always planned to be conducted with a Expectation-Maximisation approach as indicated by the referenced software we planned to use.

- Gender, EverFSM, and the KS1 result ("baseline data"; independent of whether or not they had missing data);
- The primary and secondary outcome variables ("follow-up"; which were more likely to have missing data);
- n-1 dummy variables for the schools to approximate the multilevel structure of the data as well as the described analytic approach with school-level intercepts (no missing data, since known for every student); and
- Additionally, two dummy variables which coded whether baseline data was missing (yes/ no) or only follow-up data (yes/ no; see below; no missing data since coded from available missingness patterns; see below).¹²

Interval-scaled variables were modelled with linear regressions and dichotomous variables with logistic link functions. Further, the algorithm was set to run for at least 100 updating cycles per imputed value set. In every of the $b = 1000$ bootstrap samples one imputation was performed and confidence intervals and point estimates from these analyses were then derived from the imputed data (instead of only the observed data as described above; Heymans et al., 2007; Schomaker & Heumann, 2014).

The EM algorithm does not define a specific model for the missingness mechanism, which is why it is not seen as preferable in all cases where details about missingness processes are available (especially in longitudinal studies). However, in cases such as this with very few variables and virtually no information about the specific assessment context it still allows researchers to use all available data. It further builds only on very basic tenets of the missing-at-random assumption, i.e. that conditional on observed variables, data are missing at random. To approximate the most basic of missingness processes we included two dummies which condition predictions of the EM procedure on whether or not any data for a respondent was missing at baseline (i.e. there was incomplete data obtained) or whether any data was missing at follow-up (non-attendance at the primary outcome assessment or secondary outcomes not available from the NPD).

Non-compliance with intervention

In this study only a single on-treatment analysis was performed. In order to measure the impact of compliance, teachers were scored according to the number of CPD training days they attended. This scoring was used for all schools in the intervention group instead of the school's randomisation status (and a figure of "0" entered for all control schools). The analyses for primary and secondary outcomes was re-run based on this score. As stated in the SAP it was planned to allocate schools according to the intervention that was actually delivered. However, aside from drop-out on school level meaning no primary outcome was collected from these schools, no other changes to the allocated interventions occurred.

Secondary outcome analyses

The analyses of the secondary outcomes (see footnote 13) investigated whether or not Grammar for Writing potentially impacts on other literacy outcomes (writing, reading and grammar, punctuation and spelling, as measured by KS2 SATS). The analytic approach was exactly the same procedure and model as for the primary outcome, with the only difference that instead of $KS2_{past}$ the respective secondary outcome was used as the dependent variable. The intervention was evaluated as having shown a potential effect on a secondary outcome when the 95%-bootstrap confidence interval of the coefficient (γ_{02} ; see formula 2 above) did not include 0. This result cannot be used to gauge the efficacy of the intervention and is reported purely for exploratory purposes to evaluate whether there are potential positive or negative spill-over effects on curriculum outcomes which would need further

¹² When imputing the primary outcome data set these variables were used, but due to the small variation in missingness patterns they were highly collinear with their original variables.

research. The same rules for the necessity to impute data as a sensitivity analysis were specified as for the primary outcome analysis.

Additional analyses

The only additional analysis concerns the link between teachers' grammar knowledge and programme impact. Grammar for Writing aimed to increase teachers' grammar knowledge and subsequently increase pupils' literacy outcomes. Consequently, a mediation hypothesis relating to the impact of Grammar for Writing on teacher knowledge and said grammar knowledge on student outcomes was tested.

The measure of teachers' ($N = 312$) grammar knowledge was their performance in the follow-up grammar quiz administered at the end of the intervention. The scores of this quiz are reported (Mean, Median, SD), including Cronbach- α and the pre-post correlation in the control group as reliability estimates. The study was not set up to specifically test the hypothesis whether the intervention had an effect on teachers' grammar knowledge, but we ran two tests on the teacher sample to evaluate whether there potentially had been an effect. Since the number of teachers per school was very small, no hierarchical model was applied here (see SAP). The mediation hypothesis (see below) was tested with the same approach as for the ITT analysis.

Teachers' scores obtained in the post-intervention grammar quiz were compared using a bootstrapped t -test across the two groups. If the bootstrapped 95%-confidence interval of the bootstrapped t -values did not include 0 and the average t -value was positive (indicating higher attainment in the group of teachers who received the intervention), Grammar for Writing would be evaluated as having shown a potential effect on teachers' grammar knowledge.¹³

To gauge the potential for a mediation effect of higher grammar knowledge on the side of the teachers the model used in the analysis of the primary outcome was extended by incorporating the teacher's grammar quiz performance (GQ) as a predictor on student level (for all other variables compare formulae 1-3 above).

¹³ Further analyses on the data was conducted to evaluate the validity of the grammar quiz. This entailed a linear regression model regressing the post-scores on pre-scores including an interaction effect with the intervention group to evaluate whether the intervention led to differential gains in grammar knowledge. Principal component analyses were conducted to gauge the plausibility of both quizzes representing the same trait. These analyses were post-hoc evaluations of how well the measure performed and are included in Appendix D.

$$KS2past_{ij} = \mu_{0j} + \mu_{1j}KS1_{ij} + \mu_{2j}GfW_{ij} + \varepsilon_{ij} \quad (4)$$

with $\varepsilon_{ij} \sim N(0, \sigma^2)$

$$\mu_{0j} = \gamma_{00} + \gamma_{01}REG_{0j} + \gamma_{02}GfW_{0j} + u_{00} \quad (5)$$

with $u_{00} \sim N(0, \tau_1^2)$

$$\mu_{1j} = \gamma_{10} \quad (6)$$

$$\mu_{2j} = \gamma_{20} + u_{20} \quad (7)$$

with $u_{20} \sim N(0, \tau_2^2)$

A potential mediation effect would be detected if the bootstrapped 95%-confidence interval of the product of the coefficients μ_{2j} and γ_{20} did not include 0 (details for the test can be found in: Pituch, Murphy, & Tate, 2009). As above, this analysis was purely exploratory and does not estimate the efficacy of the intervention itself.

Subgroup analyses

As specified in the protocol, subgroup analyses were carried out for:

- pupils eligible for FSM;
- boys and girls; and
- high and low achievers on the pre-test (KS1; median-split based on all observed scores).

The multilevel model described for the primary outcome was extended for each variable separately by adding the predictor itself and an interaction term between the intervention variable (*GfW*) and the variable currently analysed. The intervention was evaluated as showing a subgroup effect for the specific variable when the bootstrapped 95%-confidence interval for the coefficient for the interaction term did not include 0. As before, this analysis was purely exploratory and does not estimate the efficacy of the intervention itself.

As previously, an individual student *i*'s $KS2_{past}$ result in a specific school was modelled as depending on school *j*'s average $KS2_{past}$ attainment (random school-level intercept; μ_{0j}), previous attainment ($KS1$), and a random error term (ε_{ij}). For the test for subgroup effects, a coefficient for one of the student-level variables described above was added (*Subgroup*) as a random slope. Each school's average $KS2_{past}$ performance (μ_{0j}) was predicted by an overall intercept (average performance; γ_{00}); each school's level on the stratification variable which controls for geographical region (North East/ not-North East; *REG*); and the intervention to which the school was randomised (*GfW*) with the now added cross-level interaction with one of the sub-grouping variables (*Subgroup*) described above:

$$KS2past_{ij} = \mu_{0j} + \mu_{1j}KS1_{ij} + \mu_{2j}Subgroup_{ij} + \varepsilon_{ij} \quad (8)$$

with $\varepsilon_{ij} \sim N(0, \sigma^2)$

$$\mu_{0j} = \gamma_{00} + \gamma_{01}REG_{0j} + \gamma_{02}GfW_{0j} + u_{00} \quad (9)$$

with $u_{00} \sim N(0, \tau_1^2)$

$$\mu_{1j} = \gamma_{10} \quad (10)$$

$$\mu_{2j} = \gamma_{20} + \gamma_{21}GfW_{0j} + u_{20} \quad (11)$$

with $u_{20} \sim N(0, \tau_2^2)$

The analysis was performed in the R environment (R Core Team, 2016); specifically, the R-package `lme4` (Bates, Mächler, Bolker, & Walker, 2015) was used with the corresponding formula expression in the command `lmer()`:

```
KS2past ~ KS1 + Subgroup + REG + GfW + Subgroup:GfW + (1+Subgroup|School)
```

The intervention was evaluated as having shown a potential interaction with the specified subgroup variable when the 95%-bootstrap confidence interval of (γ_{21} ; formula 11) did not include 0.

Only when this effect was found to be statistically significant was more detailed reporting on subgroup statistics conducted (means, SDs). The exception to this was for pupils in receipt of FSM for which details are routinely reported.

The only subgroup analyses performed were those previously defined in the statistical analysis plan. An additional subgroup analysis was planned to look at high and low implementation fidelity within treatment schools using data obtained from the teacher survey. However, the large number and range of changes recorded by teachers and the difficulty in establishing the extent to which these changes were within the bounds of programme delivery as intended, meant that no fidelity measure was constructed from the teacher survey. Instead a measure of fidelity was constructed from the lesson observations undertaken by the Development and Evaluation Teams. Further details on the teacher survey, the lesson observations and the fidelity measure are given in the reporting of the process evaluation below.

Effect size calculation

Effect sizes were calculated based on the total variance in the models. Two calculations were employed. Hedge's *g* based on pooled variances (EEF, 2018) was used in all places where the effect size based on observed data are required to be reported according to the Education Endowment Foundation template. This is necessary in all places where the two intervention groups are compared descriptively, and this effect size describes how far apart (as measured in standard deviations) the averages of the intervention and control schools are.

For two-level models (see definition of error terms in formulas above) the effect size was determined based on the estimated total variance from the multilevel model (Hedges, 2007: formula 3):

$$ES = \frac{Effect}{\sqrt{\sigma^2 + \tau_1^2}}$$

Here, *Effect* was the coefficient from the estimated model (e.g., γ_{02} in the analysis of the primary outcome; formula 2). For both effect sizes confidence intervals were bootstrapped (see above). This effect size can also be interpreted as how far apart (as measured in standard deviations) the averages of the intervention and control schools are, but it controls for additional variables in the statistical model (see formulae above) as well as for clustering due to schools.

Implementation and process evaluation

The implementation and process evaluation aimed to assess implementation fidelity to the programme, any variation in fidelity and possible barriers to implementation. In addition, it aimed to explore writing instruction in the control condition schools. The process evaluation measures were primarily administered by the Evaluation Team, University of York. The measures, participants and analysis are all described in more detail below.

Teacher surveys

All Year 6 teachers in schools participating in the trial were asked by the Evaluation Team to complete an on-line survey before and after the intervention using Qualtrics (Qualtrics, 2015). The survey was designed to capture the baseline characteristics of schools and participating teachers and capture any changes that occurred during the intervention year. The pre-test survey was completed prior to randomisation and covered teacher background and experience, planned Year 6 teaching (i.e. classes, pupils, writing programmes) for the academic year 2016-2017 and some Likert-style statements relating to grammar teaching (e.g. 'I feel confident teaching grammar to Year 6' with a 5-point scale from 'Strongly Agree' to 'Strongly Disagree'). The post-test survey was administered at the end of the Summer Term 2017 and requested information on actual Year 6 teaching during the academic year 2016-2017, repeated the Likert-style statements and asked about plans regarding the programme the following academic year (2017-2018). For Year 6 teachers in intervention schools some additional questions were asked in the post-test survey relating to their experiences of the programme, including training. A grammar quiz for completion by teachers was also embedded in both the pre-test teacher survey and the post-test teacher survey, as described in the 'Outcome Measures' section above.

Two hundred and ninety-nine teacher surveys were completed at baseline from across all 155 schools involved in the evaluation. After removal of ineligible teachers (those who, in the event, were not going to be teaching Year 6 in the academic year 2016-2017) and duplicate surveys there were 263 teachers (138 intervention, 125 control) included in the baseline analysis. This compares to 312 teachers originally recruited to the study (see Pupil and School Characteristics section below). Two hundred and forty-two (242) surveys were completed at follow-up but only 232 teachers were included in the follow-up survey analysis after the data was cleaned. Sixty-seven teachers were lost between the pre- and post-test survey. The main reasons for this were school withdrawal from the study, teacher turnover, and teacher's changing Year group between administration of the pre-test and the start of the academic year (teachers completing the survey prior to September 2016-only). Emails and reminder phone calls were used to encourage completion of the teacher survey at both time points.

Table 3: Survey respondents by school allocation

	Intervention	Control	Total
	N (%)	N (%)	N
Number of teachers at baseline	168 (54)	144 (46)	312
Number of baseline surveys completed	163 (55)	136 (45)	299
Total number of eligible baseline teacher surveys	138 (52)	125 (48)	263
Number of follow-up surveys completed	130 (54)	112 (46)	242
Total number of eligible follow-up teacher surveys	125 (54)	107 (46)	232

School visits

A more detailed process evaluation was planned with a subsample of 15 schools (10 intervention and 5 control schools) in order to understand the intervention more fully as it was implemented in schools. Schools were randomly selected to participate in the subsample. This involved a lesson observation followed by an interview with a Year 6 teacher. The literacy coordinator was also interviewed in this subsample of schools where this was a different member of staff to the Year 6 teacher. Intervention schools were asked when they planned to deliver Unit 2 and control schools were asked for the day/time of their regular writing lessons. Where researchers were unable to attend Grammar for Writing lessons for pragmatic reasons another intervention school was randomly selected and approached.¹⁴ However, two intervention schools cancelled the arranged visit at short notice (due to teacher illness and a timetable change). The researchers were unable to arrange additional visits to replacement schools in these cases because the unit being observed (Unit 2) was not of a sufficient duration for these to be arranged in time. All five control school visits went ahead as planned.

Structured lesson observations

The lesson observations were designed to understand the actual implementation of the programme in the classroom. A Grammar for Writing lesson observation schedule used for the previous EEF efficacy trial was shared by the Development Team. This encompassed the three core components of the programme (use of grammar terms, linking grammar and effects in writing, and using talk to develop discussion about choices and effects). After attending training sessions and working through the lesson plans for the 2 units of work this measure was then adapted by the Evaluation Team. Lessons were assessed according to whether they used each of the central tenets 'as planned', 'partially as planned' or 'rarely'. Each was then turned into a 3-point scale with 'as planned' scoring 3 and 'rarely' scoring 1 leading to a fidelity rating from 0-9 (if a behaviour was not observed at all a rating of 0 was awarded). Where more than one lesson was observed in a school this score was then averaged. The lesson observation schedule was also designed to capture the more detailed dynamics of the programme,

¹⁴ This occurred in 4 cases.

other writing/grammar techniques taught in Year 6 and the wider classroom context e.g., levels of pupil engagement.

A comparable lesson observation measure was developed alongside the intervention school schedule for lesson observation in control classes. This mirrored that used for the intervention lesson observations but with the removal of expectations regarding programme delivery and implementation. For example, the use of the three core components mentioned above was recorded as 'often' (3), 'sometimes' (2) and 'never' (1) as opposed to 'as planned', 'partially as planned' or 'rarely'. The control school lesson observation measure was designed to enable researchers to compare the intervention and control conditions and identify any common strands in lesson delivery and content.

It was planned that each visit would focus on observing one lesson in each school although in some cases this did not occur with researchers being given the opportunity to observe one or more additional lessons. In the event 17 lesson observations (12 intervention, 5 control) occurred across the 13 schools. Two schools (one intervention and one control) were visited by both researchers conducting the school visits together to ensure that there was agreement regarding implementation and fidelity. There were no differences in ratings in either school.

Interviews with teachers/literacy coordinators

Following the lesson observation, the teacher was asked to participate in a brief (20 minute) interview. This was designed to discuss the classroom observation, in particular any adaptations made by the teacher to accommodate their pupils' needs, in order to enable greater understanding of the lesson observed. Additional questions addressed the classroom context, literacy teaching more generally and, for intervention teachers, their experiences of and attitudes towards the programme, including training. Where available (and if different to the Year 6 teacher interviewed), the literacy coordinators in the observed schools were also briefly interviewed to understand the school demographics (beyond those reported in routine administrative data), needs and context, literacy priorities and any challenges in meeting those needs. In total 20 interviews took place. The breakdown of interview participants in intervention and control schools is provided in Table 4.

Table 4: Interview participants by school allocation

	Intervention (N)	Control (N)	Total (N)
Year 6 Teacher	4	3	7
Year 6 Teacher/Literacy Coordinator combined	3	4	7
Literacy Coordinator	5	1	6
Total	12	8	20

Routinely collected data

The Development Team shared attendance data from the CPD training days, results from the end of programme evaluation survey administered at the final (4th) CPD training day (Impact Inventory data) and fidelity data with the Evaluation for the purposes of this evaluation in order to assess the take-up of the training by intervention schools and to act as a proxy fidelity measure in the impact evaluation.

Data on attendance at the CPD training days for the 168 intervention teachers involved in this study (across all 77 intervention schools) for the CPD training days was used to inform the process evaluation and to form a proxy-measure for the non-compliance analysis.

At the end of the fourth day of CPD training intervention teachers were requested to complete an Impact Inventory (i.e. an end of programme evaluation survey). This included implementation of each of the two units of work, changes in practice and perceived changes in student outcomes. These Impact Inventories were completed by 68 teachers.

Members of the Development Team made 13 visits to intervention schools during the academic year to conduct lesson observations. There was no overlap between schools visited by the Development Team and the Evaluation Team. Lessons were scored from 1-3 on 'connections made between grammar and effect' and 'discussion used to tease out thinking and choice-making', with 1=high and 3=low.

The lesson observation ratings collected by both the Development Team and the Evaluation Team were collated to form a fidelity measure for the subgroup analysis. The ratings collected by the Development Team were inverted so that 3=high fidelity and 1=low fidelity. Summed scores (maximum 6, minimum 2) on the two items collected across the lesson observations, linking grammar and effects in writing/'connections made between grammar and effect' (metalinguage), and using talk to develop discussion about choices and effects/'discussion used to tease out thinking and choice-making' (talk quality) were then used for a fidelity sub-analysis. Where more than one teacher was observed the total score for that school was the average across fidelity ratings.

Analysis

The survey data was downloaded from Qualtrics and imported into SPSS. Once in SPSS the pre-test survey and post-test survey were linked using teacher IDs. The data was analysed in SPSS using descriptive statistics and independent means t-tests. The interview data was transcribed and imported into NVivo 10 (QSR International, 2012). It was analysed inductively and thematically. The process evaluation data was then triangulated to ensure the robustness of any research findings, including an in-depth understanding of the programme as intended and as implemented within a natural setting and to identify (1) how schools can support staff in implementing Grammar for Writing, and (2) any appropriate adaptations of the programme to assist with any further development and possible scale-up of the intervention.

Costs

Cost data relating to training was collected directly from the Development Team at the University of Exeter. Intervention schools were also asked about any additional costs of training and programme delivery during the process evaluation literacy coordinator interviews. In addition, researchers reviewed the lesson plans and resources provided to schools in order to assess any underlying costs of implementing the programme. Cost per pupil was calculated by dividing the cost per school by the average number of teachers (2 teachers per school) and the average class size (25 pupils per class).

Timeline

Table 5: Timeline

Date	Activity
May-June 2012	KS1 assessments (obtained from the NPD)
May-October 2016	School Recruitment (including Head Teacher MOU & Teacher consent, and pupil UPN's)
June-October 2016	Teacher baseline survey (online collection by Evaluation Team)
July-October 2016*	Randomisation
October 2016	CPD training day 1
November 2016	CPD training day 2
December 2016	Newsletter 1 distributed to schools [#]
Spring Term 2017	Delivery of Unit 1 (narrative writing)
March 2017	CPD training day 3
March-May 2017	Supplementary consent from headteachers for in-school assessment
April 2017	Letter sent from NFER to schools to arrange writing assessment
Summer Term 2017	Delivery of Unit 2 (persuasive writing)
Summer Term 2017	Classroom observations and teacher interviews
May 2017	Newsletter 2 distributed to schools
May 2017	KS2 assessments (available from NPD October 2017)
May 2017	CPD training day 4 (after KS2 assessments)
June 2017	Writing assessments (administered by NFER, 12-30 June)
June 2017	Teacher post-test survey
June (23rd) - July (21st) 2017	Writing assessments received from NFER (in batches)
June-November 2017	Marking and moderation of Writing assessments
July 2017	NPD data application submitted

January & February 2018^{##}	NPD data received
February-November 2018	Final analysis and report

* Randomisation was conducted in batches as recruitment was on-going during the Summer Term 2016 with some final recruitment occurring in September/October 2016.

Headteachers and teachers participating in the trial in both control and intervention schools all received two newsletters from the Evaluation Team during the study period keeping them informed about the evaluation in order to encourage engagement with the study.

KS2_READMRK was received as a separate dataset in February 2018. The other secondary outcome data was received in January 2018.

Impact evaluation

Participant flow including losses and exclusions

As described above, recruitment was conducted by the Development Team. Using a geographically targeted approach, 1,571 schools in total were approached to participate in the study. 195 expressed an interest in taking part. Of these, 40 schools did not complete the requirements to be eligible for randomisation (i.e. headteacher signed MOU, teacher consent gained, baseline teacher surveys completed and pupil data provided) and did not therefore meet the inclusion criteria. Consequently, 155 schools were recruited to the study (7,239 pupils in total) and randomised. This resulted in a total of 77 schools assigned to treatment, and 78 to control.

Ten schools withdrew from the study post-randomisation; seven intervention schools and three control schools. Of these, four cited difficulties committing to the training days – either the quantity of training or the specific dates – as the principal reason. Of the remaining schools, one withdrew due to a restructuring of the curriculum, two withdrew because of staffing issues (sickness and a change of head), and three did not provide a specific reason. This withdrawal was from further participation in the study and from further data collection from the school. The use of already existing data (i.e. pupil UPN's) was not withdrawn. This was clarified by a follow-up email from the Evaluation Team to these schools.

The 145 schools remaining in the study were contacted to take part in the in-school post-intervention assessments in March 2017 (i.e. after the primary outcome was changed). Of these, 140 schools signed the supplementary consent to allow the additional assessments to take place and their details were provided to the NFER and 5 schools refused consent. At the point of arranging the assessments a further 5 schools withdrew their consent to this aspect of the study, meaning 10 of the remaining 145 schools did not take part in this aspect of the study. Consequently, the primary outcome was collected for 5,416 pupils across 135 schools (66 intervention schools, 69 control schools). The participant flow diagram for the primary outcome is given in Figure 2.

As access to pupil data via the National Pupil Database was not withdrawn from either the withdrawn schools or those schools who refused to participate in the in-school assessment higher numbers of schools and pupils completed the secondary outcomes than the primary outcome. Details of the participant flow diagram for the secondary outcomes are given in Figure 3 which shows that the following data was obtained from the National Pupil Database (NPD) and analysed:

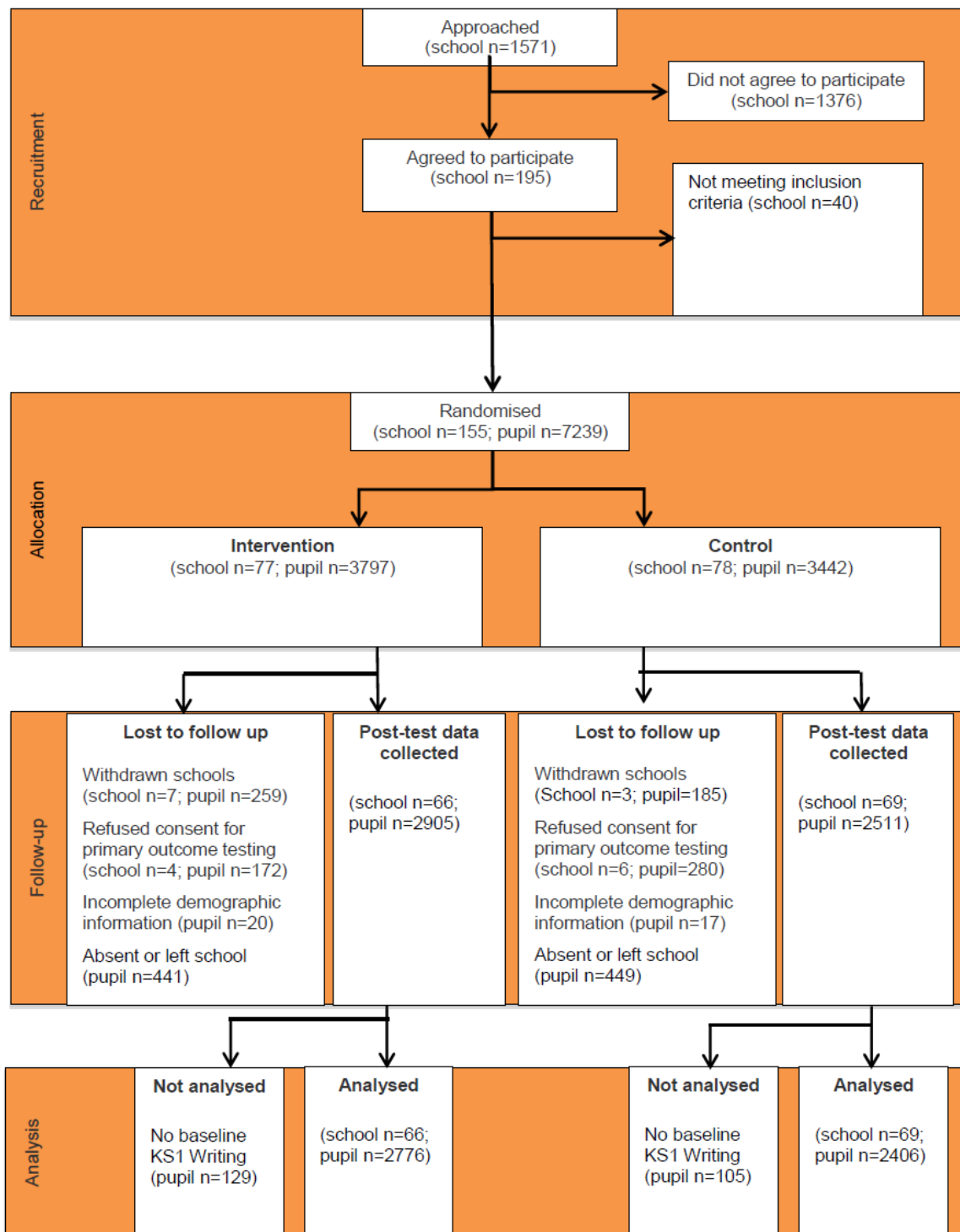
- KS2 writing assessment – 6,787 pupils;
- KS2 reading assessment – 6,646 pupils; and
- KS2 GPS assessment – 6,661 pupils.

There were a total number of 2,057 pupils lost to the study between randomisation and analysis of the primary outcome, which represents just over a quarter of the total (28%). This loss is explained by: school withdrawal from the study (10 schools, 444 pupils);

- school refusal of consent for the amended primary outcome testing (10 schools, 452 pupils);
- incomplete demographic information obtained from NPD extracts (37 pupils);
- pupil's absence on the day of the assessments or movement away from the school (890 pupils); and
- absence of baseline assessment data from NPD (234 pupils).

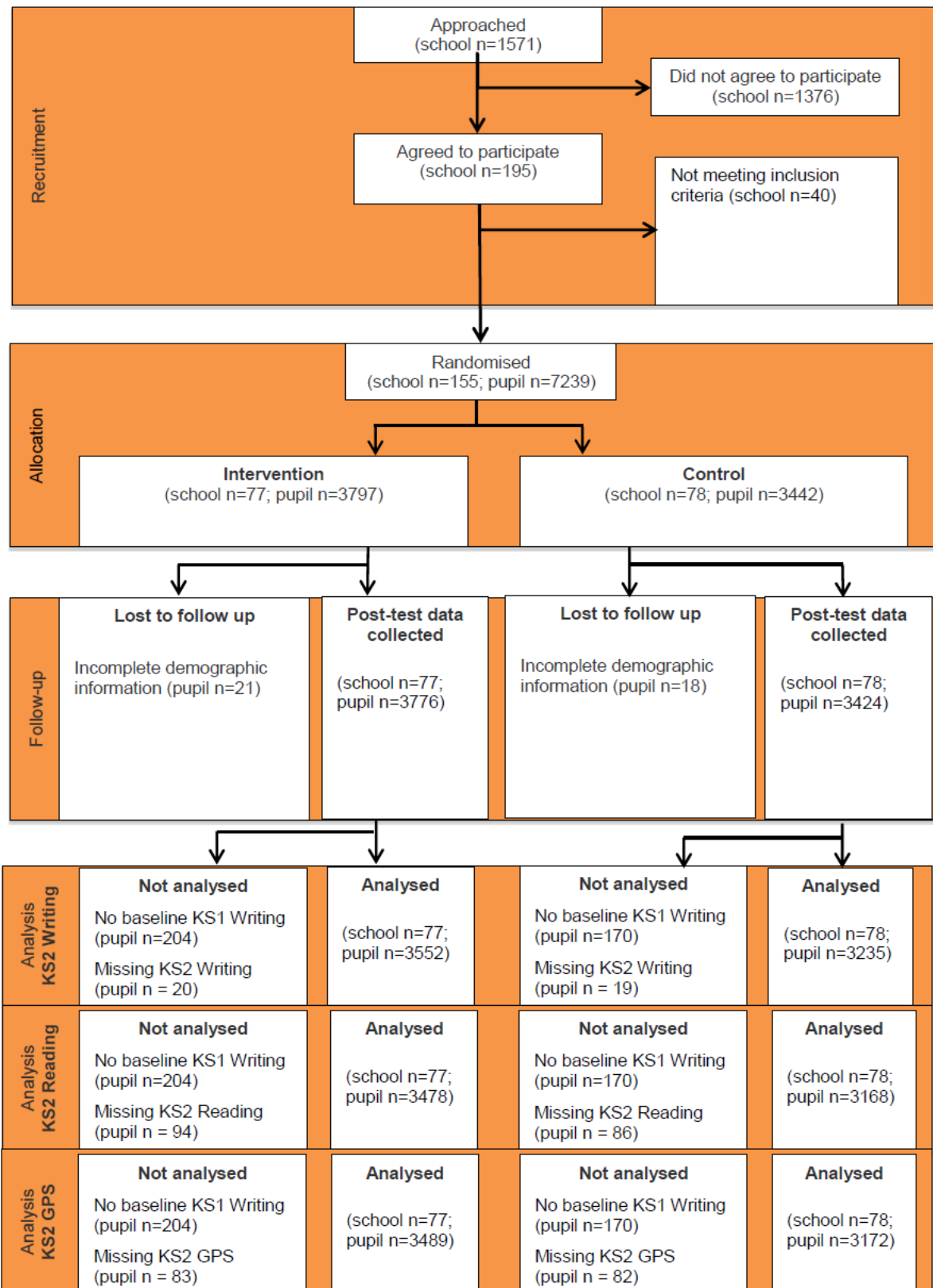
In contrast KS2 assessment data was analysed for between 92% (KS2 reading, KS2 GPS) and 94% (KS2 writing) of pupils as randomised.

Figure 2: Participant flow diagram (Primary Outcome)*



* For imputed case analysis N=6,306 (3,346 intervention, 2,960 control)

Figure 3: Participant flow diagram (Secondary Outcome)*



* For imputed case analysis N=7200 (3776 Intervention, 3424 Control).

Table 6 presents the MDES at different stages of the project. Overall, the loss in pupils and schools for the analysis does not have a sizable impact on the potential MDES. The overall goal was to be able to identify a small effect of $MDES=.20$ and this would have been possible at all stages of the design. The main reason for the small difference in MDES compared to the original proposal are the comparatively low ICCs that were found.

Table 6: Minimum detectable effect size at different stages

		Protocol		Randomisation		Analysis ^d	
		Overall	FSM	Overall	FSM	Overall	FSM
MDES		0.18	0.19	0.18	0.19	0.18	0.18
Pre-test/ post-test correlations	level 1 (pupil)	0.73	0.73	0.73	0.73	0.60	0.60
	level 2 (class)	--	--	--	--	--	--
	level 3 (school)	-- ^a	-- ^a	-- ^a	-- ^a	0.25 ^b	0.22 ^b
Intracluster correlations (ICCs)	level 2 (class)	--	--	--	--	--	--
	level 3 (school)	0.15	0.15	0.15	0.15	0.12	0.10
Alpha		0.05	0.05	0.05	0.05	0.05	0.05
Power		0.80	0.80	0.80	0.80	0.80	0.80
One-sided or two-sided?		Two	Two	Two	Two	Two	Two
Average cluster size		50	16	46.70	21.27	38.39	17.63
Number schools	of Intervention	75	75	77	77	66	66
	of Control	75	75	78	78	69	68
	Total	150	150	155	155	135	134 ^c
Number pupils	of Intervention	3750	1200	3797	1734	2776	1291
	of Control	3750	1200	3442	1541	2406	1071
	Total	7500	2400	7239	3275	5182	2362

Note. Correlations between pre- and post-test could only be evaluated on the analysis sample.

^aIt was assumed that stratification variables would explain up to 10% of the variance, but no assumption regarding the pre-post-test correlation between schools was made.

^bAverage pre-test scores per school were not used in the sample size calculation or as a predictor in the analyses and are not considered in the MDES calculation. The MDES results remain unchanged if pre-test

averages were considered due to the very small correlation of school averages in pre- and post-test measures (between-school correlation).

^c One school had only one pupil fulfilling the FSM criteria and had to be dropped for this analysis.

^d To calculate the MDES the smallest available number of pupils was used in the analysis at each stage to provide the most conservative estimate (i.e. for the post-test calculations the primary outcome was used as this was the measure with the least amount of available data (i.e. had the smallest sample size).

Attrition

As presented in Figure 2 for the primary outcome 7,239 pupils were randomised and data for 5,182 pupils were analysed (72% of the randomised sample). Across treatment groups these ratios (analysed:randomized) were 2,776:3,797 pupils (73%) for the intervention and 2,406:3,442 pupils (70%) for the control group.

As presented in Figure 3 for the secondary outcomes 7,239 pupils were randomised:

- For the KS2 SATS Writing assessment 6,787 pupils were analysed (94% of the randomised sample). Across treatment groups these ratios were 3,552:3,797 pupils (94%) for the intervention and 3,235:3,442 (94%) for the control group.
- For the KS2 Reading assessment 6,646 pupils were analysed (92% of the randomised sample). Across treatment groups these ratios were 3,478:3,797 pupils (92%) for the intervention and 3,168:3,442 (92%) for the control group.
- For the KS2 GPS assessment 6,661 pupils were analysed (92% of the randomised sample). Across treatment groups these ratios were 3,489:3,797 pupils (92%) for the intervention and 3,172:3,442 (92%) for the control group.

Pupil and school characteristics

In total 155 schools were recruited to the study (77 intervention, 78 control). This included 312 teachers (168 intervention, 144 control), an average of 2 teachers per school per condition. In addition, there were 7,239 pupils involved at randomisation (3,797 intervention, 3,442 control). Table 7 provides a baseline comparison of the recruited schools, teachers and pupils involved in the study.

Table 7: Baseline comparison

School-level (categorical)	Intervention group		Control group	
	n/N (missing)	Count (%)	n/N (missing)	Count (%)
School type	77/77 (0)		78/78 (0)	
Academy (converter and sponsor-led)	-	18 (23.4)	-	19 (24.4)
Local Authority	-	59 (76.6)	-	59 (75.6)
Maintained Schools:				
<i>Community School</i>	-	42 (54.5)	-	43 (55.1)
<i>Foundation school</i>	-	8 (10.4)	-	7 (9.0)
<i>Voluntary aided school</i>	-	7 (9.1)	-	8 (10.3)
<i>Voluntary controlled school</i>	-	2 (2.6)	-	1 (1.3)
Urban/Rural	77/77 (0)		78/78 (0)	
Rural	-	6 (7.8)	-	3 (3.8)
Urban	-	71 (92.2)	-	75 (96.2)
Ofsted rating	77/77 (0)		78/78 (0)	
Outstanding	-	13 (16.9)	-	8 (10.3)
Good	-	53 (68.8)	-	56 (71.8)
Requires improvement	-	9 (11.7)	-	10 (12.8)
Inadequate	-	0 (0)	-	0 (0)
No Ofsted assessment	-	2 (2.6)	-	4 (5.1)
N School-level (continuous)	n (missing)	Mean (SD)	n (missing)	Mean (SD)
School size				
Number of schools	75/77 (2)		77/78 (1)	
Total number of pupils	29,654 (-)	395 (285)	26,884(-)	349 (155)
Free School Meal eligibility %				
Number of schools	75/77 (2)	-	77/78 (1)	-
Proportion of pupils	-	25.7 (14)	-	26.1 (13)
School attainment				
Number of schools	75/77 (2)	-	76/78 (2)	-
Percentage of pupils reaching expected standards:				
Key stage 2 Grammar	-	65.9 (17.1)	-	70.0 (19.3)
Key stage 2 Writing	-	78.5 (16.0)	-	77.6 (18.5)
Key stage 2 Reading	-	67.1 (17.0)	-	70.7 (16.2)
Key stage 2 Maths	-	74.3 (15.7)	-	72.9 (16.6)

English as an Additional Language Number of schools	75/77 (2)	-	77/78 (1)	-
Percentage of pupils with EAL	-	21.8 (25.6)	-	22.8 (27.6)
Special Educational Needs Number of schools	75/77 (2)	-	77/78 (1)	-
Percentage of pupils with statement of SEN or EHC plan	-	1.7 (2.0)	-	1.2 (0.9)
Teacher-level (categorical)	n/N (missing)	Count (%)	n/N (missing)	Count (%)
Number of years teaching	125/168 (43)		107/144 (37)	
0-1 year	-	5 (4.0)	-	5 (4.7)
2-5 years	-	41 (32.8)	-	24 (22.6)
6-10 years	-	33 (26.4)	-	33 (31.1)
11-15 years	-	25 (20.0)	-	21 (19.8)
16-20 years	-	15 (12.0)	-	12 (11.3)
21+ years	-	6 (4.8)	-	11 (10.4)
Number of years teaching in current school	125/168 (43)		107/144 (37)	
0-1 year	-	20 (16.0)	-	23 (21.5)
2-5 years	-	61 (48.8)	-	41 (38.3)
6-10 years	-	23 (18.4)	-	25 (23.4)
11-15 years	-	14 (11.2)	-	6 (5.6)
16-20 years	-	6 (4.8)	-	7 (6.5)
21+ years	-	1 (0.8)	-	5 (4.7)
Number of years teaching Year 6 in last 3 years	125/168 (43)		107/144 (37)	
0 years	-	30 (24.0)	-	28 (26.2)
1 year	-	38 (30.4)	-	19 (17.8)
2 years	-	26 (20.8)	-	22 (20.6)
3 years	-	31 (24.8)	-	26 (35.5)
Teacher-level (continuous)	n (missing)	Mean (SD)	n (missing)	Mean (SD)
Grammar quiz, pre-test, raw data	161/168 (7)	21.17 (2.96)	136/144 (8)	20.93 (2.92)
Grammar quiz, post-test, raw data	118/168 (50)	20.79 (2.68)	104/144 (40)	21.06 (2.38)

Pupil-level (categorical)	n/N (missing)	Count (%)	n/N (missing)	Count (%)	
Eligible for FSM	3,773/3,797 (24)		3,423/3,442 (19)		
Yes	-	1,734 (46.0)	-	1,541 (45.0)	
No	-	2,039 (54.0)	-	1,882 (55.0)	
Gender	3,776/3,797 (21)		3,424/3,442 (18)		
Male	-	1,916 (50.7)	-	1,715 (50.0)	
Female	-	1,860 (49.3)	-	1,709 (50.0)	
KS2 Writing assessment outcome (WRITTAOUTCOME)^c	3,757/3,797 (40)		3,406/3,442 (36)		
EXS	-	2,247 (59.8)	-	2,033 (59.7)	
WTS	-	614 (16.3)	-	652 (19.1)	
GDS	-	624 (16.6)	-	523 (15.4)	
PKG	-	141 (3.8)	-	116 (3.4)	
PKE	-	55 (1.5)	-	34 (1.0)	
PKF	-	30 (0.8)	-	38 (1.1)	
BLW	-	30 (0.8)	-	3 (0.1)	
A	-	7 (0.2)	-	5 (0.1)	
D	-	5 (0.1)	-	2 (0.1)	
L	-	1 (0.0)	-	0 (0.0)	
M	-	3 (0.1)	-	0 (0.0)	
Percentage of Pupils with at least one missing value^d	3,797/3,797 (0)	204 (5.4)	3,442/3,442 (0)	170 (4.9)	
Pupil-level (continuous)	n (missing)	Mean (SD)	n (missing)	Mean (SD)	Effect Size
KS1 Writing Result (KS1_WRITPOINTS)	3,572 (225)	14.49 (3.80)	3,254 (188)	14.59 (3.78)	-.03

^a Only assessed in intervention schools; set to 0.0% for all control schools in the implementation fidelity analysis (Appendix D, Section D.5).

^c EXS=Working at the expected standard; WTS=Working towards the expected standard; GDS= Working at greater depth within the expected standard; PKG=Pre-key stage – growing development of the expected standard; PKE=Pre-key stage – early development of the expected standard; PKF=Pre-key stage – foundations for the expected standard; BLW=Below the standard of the interim pre-key stage standards.

^d Calculated across baseline variables Gender, EVERFSM_ALL_SPR17, KS1_WRITPOINTS

As can be seen in Table 7, approximately three-quarters of schools in both conditions were Local Authority Maintained schools, and the majority (70% in total) had a 'Good' Ofsted rating. Although there were slightly more rural schools in the control condition, 94% in total were classed as 'urban'. The schools participating in this evaluation had a higher than average proportion of pupils in receipt of FSM

than schools nationally, and a slightly higher proportion of pupils for whom English was an Additional Language (EAL) although the high reported Standard Deviation suggested this varied considerably between schools (DfE, 2017).

The average proportion of pupils meeting (or exceeding) the expected standard in the KS reading, GPS and maths assessments nationally in 2016-17 were 72%, 77% and 75%, respectively. The averages for participating schools in both conditions were lower e.g. 67% meeting the expected standard or above in reading intervention schools and 71% in control schools. Interestingly, this trend was reversed in the writing assessments where the average proportion of pupils achieving (or exceeding) the expected standard nationally was 76% compared to 79% for the intervention schools and 78% for the control schools.

Of the teachers involved in the study it is interesting to note that those in the control condition were more likely to have been in the teaching profession longer, been teaching in their current school longer, and to have more recent experience of teaching Year 6 than their counterparts in intervention schools.

Only two variables were pre-set as relevant characteristics that needed to be tested for balance at baseline, the KS1 pre-test result and FSM-status. Imbalance was evaluated for FSM-status via standardised differences of proportions which were below our pre-defined threshold for imbalance of $w \geq .05$ (Faul et al., 2007; at full baseline and for secondary outcomes: $w = .02$; primary analysis: $w = .02$); and via standardised mean differences for the KS1 pre-test result which revealed a very small standardised mean difference (using a pooled estimate of the groups variances; EEF, 2018) of .03 in favour of the control group (which was below our threshold for relevance). Further detail including a histogram of the raw scores for KS1 is provided in Appendix D.

Outcomes and analysis

This section reports the main outcomes and analysis. Tables 8-10 present the same information to the same degree of detail for different models. The columns under the heading "Raw Means" present the number of observed respondents and missing values by group, as well as the means and their confidence intervals (not cluster-corrected). The last three columns (headed "Effect Size") present the effect size estimates as derived from the statistical analysis. First, the analysed N is provided as a total and by group. In accordance with the EEF reporting template the Hedges' g estimate is then presented without taking clustering or any covariates into account. This effect size is for reporting purposes only and provides limited information about the effect size of the intervention since it does not take into account that observations are clustered by school. The final column presents the adjusted effect based on analytic model with its bootstrapped confidence interval. The effect size is described in detail in the section on 'Effect size calculation' above and is described in the EEF's guidance notes (EEF, 2018, p. 4; see also: Hedges, 2007; Xiao et al., 2016).

Since the confidence intervals for the effect sizes are determined via bootstrap, missing data occurs in the observed data analysis. Within each of the $b = 1,000$ bootstrap samples, a different number of students with missing data is selected in each run. Due to these unequal missingness patterns across bootstrap samples the N for the observed data analyses is provided as the average number of pupils analysed in these 1,000 runs (separately for total, control and intervention group, i.e. the subgroup N 's do not necessarily add up to the total N). As the tables show, the average is always very close to the number of available cases and Appendix D provides further details (e.g. standard deviations of the number of selected pupils to gauge the variability).

While this procedure could in principle lead to wider confidence intervals for the effect sizes since sometimes fewer pupils than the available complete responses are selected, based on the results this effect, if present, is judged to be minimal. Two aspects speak to this point especially: (1) The standard deviations for the number of selected pupils are small compared to the overall sample size (Appendix D), which indicates that very similar numbers of students were used in each run; and (2) the breadth of

the confidence intervals (i.e. the precision with which the effect sizes are estimated) is only minimally increased when comparing observed data results with those from the imputed data; since the imputed data analysis uses more cases a higher precision for these is expected (i.e. narrower confidence intervals), but the difference would be larger if the bootstrap procedure led to a markedly reduced precision.

Finally, all analyses are presented based on the observed data (i.e. non-imputed; the lead analysis as defined in the SAP) and with imputed results (as a sensitivity analysis). The imputed results use all available data for that specific outcome (see notes to Figures 1 and 2 above) and have the same number of cases in both groups for all runs since there are no missing data.

Primary Outcome Analysis

The primary outcome analysis was conducted as described above. As also described, only the 135 schools that participated in the post-testing were included in the analysis. As can be seen in Table 8, the adjusted effect based on the analytic model is small and the confidence interval includes "0". Therefore, the hypothesis that the two groups perform in the same way on the primary outcome measure cannot be rejected, and observed differences in average scores are likely to be due to chance variation. This result also holds up when looking at the imputed data analysis (also Table 8).

Table 8: Primary analysis based on the primary analysis set (N=5,182; imputed N=6,306)

Outcome	Raw means				Effect size		
	Intervention group	Control group		N in model (intervention; control)	Unadjusted Hedges g (95% CI)	Adjusted effect based on analytic model (95% CI)	
n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)				
KS2_{past} writing paper	2,905 (441)	16.09 (15.82, 16.35)	2,511 (449)	16.41 (16.14, 16.69)	5,181.22 (2,775.72; 2,405.50)	-.05 (-.10, .00)	-.02 (-.08, .03)
KS2_{past} writing paper, imputed data	--	--	--	--	6,306 (3,346; 2,960)	-.05 (-.10, .00)	-.03 (-.08, .02)

Note. Due to unequal missingness patterns across bootstrap samples the N for the observed data analyses is provided in averages of the analysed cases in these bootstrap runs. More detail regarding intraclass correlations and bootstrapped variance components can be found in the corresponding tables in the appendix.

Secondary Outcomes Analysis

The analysis of the three secondary outcomes was conducted as described above. As also described, all schools providing NPD data (N=155) were considered in this analysis since the data for the secondary outcomes were collected from the NPD. As can be seen in Table 9, the adjusted effect based

on the analytic model is small for the KS2 Writing and KS2 Reading assessments and their confidence intervals include zero. Therefore, the hypothesis that the two groups perform in the same way on these two secondary outcome measures cannot be rejected, and observed differences in average scores are likely to be due to chance variation. This result also holds up when looking at the imputed data analysis (also Table 9).

The results of the analysis for the KS2 GPS assessments show that, while the effect sizes are still small, both in the observed and imputed data analyses the confidence intervals do not include zero. In this case the hypothesis of equal performance can be rejected: pupils in schools that received the intervention did slightly worse in the KS2 GPS assessment than pupils in the control schools (ES= -0.06; approximately 1 month less progress). This could be a result of the GPS assessment being a decontextualized assessment whereas the Grammar for Writing intervention advocates a contextual approach to writing. It should be noted, however, that although this analysis was pre-planned, the analytic strategy was not devised to definitely test for positive and negative effects on secondary outcomes (see section on statistical analysis above). Specifically, the analysis of the secondary outcomes does not guard against false-positive discovery rates, which means that this finding could be due to chance and therefore does not necessarily indicate the presence of an effect. Therefore, this result alone cannot be seen as definite evidence for a negative effect of the Grammar for Writing programme on GPS outcomes.

Table 9: Secondary analysis based on the secondary analysis set (N=7,200)

Outcome [#]	Raw means				Effect size		
	Intervention group		Control group		N in model (intervention ; control)	Unadjusted Hedges g (95% CI)	Adjusted effect based on analytic model (95% CI)
n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)				
KS2 Writing assessment outcome^a	3,314 (462)	5.81 (5.78, 5.84)	2,939 (485)	5.81 (5.77, 5.84)	6,788.11 (3,552.52; 3,235.59)	.00 (-.05, .04)	.02 (-.02, .07)
KS2 Writing assessment outcome, imputed data	--	--	--	--	7,200 (3,776; 3,424)	.00 (-.05, .04)	.00 (-.04, .05)
KS2 Reading assessment outcome^b	3,221 (555)	29.95 (29.58, 30.31)	2,872 (552)	29.92 (29.57, 30.27)	6,647.12 (3,478.01, 3,169.11)	.00 (-.04, .04)	.01 (-.03, .06)
KS2 Reading assessment outcome, imputed data	--	--	--	--	7,200 (3,776; 3,424)	-.01 (-.05, .03)	.00 (-.04, .05)
KS2 Grammar, Punctuation and Spelling assessment outcome^c	3,229 (547)	45.23 (44.73, 45.73)	2,874 (550)	45.63 (45.11, 46.15)	6,662.24 (3,489.36; 3,173.09)	-.05 (-.10, -.01)	-.06 (-.10, -.01)
KS2 Grammar, Punctuation and Spelling assessment outcome, imputed data	--	--	--	--	7,200 (3,776; 3,424)	-.06 (-.10, -.02)	-.06 (-.11, -.02)

^a NPD variable: WRITTAOUTCOME

^b NPD variable: KS2_READMRK

^c NPD variable: GPSMRK

[#] Students whose scores were too low (in the secondary outcomes) and would have been assigned a "N" were given the lowest available scale value (rare occurrence: N=18 for reading and N=4 for GPS, very few).

Note. Due to unequal missingness patterns across bootstrap samples the N for the observed data analyses is provided in averages of the analysed cases in these bootstrap runs. More detail regarding intraclass correlations and bootstrapped variance components can be found in the corresponding tables in the appendix.

Subgroup analyses

As specified in the protocol, subgroup analyses were carried out for pupils eligible for FSM, boys and girls, and high and low achievers on the pre-test (KS1; median-split based on all observed scores). The results for these tests indicated that no statistically significant interaction was observed between the intervention and FSM-status as well as the intervention and gender, although a significant interaction was found for pupils' pre-test performance.

Table 10 presents the summary statistics for pupils eligible for FSM (which was planned to be reported) and the KS1 pre-test result as the only potentially statistically significant subgroup effect. There is no relationship between the intervention and the primary outcome measure for the pupils with higher prior attainment (indicated by very small effect sizes and confidence intervals that overlap with 0); but for the lower performing pupils a negative relationship is found. The lower performing pupils in schools that received the intervention were found to perform overall slightly worse than comparable pupils at schools that did not receive the intervention ($ES=-0.11$; equivalent to 2 months' additional progress). This result also holds when missing data are imputed. However, as the study was planned to provide evidence in the analysis of the primary outcome, not to follow-up subgroup effects in detail, this result alone cannot be used as evidence for a negative effect on students with lower KS1 writing task results because there is not sufficient power and, as such, no way of knowing if these results happened by chance.

More detailed results of the subgroup analysis can be found in Appendix D, Section D.1.

Table 10: Subgroup analysis; unadjusted and adjusted effect size estimates based on primary analysis set (N=5,182; imputed N=6,306)

Outcome	Raw means				Effect size		
	Intervention group		Control group		N in model (intervention; control) ¹⁵	Unadjusted Hedges g (95% CI)	Adjusted effect based on analytic model (95% CI)
n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)				
FSM-only							
KS2 _{past} writing paper	1,325 (241)	14.84 (14.46, 15.22)	1,095 (263)	14.84 (14.44, 15.23)	2,360.61 ¹⁶ (1,290.51; 1,070.09)	.00 (-.07, .08)	.05 (-.03, .13)
KS2 _{past} writing paper _t , imputed data	--	--	--	--	2,924.76 (1,567.38; 1,357.38)	-.01 (-.08, .06)	.03 (-.05, .10)

¹⁵ N = 2924 students provided information on their FSM status (+4 for whom no information was available; see table Table D.3.5); N = 2420 students provided information on the primary outcome variable; N = 2362 students were available for the analysis, i.e. had also a pre-test result).

¹⁶ Point estimates were estimated based on the N presented next to it. The effect sizes and their confidence intervals were bootstrapped with the numbers presented in "n in model" columns.

KS1 Writing Result, upper 50%							
<i>KS2_{past}</i> writing paper	1,807 (198)	18.7 (18.39, 19.01)	1,550 (203)	18.8 (18.43, 19.07)	3,357.66 (1,807.52; 1,550.14)	-.01 (-.06, .05)	.01 (-.05, .08)
<i>KS2_{past}</i> writing paper, imputed data	--	--	--	--	3,886.27 (2,072.01; 1814.26)	-.01 (-.07, .05)	.01 (-.05, .07)
KS1 Writing Result, lower 50%							
<i>KS2_{past}</i> writing paper	969 (199)	11.47 (11.10, 11.83)	856 (222)	12.28 (11.87, 12.68)	1,824.83 (968.80; 856.03)	-.14 (.22, -.06)	-.11 (-.20, -.03)
<i>KS2_{past}</i> writing paper imputed data	--	--	--	--	2,419.73 (1,273.99, 1,145.74)	-.14 (-.22, -.07)	-.11 (-.18, -.02)

Note. Due to unequal missingness patterns across bootstrap samples the N for the observed data analyses is provided in averages of the analysed cases in these bootstrap runs.

Non-compliance Analysis

A non-compliance analysis was conducted using attendance at CPD training days as a proxy measure for compliance. The training was planned for four days of CPD. As detailed below, the fourth day was scheduled at the time of the KS2 SATs, which impacted on teacher attendance levels. Attendance at the initial three training sessions was deemed to be compliance with the programme as these three days of CPD focused on the Grammar for Writing approach, and delivery of the two units (as described above). Of the 168 intervention teachers:

- 129 (77%) attended all three of these CPD training days;
- 18 (11%) attended two days of training;
- 7 (4%) attended one day of CPD training; and
- 14 (8%) attended none of the CPD training days offered as part of the programme.

Based on observed and imputed data no relationship was found for the primary outcome and effect sizes were generally low (-.02 for the full sample; up to .06 for FSM pupils only): There was no correlation between the number of CPD training days attended and pupil outcomes.

For the secondary outcomes, however, the main analysis already demonstrated a potentially negative effect of the intervention on KS2 GPS assessment outcomes (see Table 9). This trend is replicated for all secondary outcomes when the degree of participation in CPD is used as a proxy for compliance. In all three measures the pupils in schools whose teachers attended more of the training sessions (including the schools that were part of the control group with "0" CPD days) do on average slightly worse than those pupils in schools whose teachers did not take part or went to a smaller number of CPD training days. The effect sizes for this difference range from -.06 (WRIT) to -.14 (GPS).

Although the final (4th) day of CPD training was designed to assist teachers with delivering the programme in the future, it was scheduled at the time of the KS2 SATs, prior to the primary outcome assessments being conducted which impacted on teacher attendance levels. In addition, in the Impact Inventories collected from participants during this CPD training days a number of teachers (n=19) indicated that they had not yet delivered the second unit of work. To evaluate whether our decision to exclude the last day of training had an impact on the results, it was decided to re-run the compliance analysis to include this fourth day of training, although attendance was low with only 39% (n=66) of intervention teachers attending this fourth day and only 35% (n=59) of intervention teachers attending all 4 days of CPD training.

For the primary outcome the observed and imputed data analyses corroborate the finding from the main analysis: no treatment effect is also found with this proxy. Nevertheless, for the observed data looking at FSM eligible pupils only, a small positive potential effect is found, which is nevertheless non-significant in the sensitivity analysis with the imputed data. It repeats the findings from the main analysis which indicates that FSM-pupils potentially did better (albeit with a small effect size).

Looking at the secondary outcomes the known pattern is repeated. All secondary outcomes show potentially small effects. In all three measures the pupils in schools whose teachers went to more training days do on average slightly worse than those pupils in schools whose teachers did not take part or went to a smaller percentage of CPD training days. Further details of the non-compliance analysis are reported in Appendix D, Section D.5. Further details relating to teacher attendance at CPD training can be found in the Process Evaluation section of this report.

These analyses were planned but did not control for the number of analyses conducted, which means that the finding could be due to chance rather than indicate a real negative effect; statistically they can only be seen as a potential indication for a negative effect of the treatment. Nevertheless, the consistency of this finding across analytic strategies and outcome measures warrants further exploration. The study was not designed to follow such a finding up in detail, therefore a statistical chance finding or negative spill-over effects (spill-over effects are effects of an intervention on additional outcomes than the intended one; negative spill-over effects those that adversely affect these additional outcomes) are two possible explanations for these findings. It could also be that teachers with less experience of teaching Year 6 writing or with lower grammar knowledge attended more days of CPD training. However, the correlation between CPD training attendance and the Grammar Quiz scores on teacher / school level (all correlations between CPD training and pre or post Grammar Quiz scores below |.10|, i.e. less than 1% of the observed variance in scores was related to CPD training day attendance) does not indicate that grammar knowledge had such an effect. This could be followed up by further exploratory analyses of the data, but since the Grammar Quiz which was an actual proxy of the causal mechanism (i.e. it was hypothesized that improved grammar knowledge by teachers would lead to improved pupil writing skills) was also assessed (see below), such analyses were not planned.

It is important to remember, the proxy is not assessing the intervention itself, but participation of the teachers in CPD training. While there is a link between the two (attending more days and participating in training could lead to better knowledge and practice), it might be that teachers who felt a greater need for training went more often to the training or other selection effects that led teachers to attend CPD days might be at work.

Grammar Quiz Analysis

As discussed above, a grammar quiz was embedded in the baseline and follow-up teacher survey. Whilst it must be borne in mind that there were fewer points available in the follow-up grammar quiz compared to the baseline quiz (scored out of 30 and 29, respectively), as seen in Table 7, the control group demonstrated an increase in scores during the academic year (pre-test mean=20.79 (SD=2.68); post-test mean=21.06 (SD=2.38)) when compared to the intervention group which showed a slight decrease (pre-test mean=21.17 (SD=2.96); post-test mean=20.93 (SD=2.92)). With the largest of these differences being smaller than an effect size of Hedges $g = .12$, these changes in scores are nevertheless negligible.

The main analysis the Grammar Quiz was used for was a multilevel model as used for the primary outcome with the post-intervention quiz score as a predictor and potential mediator between the intervention and pupils' KS2 past performance. No potentially statistically significant relationship was found (See Appendix D). Analysis was also run to assess whether the percentage of CPD training days attended correlated with teachers' performance in the grammar quiz. This was not found to be the case: all correlations found were smaller than $|.10|$, i.e. they explained less than 1% of the variance in grammar quiz scores.

In addition, several analyses were performed to investigate the validity of the on-line Grammar Quiz as a proxy for teacher's grammar knowledge. It was found that the psychometric properties of the Grammar Quiz were relatively weak: Reliability estimates ranged from .35 (retest with parallel test in control group) to .55 (Kuder-Richardson at baseline), i.e. the Grammar Quiz did not measure inter-individual differences in teachers' grammar knowledge precisely. A similarly important question is, whether the performance in the grammar quiz can actually be summarised in one score. Several analyses showed that this was not the case: the explained variance based on principal component analyses was rather low (<10% with a single component) and parallel analyses indicated that it is likely that several components are needed to represent the content of the quiz sufficiently. Specialised psychometric models for educational data (Rasch and Item Response Analyses) were originally planned but their use in our application pre-supposed that the quiz data would be unidimensional (i.e. summarisable with a single weighted score) which the previous analyses indicated was not the case. Therefore, these models were not applied.

To conclude, the Grammar Quiz was able to assess individual differences, but the reliability of the quiz' scores was low. The baseline grammar quiz was developed by the Development Team and is used regularly in their CPD training to measure relevant grammar content. The post-test grammar quiz was developed based on this model by a member of the Evaluation Team with expertise in this area. Due to this development process, it is likely that the quiz had high content validity, i.e. that it was covering the relevant areas of grammar knowledge. However, the quiz did not enable us to order teachers consistently from lower to higher grammar knowledge on a single score. The quiz further did not respond to the treatment on teacher level (i.e. in our exploratory analysis no differences between teachers were found nor did they correlate with whether they received the treatment) or serve as a mediator in a multilevel model (i.e. our analyses did not show that grammar knowledge as proxied by the quiz is the causal mechanism by which the programme works). A factor affecting the results of the grammar quiz may have been that the measure was administered on-line and therefore conditions were not controlled (i.e. teachers were able to complete the quiz in their own time and could have had access to variable levels of resources or be subject to distractions).

Further details of this analysis are presented in Appendix D, Section D.6.

Missing data

Appendix D, Section D.3 presents descriptive and analytic detail regarding missing data. At the school level the schools who withdrew from the study were similarly distributed across the two treatment groups

(9/78 control and 11/77 intervention). Schools remaining in the study had on average lower proportions of pupils meeting the expected standards in Reading and Maths and lower proportions of EAL pupils, but a higher proportion of pupils with SEN than those who withdrew.

At the individual level, for the primary outcome analysis FSM-status increased the probability of reporting any primary outcome data for 'higher score in KS1 writing'; and increased the probability of reporting the primary outcome (of those schools which took part in the assessment). For the secondary outcome analysis, female pupils were more likely to return any of the three outcomes as were pupils with higher KS1 writing scores.

Due to the small number of available variables a more detailed description is not possible. Nevertheless, while some systematicity in missing data was observed, this lends at least some validity to our imputation-based sensitivity analyses which depend on the assumption that missing data can be predicted based on variables within the data set ("missing at random"; e.g., EEF, 2018). However, whether additional factors may have contributed to missingness and drop-out, especially depending on our target outcomes (i.e. "missing-not-at-random") is ultimately unclear.

Deviation from planned analysis

There were four deviations from the analysis as planned not detailed above. These, and the reasons for these deviations, were as follows:

- Given the number of schools who withdrew from the revised primary outcome but not from the use of pupil data for the secondary outcomes we used two samples for the final analysis, as detailed in Figures 1 and 2 above. This enabled the secondary outcome to be based on a larger sample size and for the analysis reported here to take into account as much of the available data as possible.
- The protocol planned to analyse the secondary outcomes both separately and combined. It was intended that by combining the KS2 writing, reading and GPS assessment outcomes it would be possible to determine if the intervention had an overall effect on pupils' literacy outcomes. However, given the difficulty of producing a composite score which included the 7-scale KS2 Writing assessment outcome, EEF latest guidance against using composite scores (EEF, 2018) and the results of the analysis the secondary outcomes individually it was decided by the Evaluation Team to not conduct this further combined analysis.
- In the SAP it was indicated that the fidelity measure would be devised by using intervention teacher responses to the post-test survey, in particular regarding self-reported adaptations made to programme delivery. It was intended that these changes would be classified as programme-conform vs. non-conform by the Evaluation Team with advice from the Development Team. This would be turned into an individual score for each teacher (0 = no or only conform changes; +1 per non-conform change) and these scores used to classify teachers per median split. However, the large number and range of changes recorded by teachers and the difficulty in establishing the extent to which these changes were within the bounds of programme delivery as intended meant that fidelity ratings from Development Team and Evaluation Team lesson observations were deemed more suitable for this measure, which was developed as described above.
- The analysis was planned to be conducted by the statistician whilst blind to condition. This took place in the initial analysis of primary and secondary outcomes. However, due to a small error being identified in the original dataset (impacting on 370 cases) the analysis reported here was conducted unblinded. The overall results were, however, little changed between the two stages.

Cost

The costs for Grammar for Writing were approximately £1,435 per school to implement in Year 6 during the evaluation year. This figure includes the following costs for the first year. A breakdown of these costs is provided in Table 11. It includes a cost of £700 per delegate for four days of CPD training

although in this trial the intervention schools received a subsidy towards this cost (as they paid only £500 per school for training).

Table 11: Breakdown of costs for programme delivery in the first year (per school)

Item	Detail	Cost
Training	£700 per delegate for four days of CPD training	£1,400
Photocopying	14 pages per pupil @ 5p per page	£35
Total		£1,435

The above figure assumes 2 teachers per school receiving training and an average of 25 children per class (based on the average for our sample of intervention schools). On that basis the average cost of the programme would be approximately £28.70 per pupil in the first year.

It should be noted, however, that this figure does not include the cost of teacher time to attend training (4 days per teacher, average of 8 days in total per school in the first year). We are aware that some schools incurred an additional financial cost for this, with interviewees in approximately half of process evaluation schools indicating that their school bought in supply cover for teacher attendance at CPD training days, whereas the other half were able to use internally available teacher cover. This latter option would nevertheless have incurred an additional 'opportunity cost' for schools in terms of use of resources. The cost of travel for training is also not included but would need to be borne in mind by schools anticipating taking up the programme. Evidence from the intervention schools suggests this figure varied widely between schools depending on distance and mode of travel (car share etc.). In addition, some teachers did not claim for travel costs.

The reported cost does include some minimal photocopying costs associated with delivery of the programme. We have included them as they are specific costs associated with delivery of the programme which, unlike training costs, would need to be repeated in subsequent years. However, schools indicated that photocopying for literacy teaching in general was high and that these costs were typically subsumed within the overall total. Schools were directed to provide 'Magpie books' for pupils but from the process evaluation data this did not appear to occur to a greater extent than was the case in control schools. Although schools were recommended to acquire one specific authentic text (RRP £6.99) in order to provide additional context for pupils in Unit 1 we have not included this cost as only one school in the process evaluation interview sample mentioned purchasing the book and it was not essential to programme delivery.

Beyond the first year the only repeat costs would have been the photocopying costs for pupils. Consequently, the cost per pupil over a three-year period is approximately £10.

Table 12: Cost per pupil per year over three years

Number of years using the programme	Cumulative cost per pupil (£)	Average cost per pupil per year (£) (cumulative costs per pupil/number of years)
1 year	£28.70	£28.70
2 years	£29.40	£14.70
3 years	£30.10	£10.03

The Development Team currently offer training in a bespoke version of the programme for £750 a day plus travel and accommodation. This training can accommodate up to 50 teachers per session and is, therefore, cheaper than the training as delivered for this version of the programme.

Process evaluation

The process evaluation was designed to:

- examine more closely the relationship between the level of implementation of the intervention and its impact on pupil outcomes;
- explain variability in implementation, including understanding the context of the implementation and social processes within schools; and
- address possible barriers to implementation.

This section firstly explores the reasons for schools becoming involved in the Grammar for Writing evaluation in order to further understand the context and motivations of the schools participating in this study. Secondly, it examines levels of implementation and programme fidelity within schools, including barriers to implementation. Thirdly, it examines the outcomes of the programme as perceived by schools and discusses this in the context of actual pupils' outcomes as measured by this evaluation. Finally, it explores usual practice in teaching writing, with a particular focus on writing instruction in the control schools. The process evaluation findings are based on analysis of teacher surveys, teacher interviews and lesson observations, supplemented by additional information collected by the Development Team, as detailed in the methodology section above.

School motivations for participating in the study

Interviewees (control and intervention) were asked about the reasons their school had expressed an interest in participating in the study, in particular, why they were interested in the Grammar for Writing programme. Of the 16 interviewees who answered this question (5 control, 11 intervention) the reason mentioned most frequently (by 12 (75%) of the 16 respondents) was the desire to improve child outcomes (including, but not exclusively, assessment results), either as a whole or for specific targeted groups. Where specific groups of children were mentioned, pupils in receipt of the pupil premium, and struggling pupils and boys were particularly mentioned, although for one school, their interest in the programme was specifically to help their higher attaining children.

We're getting the expected, but it's the greater depth that we're not achieving [and] that was a major reason why we applied for Grammar for Writing, to hopefully stretch our more able children. (Literacy Coordinator, School (Literacy Coordinator, School 21, Control)

Multiple interviewees (n=7, 44%) in both the control and intervention conditions identified the contextualised grammar approach as one of the attractions, both to the teacher and/or the school.

The more we can contextualise [grammar] for children, that's what I started to see would be the opportunities, you know, we want children to be better writers, we have to teach grammar and punctuation lessons for a test, can we bring the two together a little bit more. (Teacher, School 122, Control)

Three interviewees specifically mentioned an awareness of the previous work of the lead developer, Professor Debra Myhill:

I knew of Debra Myhill and her work, and I really liked it and when I heard about this project, I thought this would be brilliant. (Literacy Coordinator, School 136, Intervention)

Other reasons given included the opportunity to meet a perceived need in staff development (n=5, 31%), and the desire to be involved in innovation in teaching, and in contributing to research in an area of particular interest to them (n=4, 25%).

I think the view was taken that staff need support with grammar because particularly with TAs and myself much of it is new for us... as a teacher we have to know the grammar inside out really. So I think part of it was to upskill the teachers. (Teacher, School 99, Intervention)

One of the things we want to do is become more involved in the research base... we want to be current and up to date because we want to give our children the best possible education that we can. (Teacher, School 66, Control)

The reasons for interest in adopting the programme can, therefore, be seen to be similar for both control and intervention schools although, interestingly, this latter view (an interest in being involved in research) was expressed by more teachers in the control condition, despite the smaller number of interviewees in control schools (3 out of 5 control respondents, compared to 1 out of 14 intervention respondents). This may, however, be due to the fact that the interviews took place after schools were aware of their allocation to control or intervention.

Levels of Implementation and fidelity to the programme

In both the teacher survey and interviews intervention teachers were asked about attendance at, and attitudes towards, the training, levels of programme delivery and adaptations to the Grammar for Writing programme which may have impacted on levels of fidelity in order to better understand the context of programme use within schools and possible barriers to implementation. In addition, the final (fourth) day of CPD training included an end of programme Impact Inventory delivered by the Development Team. Data relating to implementation fidelity from that evaluation is also reported here.

As indicated in the participant flow section above, seven intervention schools withdrew from the study during the evaluation year, primarily due to logistical reasons e.g. difficulties attending training days, schools staffing issues. In the follow-up survey intervention teachers were asked if they had implemented the Grammar for Writing programme during the academic year. All 125 intervention teachers (from across 64 schools) who responded to the follow-up survey stated that they did implement the programme. Similarly, data collected from teacher interviews, lesson observations and by the Development Team are restricted to those teachers who reported implementing the Grammar for Writing programme. Limitations in the data, and possible bias, must therefore be acknowledged.

Training

As discussed above, three central days of CPD training were provided as part of the programme in order to train teachers in the underlying principles of the programme, improve teachers' grammar subject knowledge and train teachers to implement the programme using the lesson plans and resources provided. A fourth day of CPD training was provided at the end of the programme to provide teachers with the skills and knowledge to apply the programme principles and techniques to their teaching of writing using authentic texts in the following school year. Given the central role of training in the programme, the impact evaluation used attendance at training as a proxy measure for compliance with the intervention (see Impact Evaluation above). Tables 13 and 14 present further detail on the number of sessions teachers attended and a breakdown of attendance at each CPD training day.

Table 13: Number of CPD training days attended

Number of CPD training days attended	N (%)
0	14 (8)
1	7 (4)
2	18 (11)
3	70 (42)
4	59 (35)

Table 14: Attendance at CPD training days

Attendance at CPD training days	N (%)
Day 1	145 (86)
Day 2	147 (88)
Day 3	135 (80)
Day 4	66 (39)

As can be seen in Table 13, 129 of the 168 teachers in the 77 intervention schools (77%) attended three or more CPD training days. The main barrier to attendance at training mentioned in the interviews was the difficulty of covering the teachers' time (i.e. through paying for supply cover or reallocating school staff), indicating that, as for those schools who withdrew from the study, staffing pressures in schools are a barrier to CPD:

We couldn't get the cover for everybody for the third training session. (Literacy Coordinator, School 92, Intervention)

Table 14 demonstrates that attendance at the fourth day of CPD training was considerably lower than previous sessions (66% of intervention teachers compared with an average of 85% across the other three CPD training days). Teachers explained that this was primarily due to the timing of the sessions, being as they were, scheduled close to SATs week.

In the follow-up survey, respondents in intervention schools were asked a number of Likert-scale questions about the training provided as part of the Grammar for Writing programme. Table 15 shows that nearly all respondents agreed (or strongly agreed) that the training was useful, effective and they felt comfortable using the programme after attending the training.

This suggests that the experience of attending training was positive for teachers overall and a lack of attendance could be posited as a hindrance in programme implementation. However, as the compliance analysis indicates, attendance at training did not have an impact on pupil outcomes in this study.

Table 15: Teacher Feedback on Training and Materials

	Strongly Agree N (%)	Agree N (%)	Neither Agree or Disagree N (%)	Disagree N (%)	Strongly Disagree N (%)
The training was useful	79 (63)	42 (33)	4 (3)	-	-
The training was effective	74 (59)	44 (35)	7 (5)	-	-
I felt comfortable using the programme after attending the training	68 (54)	52 (42)	4 (3)	1 (1)	-
The materials were useful	55 (44)	60 (48)	7 (6)	3 (2)	-
The materials were effective	52 (42)	55 (44)	13 (10)	5 (4)	-
The materials were suitable for my students	48 (38)	61 (49)	8 (6)	7 (6)	1 (1)

Programme delivery

As also seen in Table 15 teachers were asked a number of Likert-scale questions in the teacher survey regarding the materials which were handed out at training and subsequently provided on-line. Responses were strongly positive about the suitability and usefulness of the materials, the number of respondents indicating any reservations being very small. However, of the 125 individual teacher responses across the 64 schools, nearly three-quarters (n=91, 73%) said they had made changes to the programme materials and/or its recommended scheme for delivery. These teachers came from 55 of the 64 intervention schools (86%) represented in the follow up survey. In schools where more than one teacher responded to the survey (n=40 schools), responses were mixed in almost half of these (n=18). Thus it is also the case that some, or all, respondents from 28 of these 64 intervention schools (43%) made no changes. The nature of these changes and the reasons for them are discussed in more detail below using data collected by the Development Team, the lesson observations and interviews conducted by the Evaluation Team and the teacher survey responses in turn.

Impact Inventory

The end of programme Impact Inventory completed by teachers at the end of the fourth day of CPD training asked a number of detailed questions relating to programme delivery for each of the two units of work. The results were collated by the Development Team and are presented in Table 16. It should be noted that here and elsewhere in the reporting of this data the sample consists of 68 teachers whereas attendance data indicates that only 66 intervention teachers participating in the study participated in this fourth day of CPD training. One explanation could be that teachers attended this CPD training day who were not included in the intervention teacher sample due to teacher turnover during the school year (i.e. they taught Year 6 writing during the evaluation year but were not the Year 6 teacher when the baseline survey was administered). An alternative reason is that some teachers attended the CPD training day who were not the Year 6 teacher during the evaluation year (2016-2017) but were intended to teach Year 6 writing in the subsequent academic year (2017-2018) and therefore attended as future teaching of the programme was the focus of this particular CPD training day. This latter explanation appears more likely as two of the 68 teachers reported not having delivered the first

unit of work and more than two teachers indicated that they had not delivered the second unit (as discussed below). The data from all 68 teachers is, however, reported here.

Table 16 indicates that there was some variation in delivery of the programme as planned, although the results reported do appear somewhat contradictory. In particular, 88% of teachers reported having taught all four lessons of Unit 1 each week and 73% reported that all of the activities were undertaken. Forty per cent of teachers indicated that they had not taught some of the lessons and half of all teachers (50%) reported skipping/combining lessons in this Unit. As will be seen below this can be explained by teachers adapting the programme; lengthening or abbreviating aspects of the programme as planned to suit the perceived needs of their own classroom context.

Teachers reported less adherence to delivering Unit 2 as planned than in Unit 1, although shorter in length (2 weeks delivery compared to 4 weeks, respectively). Only 65% reported teaching all 4 lessons each week, 59% reported that all activities were undertaken as planned and only 15% reported that the lessons broadly followed the lesson plans. However, 17 teachers reported that they had not begun to teach Unit 2 at the time of the fourth day of CPD training due to the pressures of preparing for the KS2 assessments.

Table 16: Teacher report on extent of programme delivery

	Unit 1 (Narrative writing)		Unit 2 (Persuasive writing)	
	Yes (%)	No (%)	Yes (%)	No (%)
All 4 lessons were taught each week	60 (88)	8 (12) ^a	44 (65)	24 (35) ^b
Some lessons were not taught^c	27 (40)	-	7 (10)	-
The lessons broadly followed the detailed plans	61 (89)	7 (10)	10 (15)	-
All the activities were undertaken	50 (73)	-	40 (59)	-
Most of the activities were undertaken	13 (19)	-	18 (26)	-
Several activities were undertaken	5 (7)	-	1 (2)	-

^a Includes two teachers who had not started to teach the unit.

^b Includes 17 teachers who had not started teaching Unit 2 due to the close proximity of SATs week.

^c 34 teachers reported skipping or combining lessons in Unit 1 (50%)

Note: Not all percentages equal 100 due to missing responses.

Observation and Interview Data

Lessons were observed to assess the extent to which the lessons were delivered as planned and the extent to which the materials were used in schools. The subsequent teacher interviews included discussion of the observed lesson.

In three of the eight schools visited, the materials were seen to have been adapted by the teacher or substituted with others. This was predominantly in terms of changes to PowerPoint slides although in one school a video used to provide information on Food Waste was substituted with another obtained via YouTube. However, in general, the activities in the lesson plans were retained with a mixture of teacher-led discussion and work by the children in pairs (or groups). Where the lesson plans were not adhered to this tended to be because of time constraints and the teacher moved quickly onto the next planned activity, for example because more time had been spent on an area in which the children were particularly engaged, or where they were struggling. In the interviews, some teachers mentioned that the lessons were 'pacey' and that this was a challenge:

I would say, originally with the Merlin unit, it was switching from thinking, 'It's too fast. It's too much,' to actually, 'Let's really up our game and see how the children react to it.' It's been a challenge, particularly for the middle and lower ability children, to rise up to the level of grammar required. But ... we need to set those high goals. Otherwise, it's all been incredibly positive. (Literacy Co-ordinator, School 63, Intervention)

In two observed schools, the planned lesson was not delivered in full in the time allocated and the teacher indicated that the lesson would be resumed in the next scheduled session. This lengthening of the units of work was also confirmed by one interviewee:

The scheme of work – the four-week unit took us six. But that was the editing, the redrafting, they typed it and, you know, but they absolutely loved it. (School 87 Teacher, Intervention)

Although the programme encouraged adaptations in order for teachers to differentiate within the classroom, activities tended to be delivered to the whole class with extension activities only offered to the pupils judged by their teachers as being more able at the end of the lesson once other tasks had been completed. Many teachers mentioned that they had linked the topic of Food Waste to children's own experiences as suggested by the scheme of work. In particular they had invited visitors to the classroom, primarily school canteen staff, to discuss food waste, and by making the letter-writing topic scheduled for the end of the unit relevant to their local area e.g. by writing to the local MP or the local paper. More detail on the extent and reasons for adaptations in programme delivery was given in the follow-up survey as detailed below.

The lesson observations were also designed to assess the extent to which programme delivery encompassed three core principles of the Grammar for Writing approach: use of grammar terms, linking grammar effects in writing, and using talk to develop discussion about choices and effects. More detail on this aspect of the lesson observations can be found below.

Survey Data

An open response question was included in the survey asking about the changes made to lessons. Seventy-three per cent of survey respondents indicated that adaptations to lessons had been made. The reasons given for these adaptations were grouped according to type, as shown in Table 17. Changes were described as 'schematic' when they referred to changes related to the planned delivery of the programme. The majority of these concerned the condensing or expanding of the time taken to deliver a particular element of the programme. Other adaptations included the use of more ICT (e.g. iMovie), changing the visual format to one that the children were used to, changing the format and structure of the handouts, and making the lessons more practical. A few teachers also referred to making the materials better suited to the demands of other, concurrent priorities, such as Assessment for Learning (AFL), the Grammar, Punctuation and Spelling (GPS) test, the interim framework for KS2 writing, and the National Curriculum. One respondent commented:

We tried to make the learning objectives more specific as some were very long and not in line with our perceived best practice of using specific, non-contextual learning objectives. (Clarke, 2013) (Teacher, School 6 Intervention; reference provided by respondent)

The type of changes designated as the 'addition of new elements' included building in peer- and self-assessment (even though elements of this are included in the programme), adding resources to increase/enrich children's subject knowledge, replacing some content to make it more relevant to the children's experience, and introducing more practical elements such as drama and art based activities 'that got the children moving instead of sitting and listening'.

Table 17: Type of changes made to the programme materials and/or recommended delivery (by teacher)

	N (%*)
Schematic	35 (38)
Adaptation to suit lower ability and/or SEN pupils	25 (27)
Adaptation to suit higher ability pupils	3 (3)
Adaptation to suit unspecified/multiple abilities	5 (5)
Addition of new elements	20 (22)
Other	3 (3)

* Total does not equal 100% due to rounding.

Schematic changes and the addition of new elements were all described in terms of meeting pupils' needs (albeit in some cases within larger-scale school plans). More explicitly 35% of respondents who indicated that they had made adaptations indicated that these were to meet the ability levels of their pupils, in particular the needs of lower ability and SEN pupils:

[I] tweaked some of the resources for the less able, provided further word banks etc., to support NTE and EAL pupils. (Teacher, School 136, Intervention)

[I] made the texts easier to understand. The concept was slightly out of some of the children's depths. Had to simplify some of the material. (Teacher, School 138, Intervention)

Consequently, although the majority of survey teachers (87%) stated that they strongly agreed or agreed that the programme materials were suitable for their pupils, and little differentiation was observed in the classroom visits made by researchers reported above, 20% of all survey respondents in the intervention condition stated that they adapted these materials to meet their SEN and lower ability pupils' needs. A further six per cent of the total respondents stated that they adapted the programme for higher ability pupils or for unspecified pupil abilities. This is particularly important given the large number of interviewees who indicated that they were originally attracted to the programme to meet the needs of their pupils or groups of pupils within their school. In addition, whilst non-programme related issues were not directly probed in the survey or interview questions, a few respondents referred unprompted to the counter-influence of the children's socio-cultural context, particularly in relation to the teaching of grammar, and this is clearly an important factor to take into account when considering programme implementation:

The most challenging aspect of it ... is really trying to change children's view or the way in which they talk in general ... children find it very hard sometimes to write in a way that they don't speak themselves, especially the use of tense for example - 'I done that'. No, you did that. So it is almost retraining them. That has been a challenge, I have to be honest ... (Teacher, School 136, Intervention)

In general, similar levels and types of adaptation to those reported in the teacher survey were witnessed in the lesson observations. Although it should be noted that the programme allows for adaptation to meet pupils' needs, the detail provided by teachers was insufficient to assess whether these adaptations were sufficiently extensive to compromise fidelity. For example, in discussion with the developer it was felt that adapting programme materials could be an acceptable adaptation if it was related to the text

being used. More extensive changes would, however, be more likely to impact on fidelity. Likewise, not completing all activities was deemed acceptable if it was due to pupils having already mastered the skill taught within that activity, although if activities were missed due to a general lack of time then fidelity would be compromised. The fidelity section below does suggest that the level of pupil discussion surrounding decisions and choice-making was compromised during the programme delivery and this may have been as a result of such adaptations, in particular where lessons were combined or aspects of the programme shortened or not delivered. In addition, where materials were adapted to fit the current class or school context, this may have made the delivery more like 'teaching as usual'. The number of adaptations made may also have affected overall levels of fidelity.

Finally, it should also be noted that there was little evidence of differentiation within observed classes, although the survey indicates that this was a key reason for adapting the programme and this was encouraged by the programme itself. This suggests that no adaptations were made for these pupils in the observed classes or that any adaptations that were made (i.e. changing lesson materials) were for whole class delivery rather than targeted at specific groups of pupils, either of which may have had implications for overall student outcomes. The results of the subgroup analysis for lower achieving pupils indicates that more differentiation, or more direction within the programme for the types of adaptation for lower ability pupils that could be made, may have been needed. This is supported by evidence from the teacher interviews that more detailed guidance on ways of adapting the programme that are permissible without compromising the integrity of the programme would have been useful:

We haven't been able to deliver the whole programme as it stands, you know, and we've had to adapt. And, it was like, trying to get them there, to have that 500-word legend¹⁷, it was difficult for them to get there ... so [next time I would] maybe not have you know 500 words to get them through in that time. (Teacher, School 50 Intervention)

[M]assively engaging for the children. Particularly, I have to say, the more able group who already had some knowledge and some engagement with that topic. I think the teacher [of the other Year 6 class] who had the least able group ... she found it harder to get the ideas going and the spark of enthusiasm, and she spent longer preparing children, giving them that knowledge of things to write about. Whereas my group ... they've got that bank of what a story is, haven't they, in their heads? And they've been read to, and they've got that language almost inside them anyway. (Teacher, School 63 Intervention)

Implementation fidelity

For the Development Team, fidelity to the programme relates more to adherence to the pedagogical principles underlying the programme than to the practical activities as determined by the lesson plans. As discussed above, both the Development Team and Evaluation Team conducted lesson observations in intervention schools with a view to monitoring the implementation of these principles, in particular 'connections made between grammar and effect/purpose in writing' and 'discussion used to tease out thinking and choice-making'. The Evaluation Team also rated fidelity according to the use of grammatical terminology.

The Development Team observed Unit 1 lessons in 13 schools and the Evaluation Team observed Unit 2 lessons in eight schools, although only six schools had their lesson observations rated using the coding schedule. The remaining two lesson observations focused on the other aspects of the observation measure developed for the study as described above (i.e. the detailed dynamics of the programme, other writing/grammar techniques taught in Year 6 and the wider classroom context e.g., levels of pupil engagement). Table 18 presents the fidelity ratings obtained from those observations. As can be seen, whilst the Evaluation Team observed relatively high levels of fidelity the same was not true for the Development Team. This difference can be explained by the higher levels of expertise in the programme by the Development Team, although it could also be that Unit 2 was delivered more 'as

¹⁷ The task required 'no more than 500 words' rather than specifying 500 words be written.

intended' as it was the second unit of work and teachers had possibly become more familiar with the approaches embedded within the programme. A final factor could have been school selection with higher fidelity schools being more prepared to receive a visit from the Evaluation Team, whereas visits from the Development Team could have been seen as more supportive of programme delivery and therefore more welcomed by schools struggling to implement the programme.

However, it is evident from Table 18 that the key areas of 'connections made between grammar and effect/purpose in writing' and 'discussion used to tease out thinking and choice-making' were not observed to be being implemented with high fidelity in over half of schools (36% and 26% rated as delivering these aspects 'as planned', respectively). The area in which fidelity was perceived to be weakest was '*discussion used to tease out thinking and choice-making*', suggesting that the high-quality talk that is a key feature of the programme was not being effectively used in the classroom. It should also be noted that, given the possible selection effects mentioned above, and possible observation effects (i.e. the presence of an observer resulting in teachers being more likely to adhere to programme principles) it would be reasonable to assume that there were lower levels of fidelity in other intervention schools.

Table 18: Fidelity Rating of Intervention Schools from Lesson Observations

	Unit 1: Development Team [#]		Unit 2: Evaluation Team [*]	
	n/N (missing)	Count (%)	n/N (missing)	Count (%)
Connections made between grammar and effect/purpose in writing	13/13 (0)		6/8 (2)	
'as planned'	-	3 (23.1)	-	4 (66.6)
'partially as planned'	-	4 (30.8)	-	2 (33.3)
'rarely'	-	6 (46.2)	-	-
Discussion used to tease out thinking and choice-making	13/13 (0)		6/8 (2)	
'as planned'	-	2 (15.4)	-	3 (50.0)
'partially as planned'	-	8 (61.5)	-	3 (50.0)
'rarely'	-	3 (23.1)	-	-
Grammatical Terminology Used~			6/8 (2)	
'as planned'			-	5 (83.3)
'partially as planned'			-	1 (16.7)
'rarely'			-	-

[#] one school observed did not complete the primary outcome measure.

^{*} two early school observations were not graded in this way.

~ This aspect of the programme was not measured by the Development Team as it was felt to be already embedded in the programme and associated materials. Rather, the focus of the Development Team observations was on the use of 'quality talk'.

We explored whether the fidelity rating had an effect on pupil outcomes within 18 of the 19 schools (one school could not be included as it did not participate in the primary outcome measure). For this we estimated the same model as for the primary outcome analysis but substituted the treatment variable with the fidelity rating score. Also, since there were two different teams observing, we added a variable controlling for the team which was used as a direct effect as well as interacted with the fidelity rating. The estimated effect of higher implementation fidelity on the pupils' outcomes in effect size was $-.17$ (95% bootstrapped confidence interval: $-.25, -.10$; average number of students per run $M = 762.17$, $SD = 12.49$; interaction effect non-significant, 95% bootstrapped confidence interval: $-1.02, .72$). This would constitute a small, but negative effect of higher implementation fidelity. Since the ratings between the two teams were different on average (York scores being higher than Exeter scores) and it was not clear whether this was due to different schools being tested or different implementation quality across schools or units of work (1 & 2), we centred each team's ratings on their respective average and repeated the analysis, so that the absolute levels of the ratings were not taken into account, but rather only the relative ordering within each of the team's assessments. The size of the effect remained the same (effect size $-.18$; 95% bootstrapped confidence interval: $-.25, -.11$; average number of students per run $M = 761.89$, $SD = 12.30$). However, the results of this analysis have to be taken with extreme caution given the possibility of selection bias in obtaining the sample (i.e. those schools who arranged visits as part of the evaluation may have been different from those who did not).

Programme Outcomes

As seen in the Impact Evaluation, the intervention did not significantly improve children's outcomes in the intervention condition. Teachers were asked, however, in the teacher survey, the teacher interviews and the Impact Inventory collected by the Development Team about the impact of the programme on themselves and their own teaching, and on outcomes for their pupils.

Teacher outcomes

Teachers were asked in the survey a number of Likert-scale questions relating to the programme overall. As Table 19 demonstrates, teachers overwhelmingly appreciated the value of the programme, and enjoyed teaching it:

- 92% of intervention teachers surveyed 'strongly agreed' or 'agreed' that the programme was useful for their teaching;
- 88% 'strongly agreed' or 'agreed' that they enjoyed using the materials provided in the programme; and
- 87% 'strongly agreed' or 'agreed' that they liked the programme.

Table 19: Teacher feedback on programme overall (n=128)

Variable	Strongly Agree N (%)	Agree N (%)	Neither Agree or Disagree N (%)	Disagree N (%)	Strongly Disagree N (%)
The programme was useful for my teaching	67 (54)	48 (38)	8 (6)	2 (2)	-
I enjoyed using the materials provided in the programme	63 (50)	47 (38)	10 (8)	5 (4)	-
I liked the programme	65 (52)	46 (37)	10 (8)	3 (2)	1 (1)

The sample of intervention teachers who took part in the interviews were asked about whether the training had impacted on their confidence in their linguistic and subject knowledge. Responses were overwhelmingly positive, as the examples below indicate.

[I found the training] really useful. I think at the start of the year had someone asked me, "Is your GPS knowledge good?" I'd have said, "Yes, it's very good", but having spent time with these expert grammarians you realise in fact there are lots of areas where it's a bit hazy. (Teacher, School 99, Intervention)

I found the training really interesting. It's done a lot to give me more confidence in my knowledge of grammar [and] helped us all around our subject knowledge and the way we explain things to the children. (Teacher, School 63, Intervention)

I think it's the best training I've ever been on in terms of literacy and grammar ... I feel much more confident – when speaking to staff as well ... and explaining it in a way that they can then teach it ... (Literacy Coordinator, School 87, Intervention)

There was just one teacher, recently qualified, in whom the training was reported to have induced a degree of anxiety, although the literacy coordinator felt that these anxieties were misplaced:

One of our teachers is an RQT [recently qualified teacher] [and] in terms of making a different plan your own, I think she found that quite tricky to start with ... she was a bit apprehensive, especially about the grammar terminology. She felt her grammar understanding wasn't as good, having not been teaching for very long ... but I think she has done fine with it. She has got some good writing outcomes as well. (Literacy Coordinator, School 90, Intervention)

A group of questions relating to attitudes towards the teaching of grammar were also administered in the survey at both baseline (T1) and follow-up (T2). Table 20 shows that, whilst there was a more modest increase in confidence in knowledge and understanding of grammar and confidence in teaching grammar at follow-up among the intervention group than seen in the teacher interviews, the control group also increased their confidence levels to a similar degree. This finding aligns closely with the findings from the teacher grammar quiz where no significant improvements in teacher grammar knowledge were found despite the Grammar for Writing training compared to the control group. Rather, the control group demonstrated slightly higher levels of improvement in their grammar knowledge during the academic year. Therefore, in the absence of any reporting of additional training being received by control group teachers in grammar, this perceived increase in confidence in grammar knowledge and teaching by the intervention group can be interpreted as part of the general improvement in confidence experienced by teachers during the school year as their experience of teaching grammar in Year 6 increased. Therefore, the training appears to have had no impact on teacher grammar knowledge, as measured by the grammar quiz, calling into question the value of this aspect of the training, although this finding is caveated by the low psychometric properties of the quiz noted elsewhere in this report.

Table 20 reports on teacher attitudes towards teaching grammar and writing at the start and end of the evaluation period:

- 65% of intervention teachers strongly agreed or agreed that '*It is important to teach grammar as a discrete subject to ensure that children grasp the necessary concepts*' at the start of the study compared to 61% at the end of the evaluation. The corresponding figures for the control group were 65% and 66%, respectively.
- 93% of intervention teachers strongly agreed or agreed with the statement '*I integrate grammatical concepts into all of my literacy teaching*' at the start of the study compared to 95% at the end. The corresponding figures for the control group were 95% and 93%, respectively.

This suggests that there was no effect of the programme in terms of attitudes towards teaching grammar in context in line with the grammar for writing approach, with little overall difference between the intervention and control groups.

Table 20: Pre- and post-intervention attitudes towards the teaching of grammar: Intervention and Control

		I feel confident in my own knowledge and understanding of grammar N (%)		I feel confident teaching grammar to Year 6 N (%)		It is important to teach grammar as a discrete subject to ensure that children grasp the necessary concepts N (%)		I integrate grammatical concepts into all my literacy teaching N (%)	
		Int	Con	Int	Con	Int	Con	Int	Con
Strongly Agree n (%)	T1*	26 (21)	22 (21)	28 (22)	31 (29)	22 (18)	23 (22)	52 (42)	44 (42)
	T2#	33 (26)	31 (29)	43 (34)	39 (36)	29 (23)	22 (21)	63 (49)	39 (36)
Agree n (%)	T1	79 (63)	69 (65)	70 (56)	56 (53)	59 (47)	46 (43)	64 (51)	56 (53)
	T2	86 (69)	65 (61)	80 (64)	61 (57)	47 (38)	48 (45)	59 (46)	61 (57)
Neither Agree/Disagree n (%)	T1	14 (11)	12 (11)	18 (14)	16 (15)	22 (18)	23 (22)	6 (5)	6 (6)
	T2	5 (4)	9 (8)	-	5 (5)	26 (21)	19 (18)	3 (2)	5 (5)
Disagree n (%)	T1	6 (5)	3 (3)	9 (7)	3 (3)	19 (15)	13 (12)	3 (3)	-
	T2	-	1 (1)	1 (1)	1 (1)	19 (15)	17 (16)	4 (3)	1 (1)
Strongly Disagree n (%)	T1	-	-	-	-	3 (2)	1 (1)	-	-
	T2	1 (1)	-	1 (1)	-	4 (3)	-	-	-

*T1=Baseline survey

#T2=Follow-up survey

The Impact Inventory similarly asked teachers about any changes in their teaching writing practice as a result of the programme. As can be seen in Table 21, the programme had the most reported impact on 'working with authentic texts' 'teaching grammar' with 47% and 49% of respondents reporting significant change, respectively. However, 66% of respondents indicated only some change in the 'use of talk' in their practice. This supports the finding from the lesson observations that the use of discussion and quality talk in lessons was not generally taught as planned and was a key area of non-conformity with programme principles.

Table 21: Teacher reported changes in practice

	Significant Change Things I did not know/do before N (%)	Some Change I have a new awareness of previous practice N (%)	No Change Existing practice has been affirmed N (%)
Use of talk	10 (15)	45 (66)	13 (19)
Working with authentic texts	32 (47)	24 (35)	12 (18)
'Teaching grammar'	33 (49)	33 (49)	2 (3)

Pupil outcomes

The Impact Inventory asked teachers about changes in student outcomes as a result of the programme. As can be seen in Table 22, whilst a large proportion of teachers reported significant change in pupils' subject knowledge and writing outcomes (46% and 43%, respectively) the majority reported only some change for each item. Interestingly, over two-thirds of respondents indicated only some change in 'explaining the effect in their own text' (76%) and 'explaining effects in mentor texts' (70%) which are the two items most closely related to pupils' discussion used to tease out thinking and choice-making, the aspect of the programme teachers' were observed to use less in Grammar for Writing lessons than intended by the programme.

Table 22: Teacher reported changes in student outcomes

	Significant Change N (%)	Some Change N (%)	No Change N (%)
Subject knowledge^a	31 (46)	36 (54)	-
Writing outcomes^a	29 (43)	38 (57)	-
Explaining the effect in their own text^b	12 (19)	49 (76)	3 (5)
Explaining effects in mentor texts^c	17 (29)	43 (70)	1 (1)

^a One teacher didn't respond

^b Four teachers didn't respond

^c Seven teachers didn't respond

Perceptions of pupil impact as a result of the Grammar for Writing programme were also gathered from the interviewees. Although the impact evaluation has shown no significant impact of the programme on pupil outcomes, interviewee responses were very positive. Some interviewees felt that they had seen an impact on their pupils' work as a result of the programme, and sometimes linked this to the level of pupil engagement.

The learning outcomes – verbally initially – were great to see. The children actually wanted more, and they didn't feel that they were being taught grammar. They were more interested in making the story [and] becoming better writers ... the real impact will be when they produce their [SATs] but in the interim period, looking at their books, and what they have produced, I would say that there's a lot of learning gains for the children. (Literacy Coordinator, School 50, Intervention)

If you have a look at the quality of the [pupils'] writing, it has really improved. This current Year 6 historically throughout the school have been under-performing and ... this seems to have worked really well. (Literacy Coordinator, School 136, Intervention)

For one teacher this was felt to be particularly the case for the higher ability pupils:

The children that have benefitted the most have been our higher ability children. Based on the Arthur work, their writing was just amazing - those pieces of writing were just really, really good, much better than anything else they have done. (Literacy Coordinator, School 90, Intervention)

Most teachers felt that pupils were engaged by the programme:

They've been extraordinarily motivated by it. They've been very engrossed by it, and they've really bought into it. (Literacy Coordinator, School 63, Intervention)

I think for us, well for me, the useful is to actually engage the children, because they have been very much engaged, they have loved this unit... You have chosen topics which are very engaging. (Teacher, School 50, Intervention)

Consequently, the overwhelming majority of survey respondents from the 65 intervention schools (n=111, 89%) said they would use the programme again.

However, not all teachers felt that the programme was having an impact on the quality of their pupils' work:

Stronger writing was produced by the children when not following the scheme of work. (Teacher, School 40, Intervention)

I think it worked in a very close way to how we worked before, so I don't think it's had a massive impact, but it's had lots of enjoyment. (Teacher, School 98, Intervention)

Of those teachers that did not plan to use the programme again (n=14, 11%) – the majority of reasons given tended to relate to issues of using programmes more generally.

I think that lessons created from one's own ideas are always more effective than following a prescribed scheme of work. (Teacher, School 139, Intervention)

The schemes ended up being more work for me as a teacher having to completely re-plan everything using the objectives than it would have been for me to plan from scratch. (Teacher, School 88, Intervention)

In contrast to the teachers interviewed, a small number of survey respondents also indicated that a lack of pupil engagement meant that they did not intend to use the programme again. However, these comments come from a very small proportion of the overall sample (n=3 out of 14 survey respondents who indicated that they didn't intend to use the programme again):

I wish you'd watched yesterday's lesson; they were so engaged in it, they were really excited yesterday. But then I think when you bring it back down to the grammar again they're a bit like... more writing, but yeah. (Teacher, School 92, Intervention)

My children found the topic choice boring and it did not excite them to write about food wastage. (Teacher, School 88, Intervention)

Only two survey respondents (out of 14) who indicated that they did not intend to use the programme again gave practical reasons for this decision i.e. they were moving school or year group.

Amongst the interview respondents who indicated that they would be using the programme again, some planned to make adaptations, or to select elements of it that were felt to be more appropriate or effective (as discussed in the section on Implementation above), others planned to use the programme with earlier Year groups as they felt it would be more beneficial if introduced earlier in KS2:

We feel that we need to start right back, so the teachers that have been on this course, we are going lower down in the school now ... to start the process earlier, because that is one thing that I would say is, this needed to be done in Year 5, because I am sure that [our classes] would have, not just a greater understanding but more confidence to be able to go yes I know this, I can spot the subject, I can spot the verb, I can ... even looking, they will say oh I can spot the determiners. But if they did it in Year 5, it would make it – not easier but ... they [would] have got time to consolidate, that is exactly it. (Teacher, School 136, Intervention)

It doesn't need to just be year 6 Merlin, we could adapt that to a year 5 text. ... if we can get that knowledge filtered down through school and that sort of style of teaching grammar then I think by the time the children are getting up to year 6 it will have had more of a long-term impact. (Literacy Coordinator, School 92, Intervention)

The timing of the programme as a whole was also seen to be problematic, in particular the timing of the second unit of work, coming as it did near the end of the school year, when the focus in Year 6 was on preparation for the KS2 SATs assessments:

... and again, maybe placing it at a time where, maybe right at the start of the term, rather than like at the term just before getting ready for SATS. (Teacher, School 50, Intervention)

This suggests that the programme as a whole may not have had the time needed to become embedded in pupils' practice.

Control group activity

As indicated above, there were similarities in the attitudes of the intervention and control group teachers surveyed towards teaching grammar and writing and in their confidence in their grammatical knowledge and ability to teach grammar to their Year 6 pupils, both at the start and the end of the study. In order to further understand the outcomes of the impact evaluation, interviewees in intervention schools were asked about their approach to teaching writing in the previous academic year and those in control schools were asked about their current approach. In both intervention and control schools, using real texts supplemented with separate, explicit grammar instruction in preparation for the GPS assessment was mentioned and this tended to be linked to topic work. Just as the majority of teachers in the survey strongly agreed or agreed that they integrated grammatical concepts into all their literacy teaching (93% intervention teachers, 95% control teachers in the baseline survey. See Table 17), a few teachers (n=3), spoke explicitly in the interviews about integrating grammar teaching into their writing teaching.

We do a genre of writing, it might be narrative or non-narrative writing, and then we'll do as far as possible grammar in the context of that. (Teacher, School 73, Control)

One teacher discussed how Grammar for Writing matched their existing practice:

I liked the idea of Grammar for Writing because that's the way that we tend to work anyway ... in that we never teach grammar in isolation, it's always relevant and ... the children can then use it for that particular text type that they're building up to on that learning journey. (Teacher, School 98, Intervention)

As indicated in the discussion on school motivations for participating in the study, the attractiveness of the approach and its 'fit' with existing practices were all attractions of the Grammar for Writing programme. In addition, control schools in particular were interested in taking part in research, although the numbers expressing this view were small.

In the baseline survey, all respondents were asked if they planned to use a scheme of work for their writing teaching in the academic year 2016-2017. The majority of respondents (77%) indicated that they did intend to use a programme or programme(s) in the academic year 2016-2017. This is unsurprising given that at this point schools were unaware of their allocation to control or intervention groups. However, intervention schools were more likely to report planning to use at least one programme than control schools; 81% and 29% respectively.

Table 23 shows the main programmes teachers indicated that their school both planned to use (as reported in the baseline teacher survey) and then actually used during the academic year 2016-2017 (as reported in the post-test survey), excluding the Grammar for Writing programme.

Table 23: Main programmes planned to use and used 2016-2017 (excluding Grammar for Writing)

	Planned		Used	
	Intervention N (%)	Control N (%)	Intervention N (%)	Control N (%)
The National Literacy Strategy – Grammar for Writing	13 (20)	15 (24)	13 (20)	15 (24)
Upper Key Stage 2 New Curriculum English Plans	9 (13)	11 (18)	9 (13)	11 (18)
Read, Write Inc Literacy and Language	9 (13)	4 (6)	9 (13)	4 (6)
Big Writing Adventures/Big Write	5 (7)	4 (6)	5 (7)	4 (6)
Grammar and Spelling Bug	4 (6)	4 (6)	4 (6)	4 (6)
Babcock No Nonsense Spelling and Grammar	2 (3)	1 (2)	3 (4)	1 (2)
Nelson English Skills	1 (2)	2 (3)	0 (0)	0 (0)
Nelson Grammar	2 (3)	3 (5)	0 (0)	0 (0)
Pearson Wordsmith	1 (2)	2 (3)	1 (2)	2 (3)
Power of Reading	0 (0)	5 (8)	0 (0)	5 (8)
Talk for Writing	3 (5)	2 (3)	2 (3)	2 (3)
Wordblaze	0 (0)	1 (2)	0 (0)	2 (3)

Other programmes were mentioned by only one school per condition at the most. These included; Hamilton Trust Lesson Plans, Scholastic Grammar, Writing and Punctuation, Literacy Shed, and Pearson Key Language.

As indicated in Table 23 above, literacy teaching in the control schools was very much ‘as usual’ during the experimental period. The wide range of teaching programmes outlined as planned were, with minimal variation, delivered in practice, indicating that control schools did not adopt a new programme to meet their perceived needs in teaching writing after they were allocated to the control condition. The programmes control schools reported using also closely matched those used by the intervention schools, other than the Grammar for Writing programme. The National Literacy Strategy – Grammar for Writing booklet, the most frequently reported programme used by both intervention and control schools, shares many of the principles of the Grammar for Writing programme itself, encouraging as it does embedding grammar teaching within teaching writing and discussion of writers’ choices in creating different effects in writing. Similarly, the Upper Key Stage 2 New Curriculum English Plans indicate that

pupils should be able to select 'appropriate grammar and vocabulary, understanding how such choices can change and enhance meaning'. Read write Inc. Literacy and Language also purports to teach grammar in context. All three of the more frequently reported programmes used in Table 19 therefore share some of the approaches systematically applied in the Grammar for Writing programme as evaluated in this study which may have moderated the overall impact of the programme. Interestingly, intervention schools also mentioned using more programmes in total than control schools, not including the Grammar for Writing programme (16 programmes across 65 schools, 25% compared to 11 programmes across 62 schools, 18%). This may have resulted in less consistency in writing teaching across the academic year as a number of different teaching approaches may have been adopted.

Four of the five control school lesson observations were of writing lessons, three of which were focused on persuasive writing (i.e. the same focus as Unit 2 of Grammar for Writing in the intervention schools). The fifth lesson observation focused primarily on grammar teaching. In one school the writing lesson was based on the Talk for Writing programme and in one Big Writing Adventures from Writing Owl was being used.

Table 24 shows the extent to which elements of the Grammar for Writing programme used to determine fidelity to the programme in the intervention schools were observed to be present in the control school lessons. As can be seen, in one of the four schools graded in this way discussion was often used to tease out thinking and choice making and in another connections between grammar and effect were also made. Elements of all three components were also present in these lessons, suggesting that instruction in the control schools was similar to that found in the intervention schools. In addition, given the number of adaptations reported by intervention schools the programme may have been diluted to result in similarities in this 'usual practice'.

Table 24: Use of core components used in Grammar for Writing in control school lesson observations.

School ID	Grammatical Terminology Used	Connections made between grammar and effect/purpose in writing	Discussion used to tease out thinking and choice-making	Total Rating
122	2	2	3	7
73	2	3	1	6
66	2	2	2	6
21	2	1	1	4
Total	6	7	6	

* The observed lesson with the focus on grammar was not graded in this way.

Only one control classroom had evidence of wider links to topic work within the school and only one classroom used peer assessment of writing. Four of the lessons did use pair/group work within the lesson. Three out of the five lessons also demonstrated the use of extension activities and differentiation within the lesson with two of the lessons explicitly providing different tasks to different groups of pupils (according to ability). This suggests that the control schools may have been better at meeting the needs of their different ability level pupils within their classes than the intervention schools where little within-

class differentiation was observed, although the survey did report a substantial amount of adaptation of the programme for differentiation purposes.

Conclusion

Key conclusions

1. The project found no evidence that Grammar for Writing improves writing attainment for children in Year 6, as measured by the bespoke test.
2. The project found no evidence that Grammar for Writing improves reading, writing or grammar, punctuation and spelling (GPS) as measured by KS2 SATS. Indeed, it found a small, negative effect size (equivalent to one month less progress) for the GPS outcome.
3. Pupils that have ever been eligible for free school meals made a small amount of additional progress compared to similar pupils in control schools. This result is not statistically significant. This means that the statistical evidence does not meet the threshold set by the evaluator to conclude that the true impact was not zero.
4. Grammar knowledge as measured using a teacher quiz did not improve for teachers who had done Grammar for Writing, although there was some evidence that this quiz was not a reliable measure. In contrast, more than 90% of surveyed teachers agreed that they found the programme, training and materials useful in their teaching
5. Nearly three-quarters of intervention teachers indicated that they had adapted the programme for delivery. In addition, fidelity to two of the key programme principles, 'connections made between grammar and effect/purpose in writing' and 'discussion used to tease out thinking and choice-making' was regarded by the evaluator to be compromised in a number of the schools observed.

Interpretation

This study was a two-armed effectiveness trial designed to assess the effectiveness of the Grammar for Writing programme in improving Year 6 pupils' writing skills. One hundred and fifty-five schools participated in the evaluation (7,239 pupils). Seventy-eight schools were randomised to receive the intervention and 77 were randomised to the control condition. There were twenty schools from whom primary outcome data was not collected. The main analysis was based on 135 schools and 5,415 pupils. The schools were balanced on the majority of baseline demographic characteristics, although teachers in the control conditions tended to have more years experience of teaching than those in intervention schools.

The results of the analysis found no evidence that the children in the intervention schools had better writing skills at the end of Year 6 as measured by the primary outcome than children in the control condition. In addition, there was no evidence of a statistically significant effect for children in receipt of FSM in the intervention schools. No effects were found for the secondary KS2 writing assessment and KS2 reading assessments. A small, negative effect of one-month progress when compared with the comparison group was found for the KS2 Grammar, Punctuation and Spelling assessment. As a secondary outcome analysis this result could nevertheless constitute a false-positive finding.

In addition, a potential group difference found was for previous achievement. Pupils who performed equal to or above the sample median in the pre-test were not observed to have benefitted from the intervention. However, for the lower performing pupils a potential small negative effect was found (ES=-0.11; equivalent to 2 months' progress), although this, again, was a secondary outcome analysis and could therefore constitute a false-positive finding (see below for issues relating to differentiation for lower performing pupils). No group difference was found based on gender.

Teachers completed an on-line grammar quiz pre- and post-intervention. The results were used firstly, to test whether or not the intervention led to improved grammar subject knowledge. It was found that there was no evidence of an improvement in teacher grammar knowledge as a result of the intervention. Secondly, the post-test was investigated as a potential mediator for a treatment effect i.e. whether or

not teachers' grammar knowledge, for which the post-intervention quiz acted as a proxy measure, mediated the relationship between the intervention and pupils' KS2past writing outcomes. Based on this specific measure, teachers' grammar subject knowledge was not found to be a mediating factor in children's primary writing outcome. However, it should be noted that the measure was not developed as a validated instrument and failed in several psychometric and statistical tests to indicate that it was a valid representation of teachers' inter-individual differences in grammar knowledge as well as potential change in grammar knowledge. Therefore, these results must be treated with caution.

A potential barrier to participating and delivering the programme was attendance at training. This was due to the time commitment involved and associated costs (including opportunity cost) of releasing all Year 6 teachers from the classroom for 3-4 days training over the school year. Inability to attend training was the main reason given by schools for withdrawal from the programme (40%) and of those schools who did participate approximately a quarter of teachers did not attend all 3 of the core training days deemed to be compliance by the programme developer. However, the compliance analysis reported above, which used attendance at CPD training days as a proxy measure, indicated that attendance at training did not result in higher pupil outcomes, did not improve knowledge of grammar, or change attitudes to teaching writing.

Analysis of the process evaluation data, which included a baseline and follow-up teacher survey, lesson observations, teacher interviews and teacher end of programme evaluation forms (Impact Inventories), indicated that teachers in the intervention condition in general found the programme, training and materials useful. A small number, of teachers queried the usefulness of the materials for their pupils (7% of survey respondents). However, a considerably larger number of survey respondents (73%) indicated that they adapted the programme materials and/or programme delivery suggesting that they did not meet the needs of their pupils or the school context. Such changes included lengthening the delivery period, shortening the delivery period and adapting the programme to meet the specific needs of some, or all pupils. The need for more differentiation within the programme delivery was a particular focus of the teacher feedback received (in both the survey and the teacher interviews). The process evaluation suggests that teachers would welcome more guidance on acceptable adaptations to the programme and differentiation within the classroom, particularly for lower ability pupils. This is supported by the potentially small, negative effect sizes found for lower ability pupils in the intervention condition when compared with their peers in control schools.

Although it should be noted that the programme allows for adaptation to meet pupils' needs and provides pointers for differentiation, the detail provided by teachers in the survey was insufficient to assess whether the adaptations reported by teachers were sufficiently extensive to compromise fidelity. The fidelity assessment focusing on two of the core principles of the programme 'pupil discussion surrounding decisions and choice-making' and 'connections made between grammar and effect' suggests that these approaches were compromised during the programme delivery. This may have been as a result of the adaptations made, including the number of adaptations made, to the programme. Although the exploratory analysis reported in the Process Evaluation suggests that implementation fidelity did not impact on pupil outcomes these findings must be interpreted with caution and the low levels of fidelity found by the Development Team in particular should not be ignored when interpreting the overall results of this study. Two of the core components of the programme ('connections made between grammar and effect/purpose in writing' and 'discussion used to tease out thinking and choice-making') do appear to have been compromised in this study which, taken with the number of adaptations made by teachers questions the extent to which the programme evaluated was that as planned by the developer.

The school visits suggest that more differentiation was occurring in the control classes observed than in the intervention classes although the sample was extremely small. When attitudes towards teaching writing in Year 6 in the intervention and control conditions were compared at both baseline and follow up there was no significant differences, with the majority of all respondents agreeing (or strongly

agreeing) that they integrated grammatical concepts into all their literacy teaching (95% intervention, 93% control at follow-up). The lesson observations suggested that elements of the Grammar for Writing approach were partially embedded within the control schools and that alternative writing programmes used by both control and intervention schools also had similar approaches in terms of embedding grammar teaching within writing, encouraging pupil choice in writing, and encouraging high quality talk in the classroom. Where Grammar for Writing programme materials were adapted by intervention schools to fit class or school contexts, this may have diluted the programme content and central principles and made the delivery more like 'teaching as usual' (i.e. the control group). This is supported by the survey finding that high proportions of teachers in both the control and intervention conditions agreed that they embedded grammatical concepts in all their literacy teaching.

Although the schools were matched on baseline factors it should also be noted that the control teachers who participated in the survey were, in general, more experienced teachers, both in terms of number of years in the profession and in recent number of years teaching Year 6 which may have impacted on Year 6 writing teaching more generally. In addition, intervention schools were more likely than control schools to use other programmes in their writing teaching and to use a greater number. Given that Grammar for Writing is only delivered for a 6-week period using a particular teaching approach, the short period of delivery, and the number of other programmes used may have also had an impact on overall outcomes.

The previous EEF trial led to the conclusion that the Grammar for Writing programme lacks conclusive evidence of effectiveness at the primary phase (Torgerson & Torgerson, 2014). This RCT does not provide that conclusive evidence, as it indicates no significant positive effects of the programme. The delivery period was, however, small, the timing overlapped to some extent with the KS2 assessments and the process evaluation suggests implementation fidelity was compromised.

Limitations

There are a number of limitations to this study, in particular questions relating to the primary outcome measure and implementation fidelity.

The primary outcome measure was changed during the year of the evaluation which meant that those schools who withdrew (n=10) prior to this did not supply data for the main analysis. This means that the main analysis was, as in the developer-led RCT in secondary schools, per protocol as opposed to intention to treat, although an intention to treat analysis would have been unlikely to have resulted in significant effects in favour of the programme. More importantly the refusal of additional consent to undertake the primary outcome from a further 10 schools may have impacted on the main analysis if the withdrawn schools and/or schools that did not undertake the primary outcome (n=20 in total) were different from those that did consent. As seen in the missing data analysis, schools remaining in the study had, on average, lower proportions of pupils meeting the expected standards in Reading and Maths and lower proportions of EAL pupils but higher proportions of pupils with SEN than those that withdrew. There could also be unquantifiable differences, operating at a school organisational level. Either way, the levels of attrition from the main analysis were high (approximately 25%) and affect the security of the findings.

The change in the primary outcome during the evaluation year highlights a wider issue, that of assessing children's writing skills in general. The KS2 writing assessments currently consists of a teacher-assessed portfolio of writing and pupils are judged to be 'working towards', 'working at', or 'working above' the expected standard for the end of KS2. As it was not externally moderated or sufficiently graded it was deemed to be an inappropriate measure to assess children's writing skills as a primary outcome. Whilst the independently administered and externally moderated past-KS2 papers were deemed more suitable it must be recognised that these, as with other similar possible measures of writing, are taken under different conditions to those in which Year 6 children's writing is generally undertaken (i.e. under timed assessment conditions). In addition, they are not aligned to the current

national curriculum or the KS2 assessment rubrics teachers were using during the year, although this would hold true for both control and intervention schools. Nevertheless, all stakeholders agreed that the chosen approach was the most appropriate available at the time.

Similarly, the analysis of the on-line grammar quiz suggested that reliability of the quiz' scores was low, and the principal component analysis showed that it likely assessed multiple components instead of a single, strong score representing grammar knowledge, suggesting the measure was not useful in this instance. One factor within this may have been that the measure was administered on-line and therefore conditions were not controlled.

In addition, the intervention was delivered to Year 6 pupils, a year group where teaching is often focused on the KS2 SATs GPS assessment, which is decontextualised (and therefore in opposition to the central programme principles) and on the teaching of language features required for the KS2 SATs teacher-assessment of writing. Preparation for and administration of the KS2 SATs also meant that some schools did not implement the second unit of work, and meant in low teacher attendance at the fourth CPD training day,

Although there were a number of limitations to the evaluation, as detailed above, it is considered that the results can be considered generalizable to a wider population of Year 6 pupils. This is because issues relating to the primary outcome would have been shared by intervention and control schools alike and the RCT was an effectiveness trial: testing a scalable model of the programme under everyday conditions in a large number of schools.

Future research and publications

Any future research questions would relate to implementation of the Grammar for Writing programme and which adaptations are acceptable to the programme alongside which adaptations are deemed necessary by teachers in the classroom. Further research on the teacher management of discussion of grammatical effects would be valuable to address the weaknesses in implementation identified here. Other considerations possibly worth exploring would be if there was a greater impact on pupils with higher prior attainment, as found in the developer-led RCT and if a more embedded version of the programme, introduced earlier in the primary phase may have longer-term impacts.

Further publications are planned by the Evaluation Team on the results of this study and by the Development team on more detailed data collected relating to pupils' writing collected from a subsample of intervention schools.

References

- Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research*, 20(1), 40-49. DOI: 10.1002/mpr.329
- Babcock LDP. (2016). *Non Nonsense Grammar; A Complete Grammar Programme* (Oxford: Raintree) pp.72.
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*; Vol 1, Issue 1 (2015). <https://doi.org/10.18637/jss.v067.i01>
- Department for Education (DfE) (2017). *Schools, pupils and their characteristics: January 2017*. SFR 28/2017, 29 June 2017. Accessed at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/650547/SFR28_2017_Main_Text.pdf, 27 April, 2018.
- Costa, D.S. (2015). Reflective, causal, and composite indicators of quality of life: A conceptual or an empirical distinction? *Qual Life Res.* 2015 Sep;24(9):2057-65. doi: 10.1007/s11136-015-0954-2. Epub 2015 Mar 1. <https://www.ncbi.nlm.nih.gov/pubmed/25725599>
- Education Endowment Foundation (EEF) (2013). *Pre-testing in EEF evaluations*. Accessed at: https://v1.educationendowmentfoundation.org.uk/uploads/pdf/Pre-testing_paper.pdf, 17 January, 2017.
- Education Endowment Foundation (EEF) (2018). *Statistical analysis guidance for EEF evaluations*. March 2018. Accessed at: https://educationendowmentfoundation.org.uk/public/files/Evaluation/Writing_a_Protocol_or_SAP/EEF_statistical_analysis_guidance_2018.pdf
- Hanley, P., Böhnke, J. R., Slavin, R., Elliott, L., & Croudace, T. J. (2016). *Let's Think Secondary Science: Evaluation Report and Executive Summary*. London: Education Endowment Foundation. Retrieved from <https://educationendowmentfoundation.org.uk/evaluation/projects/lets-think-secondary-science/>
- Hedges, L. V. (2007). Effect Sizes in Cluster-Randomized Designs. *Journal of Educational and Behavioral Statistics*, 32(4), 341–370. <https://doi.org/10.3102/1076998606298043>
- Heymans, M.W., van Buuren, S., Knol, D.L., van Mechelen, W., & de Vet, H.C.W. (2007). Variable selection under multiple imputation using the bootstrap in a prognostic study. *BMC Medical Research Methodology*, 7:33.
- Hoffmann, T. C., Glasziou, P. P., Boutron, I., Milne, R., Perera, R., Moher, D., ... Michie, S. (2014). Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *BMJ (Clinical Research Ed.)*, 348, g1687. <https://doi.org/10.1136/bmj.g1687>
- Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A Program for Missing Data. *Journal of Statistical Software*; Vol 1, Issue 7 (2011). <https://doi.org/10.18637/jss.v045.i07>
- Huang, F. L. (2016). Using Cluster Bootstrapping to Analyze Nested Data With a Few Clusters. *Educational and Psychological Measurement*, 0(0)2, 297-3180013164416678980.
- Jones, S., Myhill, D. A., & Bailey, T. (2013). Grammar for writing? An investigation into the effect of contextualised grammar teaching on student writing. *Reading and Writing: An Interdisciplinary Journal*, 26(8), 1241-1263.

King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation. *American Political Science Review*, 95(1), 49–69.

Minimpy, 2013. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3317766/>

Myhill, D.A. Jones, S.M., Lines, H. & Watson A. (2012). Re-thinking grammar: the impact of embedded grammar teaching on students' writing and students' metalinguistic understanding. *Research Papers in Education*, 27(2), 1-28.

Myhill, D.A. Jones, S. & Watson, A. (2013). Grammar matters: How teachers' grammatical subject knowledge impacts on the teaching of writing. *Teaching and Teacher Education*, 36, 77-91.

Pituch, K. A., Murphy, D. L., & Tate, R. L. (2009). Three-Level Models for Indirect Effects in School- and Class-Randomized Experiments in Education. *The Journal of Experimental Education*, 78(1), 60–95. <https://doi.org/10.1080/00220970903224685>

R Core Team. (2016). R: A language and environment for statistical computing (Version 3.2.5). Vienna: R Foundation for Statistical Computing. Retrieved from <http://www.R-project.org>

Saghaei, M., & Saghaei, S. (2011). Implementation of an open-source customizable minimization program for allocation of patients to parallel groups in clinical trials. *Journal of Biomedical Science and Engineering*, 4, 734–739.

Schomaker, M., & Heumann, C. (2014). Model selection and model averaging after multiple imputation. *Computational Statistics & Data Analysis*, 71, 758-770.

Standards & Testing Agency (2017). Teacher assessment frameworks at the end of key stage 2. For use in the 2017 to 2018 academic year.

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/738789/2017_Teacher_Assessment_Frameworks_at_the_end_of_key_stage_2_WEBHO.pdf

Torgerson, D. J., Torgerson, C., Mitchell, N., Buckley, H., Heaps, C., & Jefferson, L. (2014). Grammar for Writing. Evaluation Report and Executive Summary. London: Education Endowment Foundation. Retrieved from

https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Campaigns/Evaluation_Reports/EEF_Project_Report_GrammarForWriting.pdf

Tracey, L., Elliott, L., Boehnke, J., & Bowyer-Crane, C. (2017). Grammar for Writing. Protocol.

https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Protocols/Regrant_-_Grammar_for_Writing_effectiveness.pdf

Wirth, R. J., & Edwards, M. C. (2007). Item Factor Analysis: Current Approaches and Future Directions. *Psychological Methods*, 12(1), 58–79. <http://doi.org/10.1037/1082-989X.12.1.58>

Wyse, D. & Torgerson, C. (2017). Experimental trials and 'what works?' in education: The case of grammar for writing. *British Educational Research Journal*, 43(6), 1019-1047.

Xiao Z., Kasim, A., Higgins, S.E. (2016) Same Difference? Understanding Variation in the Estimation of Effect Sizes from Educational Trials *International Journal of Educational Research* 77: 1-14 <http://dx.doi.org/10.1016/j.ijer.2016.02.001>

Appendix A: EEF cost rating

Cost ratings are based on the approximate cost per pupil per year of implementing the intervention over three years. More information about the EEF's approach to cost evaluation can be found [here](#). Cost ratings are awarded as follows:

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

Appendix B: Security classification of trial findings

Rating	Criteria for rating			Initial score	Adjust	Final score	
	Design	Power	Attrition *				
5	Well conducted experimental design with appropriate analysis	MDES < 0.2	0-10%		Adjustment for Balance [n/a]		
4	Fair and clear quasi-experimental design for comparison (e.g. RDD) with appropriate analysis, or experimental design with minor concerns about validity	MDES < 0.3	11-20%				
3	Well-matched comparison (using propensity score matching, or similar) or experimental design with moderate concerns about validity	MDES < 0.4	21-30%	3		Adjustment for threats to internal validity [n/a]	3
2	Weakly matched comparison or experimental design with major flaws	MDES < 0.5	31-40%				
1	Comparison group with poor or no matching (E.g. volunteer versus others)	MDES < 0.6	41-50%				
0	No comparator	MDES > 0.6	>50%				

- **Initial padlock score:** lowest of the three ratings for design, power and attrition = the design and MDES at randomisation were good, however, the necessity of changing the testing plans (i.e. testing pupils as opposed to using administrative data as was originally intended) meant that a number of schools (28%) opted out of testing.
- **Reason for adjustment for balance** (if made): n/a
- **Reason for adjustment for threats to validity** (if made): n/a
- **Final padlock score:** initial score adjusted for balance and internal validity = 3 padlocks

Appendix C: Information and Consent Forms

C.1 Head Teacher Memorandum of Understanding

THE UNIVERSITY *of York*



UNIVERSITY OF
EXETER

Grammar for Writing Randomised Controlled Trial Study

MEMORANDUM OF UNDERSTANDING

This project is designed to study the teaching and learning of writing in primary schools. The new approach, Grammar for Writing, aims to improve writing by developing pupils' understanding of grammatical choices. Its impact will be evaluated by comparing it with the "business as usual" approach using a randomised controlled trial (RCT).

During this project, you will be contacted by both the **Project Team** (University of Exeter), who are responsible for developing and supporting the new teaching approach, and by the **Evaluation Team** (University of York), 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. If you agree to take part and accept the terms and conditions outlined, please sign a copy of this form and return by email or mail to the contact provided at the end of this letter.

Randomised Controlled Trial (September 2016 – July 2017)

The trial will involve your school being *randomly* assigned either to deliver Grammar for Writing (the intervention group) or to continue with your normal teaching approach (the comparison group). Teachers in the intervention group will be asked to attend four training days across the year, and to deliver two Grammar for Writing units in the spring term, one on narrative writing (four weeks) and one on persuasive writing (two weeks). Schools in the intervention group will be asked to pay £500 to participate as a partial contribution to the costs of the 4 CPD days and the teaching materials. Schools in the comparison group will receive a £500 payment as a partial contribution towards buying in the CPD 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 contact details of a main contact person and of Year 6 teachers (valid email addresses and telephone numbers) to the Project Team.
- Provide names of teachers and details of classes (including UPNs), along with details of any setting or streaming by attainment, to the Evaluation Team by the end of the summer term.
- Facilitate the participation of teachers to complete a short on-line grammar quiz.

During the evaluation

Participating teachers will:

- Complete a short on-line grammar quiz and teacher survey at the end of year, and will receive a £10 gift voucher for successful completion.
- Update UPNs of Year 6 pupils by the end of September 2016 and contact details (if appropriate) during the course of the evaluation
- Facilitate a visit by the Project Team to observe one Grammar for Writing lesson of a sample of participating teachers (*intervention group only*).
- For a randomly selected sub-sample: facilitate a school visit by one or two researchers from the Evaluation Team to observe an English lesson (Grammar for Writing in schools trialling the new approach, a lesson focusing on grammar/writing in other schools) during the study year, followed by short discussions with some of the Year 6 teaching staff and a member of the senior management team.
- Provide a breakdown of the KS2 writing results on narrative and persuasive writing for each participating Year 6 pupil.

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). 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. Confidentiality will be maintained and no one outside the Evaluation Team will have access to the database. Identifying data will be retained for one year after the end of the evaluation and anonymised for a maximum of 3 years.

Requirement 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 6 writing.
- All the Year 6 pupils and their English teachers will participate in the project.
- Participating teachers will complete the training provided and seek help and advice from the Project Team if they have any queries or uncertainties about implementing Grammar for Writing.
- 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 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 prevent the effective use of the approach.

- ❑ If the school has to withdraw from the project for operational or other unavoidable reasons, it will notify the Evaluation Team straight away and, wherever possible, still provide test data for the evaluation.

Responsibilities of the Project Team:

- To set-up a training course to inform teachers on how to implement Grammar for Writing
- Act as the first point of contact for any questions about the evaluation
- Provide on-going support to the school
- Provide information sheets for parents
- Collect participating teacher and lead contact names and email details.

Responsibilities of the Evaluation Team:

- Conduct the random allocation
- Collect class and pupil level data (including name, date of birth, UPN)
- Request NPD data using pupil details
- Analyse the data from the project
- Disseminate the research findings

HEADTEACHER AGREEMENT

Please initial each box and sign below:

- I confirm that I have read and understood the information sheet for the above evaluation and have had the opportunity to ask questions;
- I agree to providing the data as specified in the attached information sheet and in the format requested by the Evaluation Team;
- I understand that failure to provide all the data specified as required prior to randomisation will prevent participation in the study. Any data already provided will then be destroyed by the Evaluation Team.
- I agree to the Evaluation Team obtaining data on the evaluation cohort's KS1 and KS2 results from the National Pupil Database;
- I agree to providing an information letter to all parents of children in Year 6 and to inform the Evaluation Team of any parental opt-out from the study;
- I agree to random allocation to implement 'Grammar for Writing' or continue 'teaching as usual';
- I understand that all data will be kept in accordance with the Data Protection Act (1998) and that no material which could identify individual children, teacher's or the school will be used in any reports of this evaluation;
- I agree to staff attending professional development days.

I agree for my school _____ to take part in the Grammar for Writing study and I accept the eligibility terms and conditions as described above.

Signature of Head Teacher: _____

Name of Head Teacher: _____ **Date:**
 ___/___/_____

PLEASE RETAIN A COPY FOR YOUR RECORDS AND RETURN A COPY TO: Name/email/postal address

C.2 Teacher Consent

THE UNIVERSITY *of York*



Grammar for Writing Randomised Controlled Trial Study

MEMORANDUM OF UNDERSTANDING

This project is designed to study the teaching and learning of writing in primary schools. The new approach, Grammar for Writing, aims to improve writing by developing pupils' understanding of grammatical choices. Its impact will be evaluated by comparing it with the "business as usual" approach using a randomised controlled trial (RCT).

During this project, you will be contacted by both the **Project Team** (University of Exeter), who are responsible for developing and supporting the new teaching approach, and by the **Evaluation Team** (University of York), 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. If you agree to take part and accept the terms and conditions outlined, please sign a copy of this form and return by email or mail to the contact provided at the end of this letter.

Randomised Controlled Trial (September 2016 – July 2017)

The trial will involve your school being *randomly* assigned either to deliver Grammar for Writing (the intervention group) or to continue with your normal teaching approach (the comparison group). Teachers in the intervention group will be asked to attend four training days across the year, and to deliver two Grammar for Writing units in the spring term, one on narrative writing (four weeks) and one on persuasive writing (two weeks). Schools in the intervention group will be asked to pay £500 to participate as a partial contribution to the costs of the 4 CPD days and the teaching materials. Schools in the comparison group will receive a £500 payment as a partial contribution towards buying in the CPD 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 contact details of a main contact person and of Year 6 teachers (valid email addresses and telephone numbers) to the Project Team.
- Provide names of teachers and details of classes (including UPNs), along with details of any setting or streaming by attainment, to the Evaluation Team by the end of the summer term.
- Facilitate the participation of teachers to complete a short on-line grammar quiz.

During the evaluation

Participating teachers will:

- Complete a short on-line grammar quiz and teacher survey at the end of year, and will receive a £20 gift voucher for successful completion.
- Update UPNs of Year 6 pupils by the end of September 2016 and contact details (if appropriate) during the course of the evaluation.

- Facilitate a visit by the Project Team to observe one Grammar for Writing lesson of a sample of participating teachers (*intervention group only*).
- For a randomly selected sub-sample: facilitate a school visit by one or two researchers from the Evaluation Team to observe an English lesson (Grammar for Writing in schools trialling the new approach, a lesson focusing on grammar/writing in other schools) during the study year, followed by short discussions with some of the Year 6 teaching staff and a member of the senior management team.
- Provide a breakdown of the KS2 writing results on narrative and persuasive writing for each participating Year 6 pupil.

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). 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. Confidentiality will be maintained and no one outside the Evaluation Team will have access to the database. Identifying data will be retained for one year after the end of the evaluation and anonymised for a maximum of 3 years.

Requirement 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 6 writing.
- All the Year 6 pupils and their English teachers will participate in the project.
- Participating teachers will complete the training provided and seek help and advice from the Project Team if they have any queries or uncertainties about implementing Grammar for Writing.
- 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 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 prevent the effective use of the approach.
- If the school has to withdraw from the project for operational or other unavoidable reasons, it will notify the Evaluation Team straight away and, wherever possible, still provide test data for the evaluation.

Responsibilities of the Project Team:

- To set-up a training course to inform teachers on how to implement Grammar for Writing
- Act as the first point of contact for any questions about the evaluation
- Provide on-going support to the school
- Provide information sheets for parents
- Collect participating teacher and lead contact names and email details.

Responsibilities of the Evaluation Team:

- Conduct the random allocation
- Collect class and pupil level data (including name, date of birth, UPN)
- Request NPD data using pupil details
- Analyse the data from the project
- Disseminate the research findings

TEACHER AGREEMENT

Please initial each box and sign below:

- I confirm that I have read and understood the information sheet for the above evaluation and have had the opportunity to ask questions;
- I agree to providing the data as specified in the attached information sheet and in the format requested by the Evaluation Team, including completing a short, on-line grammar quiz and on-line teacher survey in Summer 2016 and Summer 2017 and I will receive a £20 voucher at the end of the evaluation for doing so;
- I agree to providing an information letter to all parents of children in Year 6;
- I agree to implement 'Grammar for Writing' if randomly allocated to do so or continue 'teaching as usual';
- I agree to facilitate visits by the Project and Evaluation Teams, if requested;
- I understand that all data will be kept in accordance with the Data Protection Act (1998) and that no material which could identify individual children, teacher's or the school will be used in any reports of this evaluation.

I agree to take part in the Grammar for Writing study and I accept the eligibility terms and conditions as described above.

Signature of Teacher: _____

Name of Teacher: _____ **Date:** ___/___/___

Name of School: _____

PLEASE RETAIN A COPY FOR YOUR RECORDS AND RETURN A COPY TO: Name/email/postal address

C.3 Parent/Guardian Information Sheet

THE UNIVERSITY *of York*



September 2016

Dear Parent/Guardian,

We would like to ask permission for your child to take part in an educational research study. This study is being done to assess the effectiveness of Grammar for Writing, a new approach to teaching grammar to help pupils make informed grammatical choices in their writing. Teachers will undergo four days of training and deliver two sets of lessons, a four-week unit on narrative writing and a two-week unit on persuasive writing.

Your child's school has agreed to participate in the study. The units will be taught in the spring term of Year 6. Schools will be assigned at random to either use Grammar for Writing or to continue teaching in their usual way. **Your child has a 50% chance of being in a school that tries out the new approach.**

To judge the effectiveness of Grammar for Writing compared with schools' usual teaching, we will look at pupils' performance in the Key Stage 1 and Key Stage 2 English SATs. To do this, we will need to obtain SATs scores for your child from the National Pupil Database (held by the Department for Education) or from the school and share them with: 1) the Department for Education, 2) the Education Endowment Foundation (EEF) 3) the EEF data contractor and (in an anonymised form) 4) the UK Data Archive. No individual pupil's data will appear in any report about the research study.

Your child's data will be treated in the strictest confidence. It will be stored in accordance with the Data Protection Act and any individually-identifiable data will be destroyed by the end of 2018. We will not use your child's name or the name of the school in any report arising from the research. If you prefer your child's SATs scores **NOT** to be used, please complete and return the opt-out form to your child's teacher within a week of receiving this letter. If you are happy that we use your child's SATs scores for the purposes of this research, then you do not need to return the form.

Signing the opt-out form will mean that your child will still be taught using Grammar for Writing if they are in this group, but we will not use their SATs scores to evaluate the programme.

If you would like more information, please contact Louise Tracey (e-mail: louise.tracey@york.ac.uk Tel:01904 328160) or the Education Ethics Committee (education-research-administrator@york.ac.uk)

With thanks and best wishes

Louise Tracey (York Evaluation Team)
Debra Myhill (Exeter Project Team)

GRAMMAR FOR WRITING EVALUATION

Parent/Guardian opt-out form

If you do not permit your child's Key Stage SATs scores to be used in the study, please complete this form and return it to your child's teacher

I **do not** wish my child's test scores to be used in the research project.

Pupil's name:
(Please print clearly)

School name:

Class teacher

Parent's/Guardian's name:
(Please print clearly)

Parent's/Guardian's signature:

Date:.....

This form will be returned by the school to:

Louise Tracey,
Department of Education,
Berrick Saul Building,
University of York
YO10 5DD.
Email: gfw-evaluation-team@york.ac.uk

C.4 Head Teacher Supplementary Digital Consent

THE UNIVERSITY *of York*



Dear Headteacher

Thank you for your school's participation in the *Grammar for Writing* project: the project is progressing well and the teachers who have attended the training seem to have found it very useful. The intervention schools are now teaching or about to teach the second teaching unit, and many of the comparison schools have now signed up for their training.

We have had to make a minor modification to the data we collect. As you will be aware, last year's KS2 writing assessment was new and there was some confusion surrounding it, with the consequence that it is not reliable data for research. We are not yet confident that this year's data will be reliable. So instead of using this data, we are going to ask children to do a writing assessment in school which will be marked independently. The assessment will be after the KS2 tests so there will be no risk of any impact on those.

To reduce any pressure on your staff, the NfER will administer these assessments, arrange for someone to come in, set the assessment and take the scripts away on the day. In order for them to do so the project team will be sharing information with the NfER relating to the project, including school contact details and pupil names. All data will be treated as confidential and handled in accordance with the Data Protection Act. If you would like to keep the scripts for the portfolio assessment, we can arrange for you to keep a copy. Our goal is to capture an accurate assessment of writing, but to minimise any pressure on the children or any additional workload for your staff.

If you would like more information about the project, please follow this link:

<http://socialsciences.exeter.ac.uk/education/research/centres/centreforresearchinwriting/projects/grammarforwriting/>

As this is a change to our agreed plans, our ethical procedures require us to seek your consent to this below. We hope that you will agree as this is important data for the project. Please could you simply click on the appropriate box below.

YES - I CONSENT TO THIS DATA BEING COLLECTED

NO – I DO NOT CONSENT TO THIS DATA BEING COLLECTED

Thank you

Debra Myhill – University of Exeter

Louise Tracey – University of York

C.5 Parent/Guardian Supplementary Information Sheet

THE UNIVERSITY *of York*UNIVERSITY OF
EXETER

[Date]

Dear Parent/Guardian,

In the summer last year we sent you information about an educational research study which your child's school is participating in. It is assessing the effectiveness of Grammar for Writing, a new approach to teaching grammar to help pupils make informed grammatical choices in their writing.

We sent you information about the data we would be collecting (using the National Pupil Database). We have had to make a minor change to this because last year's national assessment of writing was new and the results are not reliable. So we will instead be arranging for an in-school test of writing, administered by the NFER.

Your child's data will be treated in the strictest confidence. For the purposes of test administration your child's name and UPN number will be shared with the NFER using secure servers. It will be stored in accordance with the Data Protection Act and any individually-identifiable data will be destroyed by the end of 2018. We will not use your child's name or the name of the school in any report arising from the research. If you prefer your child's in-school test scores **NOT** to be used, please complete and return the opt-out form to your child's teacher within a week of receiving this letter. You may also withdraw your child's data at any point prior to the end of July 2017 by returning this form or contacting Louise Tracey (details below). If you are happy that we use your child's test scores for the purposes of this research, then you do not need to return the form. Signing the opt-out form will mean that your child will still be taught using Grammar for Writing if they are in this group, but we will not use their writing test scores to evaluate the programme.

If you would like more information, please contact Louise Tracey (e-mail: louise.tracey@york.ac.uk Tel:01904 328160) or the Chair of the Education Ethics Committee at the University of York (education-research-administrator@york.ac.uk)

With thanks and best wishes,

Louise Tracey (York Evaluation Team)

Debra Myhill (Exeter Project Team)

GRAMMAR FOR WRITING EVALUATION**Parent/Guardian opt-out form**

If you do not permit your child's in-school test scores to be used in the study, please complete this form and return it to your child's teacher.

I **do not** wish my child's test scores to be used in the research project.

Pupil's name: (Please print clearly)

School name:

Class teacher

Parent's/Guardian's name: (Please print clearly)

Parent's/Guardian's signature: **Date:**

This form will be returned by the school to: **Louise Tracey, Department of Education**, Berrick Saul Building, University of York YO10 5DD. Email: iee@york.ac.uk

Appendix D: Technical Report

For a full appraisal of the results a more detailed presentation of the analyses is needed. Since the main report only presents headline findings which briefly summarise the main tests of the study hypotheses and presents only the statistics that are necessary for understanding the specific results, more detail on the statistical analyses is presented in this Technical Report. This Appendix reports on the affordances of reporting for the EEF, for example a more detailed presentation of missing data patterns and descriptives than described in the main body of the report, additional details for the appraisal of the hierarchical models (e.g., variance components, coefficients of the other involved variables in model estimation), and, finally, the necessary code to estimate and reproduce any of the analyses undertaken. Repetition has been kept to a minimum although where it has been deemed suitable some repetition does occur e.g. some repetition of details regarding how the models are estimated and which approach was chosen, for ease of reading.

D.1. Analysis of the primary outcome including imputation and subgroup analyses

In the following a detailed description of the models and results for the estimation of the primary outcome of the study is presented. Details on the analysed population were presented in the CONSORT diagram (Figure 2) in the main report. Overall, for the analysis of the raw data $N = 5182$ cases of pupil-level data were available and for the imputed analyses $N = 6306$. The main report and the Statistical Analysis Plan present details of the statistical model used to analyse the primary outcome.

The analysis was cluster-bootstrapped: From each school a random sample of the same size as its actual sample was drawn (with replacement) and across these school-wise bootstrap samples, the mixed model was then estimated. This process was repeated $b = 1000$ times and for a 95%-confidence interval the statistical estimates were saved and their top and bottom 2.5%-quantiles were identified. The average of the bootstrapped values was treated as the point estimate and is reported in the following tables. As stated in the statistical analysis plan (SAP), no p-values are reported for any analysis.

Since 14.1% of missing values were observed for the primary outcome in this dataset, the SAP stated that a sensitivity analysis of this result based on multiply imputed data would be carried out. Details of the imputation model are presented in the main report and are not repeated here.

1. Results for the primary outcome for all pupils

1.1. Primary outcome with available data

The estimated coefficients (averages across bootstrap runs) and their confidence intervals (2.5% and 97.5% percentiles of the distribution of bootstrapped values) are displayed in Table D.1.1. The average KS2past result for a student with a KS1 result at the grand mean (in control group in the North-East) was 16.80 KS2past points. The pre-test has an effect on the outcome: per point increase in KS1 on average an increase of 1.15 points in KS2past is expected. The effect of region was different from zero since the confidence interval did not include this value, i.e. a small regional effect on average KS2past performance was found with schools in the "other" regions (i.e. not North-East) doing on average .32 points worse.

The estimated coefficient of -0.14 states that the pupils in the intervention group showed on average KS2past scores that were 0.14 points lower than those reached in the control group. The confidence interval includes zero. Therefore the primary outcome analysis concludes that the treatment showed in

this study building on available schools no effect that could be statistically differentiated from "0". The Null hypothesis of "no effect" could not be rejected.

The associated effect size was $ES = -0.02$ with its 95%-bootstrap confidence interval ranging from (lower border) $LB_{ES} = -0.08$ to (upper border) $UB_{ES} = 0.03$. The estimated effect sizes are very small and the confidence interval includes 0, indicating that the Null hypothesis of "no effect" could not be rejected.

Table D.1.1. Bootstrapped coefficients for the primary outcome, observed data

Variable	Average Coefficient (SD)	95% Confidence Interval
Intercept	16.80 (.30)	16.24, 17.39
KS1 Writing Result	1.15 (.02)	1.10, 1.20
Region	-.32 (.17)	-.65, -.004
Treatment	-.14 (.16)	-.47, .16

Note. The analysis is based on $b = 1000$ bootstrap samples. The estimated intraclass correlations were for KS1 $ICC = .091$ ($SD = .008$) and for $KS2_{past}$ $ICC = .137$ ($SD = .010$); the average level-1 variance/ residual was $resid = 27.88$ and the average level-2 variance was $varl2 = 7.95$; $N = 5181.22$ ($SD = 29.67$), intervention $N = 2775.72$ ($SD = 20.34$), control $N = 2405.50$ ($SD = 21.18$)

1.2. Primary outcome with imputed data

The estimated coefficients (averages across bootstrap runs) and their confidence intervals (2.5% and 97.5% percentiles of the distribution of bootstrapped values) are displayed in Table D.1.2. The average KS_{past} result for a student with a KS1 result at the grand mean was 16.83 $KS2_{past}$ points. The pre-test has an effect on the outcome: per point increase in KS1 on average an increase of 1.15 points in $KS2_{past}$ is expected. The effect of region was different from zero since the confidence interval did not include this value, i.e. a small regional effect on average KS_{past} performance was found with schools in the "other" regions (i.e. not North-East) doing on average .38 points worse.

The estimated coefficient of -0.15 states that the pupils in the intervention group showed on average $KS2_{past}$ scores that were 0.15 points lower than those reached in the control group. The confidence interval includes zero. Therefore this sensitivity analysis for the primary outcome analysis concludes that the treatment showed in this study building on available schools no effect that could be statistically differentiated from "0". The Null hypothesis of "no effect" could not be rejected.

The associated effect size was $ES = -0.03$ with its 95%-bootstrap confidence interval ranging from $LB_{ES} = -0.08$ to $UB_{ES} = 0.02$. The estimated effect sizes are very small and the confidence interval includes 0, indicating that the Null hypothesis of "no effect" could not be rejected.

Table D.1.2. Bootstrapped coefficients for the primary outcome, imputed data

Variable	Average Coefficient (SD)	95% Confidence Interval
Intercept	16.83 (.31)	16.25, 17.45
KS1 Writing Result	1.15 (.02)	1.11, 1.19
Region	-.38 (.16)	-.69, -.07
Treatment	-.15 (.15)	-.45, .12

Note. The analysis is based on $b = 1000$ bootstrap samples. The estimated intraclass correlations were for KS1 $ICC = .093$ ($SD = .007$) and for KS2past $ICC = .136$ ($SD = .009$); the average level-1 variance/ residual was $resid = 27.79$ and the average level-2 variance was $varl2 = 7.81$; $N = 6306$, intervention $N = 3346$; $N_{control} = 2960$

1.3. Primary outcome with available data within FSM population only

The estimated coefficients (averages across bootstrap runs) and their confidence intervals (2.5% and 97.5% percentiles of the distribution of bootstrapped values) are displayed in Table D.1.3. The average KSpast result for a student (EVER_FSM) with a KS1 result at the grand mean was 15.85 KS2past points. The pre-test has an effect on the outcome: per point increase in KS1 on average an increase of 1.11 points in KS2past is expected. The effect of region is not different from zero since the confidence interval includes this value (i.e. no regional effects on average KSpast performance were found for the FSM-only pupils).

The estimated coefficient of 0.30 states that the pupils in receipt of FSM in the intervention group showed on average KS2past scores that were 0.30 points higher than those reached in the control group. The confidence interval includes zero. Therefore this subgroup analysis of the primary outcome analysis concludes that the treatment showed in this study building on available schools no effect that could be statistically differentiated from "0". The Null hypothesis of "no effect" could not be rejected.

The associated effect size was $ES = 0.05$ with its 95%-bootstrap confidence interval ranging from $LB_{ES} = -0.03$ to $UB_{ES} = 0.13$. The estimated effect sizes are very small and the confidence interval includes 0, indicating that the Null hypothesis of "no effect" could not be rejected.

Table D.1.3. Bootstrapped coefficients for the primary outcome for FSM population, observed data

Variable	Average Coefficient (SD)	95% Confidence Interval
Intercept	15.85 (.41)	15.05, 16.63
KS1 Writing Result	1.11 (.03)	1.04, 1.18
Region	-.24 (.24)	-.72, .23
Treatment	.30 (.23)	-.16, .74

Note. The analysis is based on $b = 1000$ bootstrap samples. The estimated intraclass correlations were for KS1 $ICC = .096$ ($SD = .012$) and for KS2past $ICC = .154$ ($SD = .016$); the average level-1 variance/ residual was $resid = 25.81$ and the average level-2 variance was $varl2 = 8.09$; average bootstrapped $N = 2360.61$ ($SD = 35.52$) $N_{intervention} = 1290.51$ ($SD = 27.08$); $N_{control} = 1070.09$ ($SD = 25.31$)

1.4. Primary outcome within FSM population only, imputed data

The estimated coefficients (averages across bootstrap runs) and their confidence intervals (2.5% and 97.5% percentiles of the distribution of bootstrapped values) are displayed in Table D.1.4. The average KSpast result for a student (EVER_FSM) with a KS1 result at the grand mean was 15.83. The pre-test has an effect on the outcome: per point increase in KS1 on average an increase of 1.11 points in KS2past is expected. The effect of region is not different from zero since the confidence interval includes this value (i.e. no regional effects on average KSpast performance was found).

The estimated coefficient of 0.15 states that the pupils in the intervention group showed on average KS2past scores that were 0.15 points higher than those reached in the control group. The confidence interval includes zero. Therefore this subgroup analysis of the primary outcome analysis concludes that the treatment showed in this study building on available schools no effect that could be statistically differentiated from "0". The Null hypothesis of "no effect" could not be rejected.

The associated effect size was $ES = 0.03$ with its 95%-bootstrap confidence interval ranging from $LB_{ES} = -0.05$ to $UB_{ES} = 0.10$. The estimated effect sizes are very small and the confidence interval includes 0, indicating that the Null hypothesis of "no effect" could not be rejected.

Table D.1.4. Bootstrapped coefficients for the primary outcome for FSM population, imputed data

Variable	Average Coefficient (SD)	95% Confidence Interval
Intercept	15.83 (.41)	15.00, 16.60
KS1 Writing Result	1.11 (.03)	1.05, 1.16
Region	-.19 (.23)	-.63, .26
Treatment	.15 (.22)	-.31, .58

Note. The analysis is based on $b = 1000$ bootstrap samples. The estimated intraclass correlations were for KS1 ICC = .096 (SD = .012) and for KS2past ICC = .143 (SD = .013); average bootstrapped $N = 2924.76$ (SD = .99) $N_{intervention} = 1567.38$ (SD = .86), $N_{control} = 1357.38$ (SD = .49); the average level-1 variance/ residual was $resid = 25.94$ and the average level-2 variance was $var2 = 7.90$; since for four cases the FSM status needed to be imputed, the sample sizes vary in this analysis; School 280 excluded since only one FSM student attended it.

2. Other subgroup analyses for the primary outcome

As specified in the protocol, subgroup analyses were carried out for:

- students eligible for FSM;
- boys and girls; and
- high and low achievers on the pre-test (KS1; median-split based on all observed scores)

An additional subgroup analysis was planned to look at high and low implementation fidelity within treatment schools. However, given the large number and range of changes recorded by teachers as detailed in the reporting of the process evaluation in the main report (73% of intervention teachers in the follow-up survey reported making changes to the programme), and the difficulty in establishing the extent to which these changes were within the bounds of programme delivery as intended, no fidelity measure was constructed from the teacher survey.

The multilevel model described for the primary outcome was extended for each variable separately by adding the predictor itself and an interaction term between the intervention variable (GfW) and the variable currently analysed. The intervention was to be evaluated as showing a subgroup effect for the specific variable when the bootstrapped 95%-confidence interval for the coefficient for the interaction term does not include 0. As before, this analysis is purely exploratory and does not estimate the efficacy of the intervention itself.

As previously, an individual student i 's KS2past result in a specific school was modelled as depending on school j 's average KS2past attainment (random school-level intercept; μ_{0j}), previous attainment (KS1), and a random error term (ε_{ij}). For the test for subgroup effects, a coefficient for one of the student-level variables described above was added (Subgroup) as a random slope. Each school's average KS2past performance (μ_{0j}) was predicted by an overall intercept (average performance; γ_{00}); each school's level on the stratification variable which controls for geographical region (North East/ not-North East; REG); and the intervention to which the school was randomised (GfW) with the now added cross-level interaction with one of the sub-grouping variables (Subgroup) described above (formulas are a direct quote from statistical analysis plan):

$$KS2past_{ij} = \mu_{0j} + \mu_{1j}KS1_{ij} + \mu_{2j}Subgroup + \varepsilon_{ij} \quad (8)$$

$$\text{with } \varepsilon_{ij} \sim N(0, \sigma^2)$$

$$\mu_{0j} = \gamma_{00} + \gamma_{01}REG_{0j} + \gamma_{02}GfW_{0j} + u_{00} \quad (9)$$

$$\mu_{1j} = \gamma_{10} \quad (10)$$

$$\mu_{2j} = \gamma_{20} + \gamma_{21}GfW_{0j} + u_{20} \quad (11)$$

$$\text{with } u_{00} \sim N(0, \tau_1^2)$$

The analysis was done in the R environment (R Core Team, 2016); specifically the R-package lme4 (Bates, Mächler, Bolker, & Walker, 2015) was used with the corresponding formula expression in the command lmer():

KS2past ~ KS1 + Subgroup + REG + GfW + Subgroup:GfW + (1+Subgroup|School)

The intervention is evaluated as having shown a potential interaction with the specified subgroup variable when the 95%-bootstrap confidence interval of (γ_{21} ; formula 11) does not include 0. Only in this case was more detailed reporting on subgroup statistics undertaken in the main report (Table 10). The exception is FSM for which details were routinely reported.

Table D.1.5 presents the estimated coefficients for the fidelity variable and their respective confidence intervals. For FSM and gender for imputed as well as observed data analyses the confidence intervals

clearly include 0, i.e. no interaction effect between the subgroup and the treatment was found, the potential strength and/or direction of the effect of GfW on KS2post did not differ across the subgroups.

The case is different for the dichotomised pre-test measure. The median value of the KS1 pre-test measure (observed values) was used to split the sample into students whose scores were below the median ($N = 2246$) vs. equal and above the median ($N = 3758$). This variable was entered into the analysis detailed above instead of the KS1 continuous pre-test as well as its interaction with the treatment. The confidence intervals of the observed data analysis do not include "0", i.e. high and low achievers benefitted potentially differently from the treatment.

Table D.1.5. Summary of results obtained for the subgroup analyses for the primary outcome.

	Coefficient	N (SD)
FSM	.48 (-.18, 1.18)	5,180.98 (29.57)
FSM – imputed	.49 (-.16, 1.08)	6,306
Gender	-.08 (-.62, .47)	5,181.10 (29.75)
Gender – imputed	-.16 (-.74, .39)	6,306
Pre-test, dichotomised	.72 (.12, 1.40)	5,181.17 (29.84)
Pre-test, dichotomised – imputed	.65 (-.03, 1.27)	6,306

Note. All analyses bootstrapped with $b = 1000$ samples; due to unequal missingness patterns across bootstrap samples the N for the observed data analyses is an averages (SD) of the analysed cases in these bootstrap runs.

Tables D.1.6 and D.1.7 present the estimated coefficients from the bootstrapped analyses in the observed data (for effect sizes see Table 10 in main report). There is no relationship between the intervention and KS_{past} in the high performing pupils (indicated by very small coefficients and confidence intervals that overlap with 0); but in the population of the lower performing pupils a negative relationship is found: lower performing pupils in schools that received the intervention are doing slightly worse (about .10 KS_{2past} points) than comparable pupils at schools that did not receive the intervention. This result also holds when missing data are imputed (Tables D.1.8 and D.1.9).

Table D.1.6. Bootstrapped coefficients for the estimated model to estimate the treatment effect in the group of students performing in the top 50% of the pre-test distribution (KS1 Writing Result); observed data

Variable	Average Coefficient (SD)	95% Confidence Interval
Intercept	18.99 (.44)	18.14, 19.84

Region	-.04 (.19)	-.50, .45
Treatment	.09 (.18)	-.36, .51

Note. The analysis is based on $b = 1000$ bootstrap samples. $N=3357.66$ (18.31); Average N in control = 1550.14 SD=18.31; Average N in intervention group 1807.52, SD=12.25

Table D.1.7. Bootstrapped coefficients for the estimated model to estimate the treatment effect in the group of students performing in the lower 50% of the pre-test distribution (KS1 Writing Result); observed data

Variable	Average Coefficient (SD)	95% Confidence Interval
Intercept	11.87 (.49)	10.96, 12.81
Region	.22 (.26)	-.28, .72
Treatment	-.66 (.25)	-1.14, -.16

Note. The analysis is based on $b = 1000$ bootstrap samples. $N=1824.83$ (17.37); Average N in control = 856.03 SD=11.93; Average N in intervention group 968.80 SD=11.93

Table D.1.8. Bootstrapped coefficients for the estimated model to estimate the treatment effect in the group of students performing in the top 50% of the pre-test distribution (KS1 Writing Result); imputed data

Variable	Average Coefficient (SD)	95% Confidence Interval
Intercept	19.02 (.40)	18.27, 19.79
Region	-.08 (.22)	-.50, .38
Treatment	.04 (.21)	-.36, .48

Note. The analysis is based on $b = 1000$ bootstrap samples. $N=3886.27$ (7.35); Average N in control = 1814.26 SD=4.35; Average N in intervention group 2072.01, SD=4.96; since the pre-test needs to be imputed as well, sample sizes vary across bootstrap samples also in the imputed data condition since some students are sometimes imputed as above, sometimes as below the median performance.

Table D.1.9. Bootstrapped coefficients for the estimated model to estimate the treatment effect in the group of students performing in the lower 50% of the pre-test distribution (KS1 Writing Result); imputed data

Variable	Average Coefficient (SD)	95% Confidence Interval
Intercept	11.95 (.48)	10.99, 12.86
Region	-.08 (.25)	-.57, .42
Treatment	-.65 (.25)	-1.11, -.12

Note. The analysis is based on $b = 1000$ bootstrap samples. $N=2419.73$ (7.35); Average N in control = 1273.99 $SD=4.35$; Average N in intervention group 1273.99 $SD=4.96$; since the pre-test needs to be imputed as well, sample sizes vary across bootstrap samples also in the imputed data condition since some students are sometimes imputed as above, sometimes as below the median performance.

D.2 Secondary Outcomes

Below summarises the results for the analyses of the secondary outcomes. How to interpret the coefficients is presented in detail in Section D.1. In the following only details regarding the tests of the outcome and its effect size are presented. The reader should note that although these analyses were pre-planned, their family-wise error rate was not controlled for multiple testing, i.e. the rate of false-positive findings could be higher than expected. The results are presented for purely exploratory purposes to investigate potential spill-over and adverse effects that might merit further investigation. The N is provided in the note to each table. For the non-imputed analyses the N varies for each analysis depending on how many cases with missing data were selected in the respective bootstrap runs. Therefore the averages and their standard deviation are provided.

Tables D.2.1 and D.2.2 present the results regarding the writing task (Table D.2.1: complete case analysis; Table D.2.2: imputed data). The estimated coefficient indicates that pupils receiving Grammar for Writing did on average .01 points better than pupils not receiving the programme (imputed data: .003 points on average better in intervention group). The confidence interval for the coefficient includes "0" for observed and imputed data, which indicates that the Null hypothesis of "no effect" could not be rejected. The associated effect size was $ES = 0.02$ with its 95%-bootstrap confidence interval ranging from $LB_{ES} = -0.02$ to $UB_{ES} = 0.07$ (imputed: .004; $LB_{ES} = -0.04$ to $UB_{ES} = 0.05$). The estimated effect sizes are very small and the confidence interval includes 0, indicating that the Null hypothesis of "no effect" could not be rejected.

In both analyses (observed and imputed) a small effect for region is found: pupils in schools in other parts than the North East were doing on average .04 points worse (imputed: .06) than pupils in schools in the North East.

Table D.2.1. Bootstrapped coefficients for the secondary outcome (KS2 Writing assessment), observed data

Variable	Average Coefficient (SD)	95% Confidence Interval
Intercept	5.91 (.03)	5.85, 5.96
KS1 Writing Result	.15 (.003)	.15, .16
Region	-.04 (.02)	-.07, -.01

Treatment	.01 (.01)	-.01, .04
------------------	-----------	-----------

Note. The analysis is based on $b = 1000$ bootstrap samples. The estimated intraclass correlations were for the outcome $ICC = .103$ ($SD = .009$); average $N = 6788.11$ ($SD = 19.28$), N intervention = 3552.52 (13.98); N control = 3235.59 ($SD = 12.91$); the average level-1 variance/ residual was $resid = .37$ and the average level-2 variance was $varl2 = .06$

Table D.2.2. Bootstrapped coefficients for the secondary outcome (KS2 Writing assessment), imputed data

Variable	Average Coefficient (SD)	95% Confidence Interval
Intercept	5.92 (.03)	5.86, 5.98
KS1 Writing Result	.16 (.003)	.15, .16
Region	-.06 (.02)	-.09, -.03
Treatment	.003 (.02)	-.03, .03

Note. The analysis is based on $b = 1000$ bootstrap samples. The estimated intraclass correlations were for the outcome $ICC = .104$ ($SD = .009$); $N = 7200$, N intervention = 3776; N control = 3424; the average level-1 variance/ residual was $resid = .41$ and the average level-2 variance was $varl2 = .06$

Tables D.2.3 and D.2.4 present the results regarding the reading task (Table D.2.3: complete case analysis; Table D.2.4: imputed data). The estimated coefficient indicates that pupils receiving Grammar for Writing did on average .12 points better than pupils not receiving the programme (imputed data: .03 points on average better in the intervention group). The confidence interval for the coefficient includes "0" for observed and imputed data, which indicates that the Null hypothesis of "no effect" could not be rejected. The associated effect size was $ES = -0.01$ with its 95%-bootstrap confidence interval ranging from $LB_{ES} = -0.03$ to $UB_{ES} = 0.06$ (imputed: .004; $LB_{ES} = -0.04$ to $UB_{ES} = 0.05$). The estimated effect sizes are very small and the confidence interval includes 0, indicating that the Null hypothesis of "no effect" could not be rejected.

Table D.2.3. Bootstrapped coefficients for the secondary outcome (KS2 Reading assessment), observed data

Variable	Average Coefficient (SD)	95% Confidence Interval
Intercept	29.82 (.36)	29.12, 30.51
KS1 Writing Result	1.65 (.03)	1.59, 1.71
Region	-.02 (.20)	-.39, .37
Treatment	.12 (.19)	-.25, .48

Note. The analysis is based on $b = 1000$ bootstrap samples. The estimated intraclass correlations were for the outcome $ICC = .162$ ($SD = .008$); average $N = 6647.12$ ($SD = 22.13$), N intervention = 3478.01 ($SD = 15.99$);

N control = 3169.11 (SD = 15.04); the average level-1 variance/ residual was $resid = 50.38$ and the average level-2 variance was $varl2 = 13.11$

Table D.2.4. Bootstrapped coefficients for the secondary outcome (KS2 Reading assessment), imputed data

Variable	Average Coefficient (SD)	95% Confidence Interval
Intercept	29.61 (.36)	28.93, 30.26
KS1 Writing Result	1.73 (.03)	1.68, 1.78
Region	-.08 (.20)	-.45, .32
Treatment	.03 (.19)	-.34, .42

Note. The analysis is based on $b = 1000$ bootstrap samples. The estimated intraclass correlations were for the outcome $ICC = .159$ ($SD = .008$); $N = 7200$, N intervention = 3776; N control = 3424; the average level-1 variance/ residual was $resid = 52.32$ and the average level-2 variance was $varl2 = 13.46$

Tables D.2.5 and D.2.6 present the results regarding the grammar, punctuation and spelling task (Table D.2.5: complete case analysis; Table D.2.6: imputed data). The estimated coefficient indicates that pupils receiving Grammar for Writing did on average .62 points worse than pupils not receiving the programme (imputed data: .65 points on average worse in intervention group). The confidence interval for the coefficient excludes "0" for observed and imputed data, which indicates that the Null hypothesis of "no effect" could be rejected. The associated effect size was $ES = -0.06$ with its 95%-bootstrap confidence interval ranging from $LB_{ES} = -0.10$ to $UB_{ES} = -0.01$ (imputed: $-.06$; $LB_{ES} = -0.11$ to $UB_{ES} = -0.02$). The estimated effect sizes are small.

Table D.2.5. Bootstrapped coefficients for the secondary outcome (KS2 Grammar, Punctuation and Spelling assessment), observed data

Variable	Average Coefficient (SD)	95% Confidence Interval
Intercept	45.59 (.44)	44.75, 46.50
KS1 Writing Result	2.70 (.04)	2.63, 2.77
Region	.09 (.25)	-.40, .57
Treatment	-.62 (.24)	-1.10, -.13

Note. The analysis is based on $b = 1000$ bootstrap samples. The estimated intraclass correlations were for the outcome $ICC = .138$ ($SD = .008$); average $N = 6662.24$ ($SD = 22.29$), N intervention = 3489.36 ($SD = 15.52$); N control = 3173.09 ($SD = 15.11$); the average level-1 variance/ residual was $resid = 84.63$ and the average level-2 variance was $varl2 = 26.17$

Table D.2.6. Bootstrapped coefficients for the secondary outcome (KS2 Grammar, Punctuation and Spelling assessment, imputed data)

Variable	Average Coefficient (SD)	95% Confidence Interval
Intercept	45.33 (.43)	44.48, 46.15
KS1 Writing Result	2.82 (.03)	2.75, 2.88
Region	-.02 (.24)	-.47, .44
Treatment	-.65 (.23)	-1.13, -.18

Note. The analysis is based on $b = 1000$ bootstrap samples. The estimated intraclass correlations were for the outcome $ICC = .130$ ($SD = .008$); average $N = 7200$, N intervention = 3776; N control = 3424; the average level-1 variance/ residual was $resid = 88.85$ and the average level-2 variance was $varl2 = 25.64$

D.3 Report on Missing Data

Specified analyses according to the Statistical Analysis Plan (SAP)

The SAP stated that the amount of missing data will be documented for each variable individually as well as for the patterns of missing values which occur. Further, the relative frequency of pupils with any missing data will also be presented by school. To evaluate the impact of missing data on the robustness of findings from the ITT analyses of the primary outcome, sensitivity analyses will be run to evaluate the robustness of the results if either $> 5\%$ missing data for the primary outcome analysis are encountered (i.e. 5% of cases would have to be deleted listwise for that analysis); or if at least one school which enters the ITT analysis has more than $> 15\%$ missing responses for the primary outcome.

Primary outcome analysis set

Twenty schools did not provide any data on the primary outcome, i.e. their pupils were not assessed using the primary outcome measure at all. Due to this large amount of systematically missing data without any reference values for the outcome within each randomisation unit it was decided not to impute. We therefore present first the school-level differences between those schools which returned the primary outcome and those which did not.

Schools withdrawn from the study were similarly distributed across the two treatment groups with 9/78 (11.5%) dropping out in the control group and 11/77 (14.3%) in the intervention group ($X^2(df=1) = .26$, $p = .61$). Because the withdrawn schools were distributed equally across the intervention groups and also because within 20 schools the numbers would be too small to describe by-group statistics, in the following data on the $N=135$ schools that provided primary outcome assessments and the $N=20$ which left is provided. Not all schools, for example some new academies, had school-level data available for this analysis. Table D.3.1 shows the descriptive statistics between the schools that stayed and those that withdrew. For several criteria the differences in the averages are below a Hedges $g = .20$, which is generally considered as a threshold for group differences that are likely negligible (also for school type and Ofsted rating, tables D.3.2 and D.3.3). Nevertheless, the schools remaining in the trial had on average lower percentages achieving the expected levels in Reading, Maths and lower proportions of EAL pupils than those who withdrew from the study, but a higher percentage of SEN pupils. Table D.3.2 further shows the distribution of school types across participating and withdrawn schools which differs only minimally across the two intervention groups ($X^2(df=5) = 2.00$) as do the Ofsted ratings ($X^2(df=3) = 2.22$; table D.3.3)

Table D.3.1. Averages and standard deviations for school-level variables for schools contributing to the primary outcome analysis and those withdrawn from the study

	Schools in Primary Outcome Analysis		Withdrawn schools N=19*		Effect size Hedges <i>g</i>
	M	SD	M	SD	
School Size	N=133 377.34	233.01	334.32	202.33	.19
%FSM	N=133 26.20	13.40	23.98	14.23	.16
Grammar	N=134 .67	.18	0.71	0.18	.22
Writing	N=134 .77	.18	.80	.15	.17
Reading	N=134 .67	.17	.74	.16	.41
Maths	N=134 .72	.17	.79	.10	.43
EAL	N=133 21.24	25.61	29.75	32.26	.32
SEN	N=133 1.43	1.61	.88	.76	.36

* Data was not available for one school.

Table D.3.2. Percentages of school types within the groups of participating and withdrawn schools

School type	Primary Outcome	Withdrawn	Count
Academy - Converter Mainstream	17.8	10	26
Academy Sponsor Led Mainstream	7.4	5	11
Community School	54.1	60	85
Foundation School	9.6	10	15
Voluntary Aided School	8.9	15	15
Voluntary Controlled School	2.2	0	

Table D.3.3 Percentages of Ofsted ratings within the groups of participating and withdrawn schools

Ofsted Rating	Primary Outcome	Withdrawn	Count
Outstanding	13.3	15	21
Good	68.9	80	109
Requires improvement	13.3	5	19
No Ofsted assessment	4.4	0	6

In the primary outcome analysis all pupils that provided at least demographic information at baseline were taken into account. The only pieces of demographic information read from NPD in this study are gender ($n = 39$ missing entries) and EVERFSM ($n = 43$ missing entries); of these $n = 39$ cases had missing values on both variables and were therefore excluded from the analyses. For this analysis the data of $N = 135$ schools were available, which on average contributed $N = 46.71$ ($SD = 21.63$) pupils and overall $N = 6306$ pupils were documented in the NPD data file.

Tables D.3.1 and D.3.2 present descriptive information based on the available cases for continuous and categorical data, respectively. The analysis of missing values below shows that the threshold was crossed for the primary outcome with 14.1% missing values overall. In addition to the analysis of the primary outcome on available cases assuming missing at random given predictors (i.e. pre-test, region and school average) a sensitivity analysis based on multiply imputed data needed to be conducted.

The bottom rows of Table D.3.4 and Figure D.3.5 provide further information on the patterns of missingness observed in the dataset. In Figure D.3.1, the variables are ordered from the one with the highest percentage of missing values to the lowest. The plot presents in principle horizontal lines for each student in the sample, which are red when data are observed and light red if data are missing. Further, most of the continuous variables also show skewed distributions. These are addressed by the bootstrap procedure implemented in the analytic strategy.

Table D.3.4. Missing data, descriptive information and qualitative assessment of distribution form for continuous data in the analysis sample for the primary outcome.

Variable	Missing	Mean (SD)	Median	Comment
KS2past writing paper (CalcTotal_Overall2)	890	16.24 (7.15)	16	slightly right-skewed
KS2 Writing assessment outcome (KS2_WRITTAOUTCOME_Code)	53	5.81 (.92)	6	strongly left-skewed
KS2 Grammar, Punctuation and Spelling assessment outcome (KS2_GPSMRK)	203	45.42 (14.39)	48	left-skewed
KS2 Reading assessment outcome (KS2_READMRK)	213	29.93 (10.13)	31	left-skewed

KS1 Writing Result (KS1_WRITPOINTS)	302	14.52 (3.79)	15	symmetric
--	-----	--------------	----	-----------

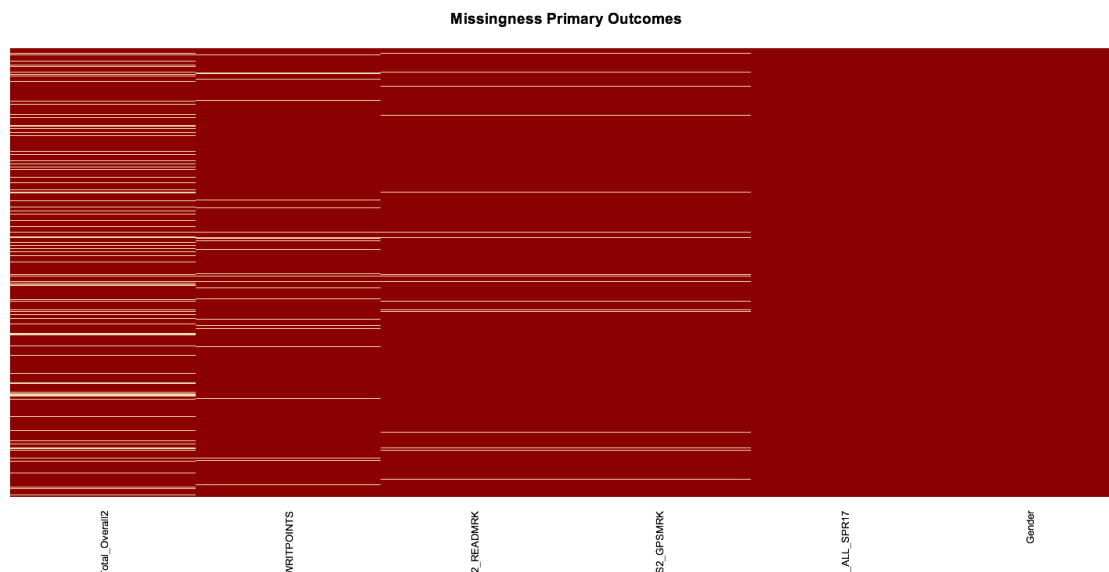
Note. N = 6306; 5% of $n = 316$

Table D.3.5. Frequency distributions for categorical data incl. missing values

	Observed	Missing
Treatment	Control=2,960 Intervention=3,346	0
Gender	1=3,188 2=3,118	0
EVERFSM_ALL_SPR17	0=3,378 1=2,924	4
Region	1=2,218 2=4,088	0
Any baseline data missing	0=6,004 1=302	--
Any secondary outcome missing	0=6,073 1=233	--
Any follow-up data missing	0=5,342 1=964	--

Note. N = 6306; 5% of $n = 316$

Figure D.3.1: Plot illustrating the patterns of missingness in the analysis sample for the primary outcome analysis; each red line represents a single case and where depicted in light red a missing value on the specific variable is observed.



Although the number of available predictors is very small, we checked whether variables predicted that any follow-up measure was missing in this sample. Gender was not predictive ($b = .02$, $p = .78$; regression weights on logit scale), but FSM-status increased the probability of not reporting any outcome data in the primary analysis sample ($b = .33$, $p < .001$) and a higher score in KS1 writing increased the probability of reporting results ($b = -.15$, $p < .001$).

Secondary outcome analysis set

In this analysis all pupils that provided at least demographic information should be taken into account. The only pieces of demographic information read from NPD in this study are gender ($n = 39$ missing entries) and EVERFSM ($n = 43$ missing entries); of these $n = 39$ cases had missing values on both variables and were therefore excluded from the analyses.

For this analysis the data of $N = 155$ schools were available, which on average contributed $N = 46.45$ ($SD = 22.50$) pupils and overall $N = 7200$ pupils were documented in the NPD data file. As the following table shows, the amount of missing data on individual variables relevant for the analysis of the secondary outcomes was below the pre-defined threshold apart from the baseline measure (KS1_WRITPOINTS). In addition to the analysis of the secondary outcomes on available cases assuming missing at random given predictors (i.e. pre-test, region and school average) a sensitivity analysis based on multiply imputed data needs to be conducted. See Table D.3.6 and Table DS.3.7 for further details.

Further, most of the continuous variables also show skewed distributions. These are addressed by the bootstrap procedure implemented in the analytic strategy.

Table D.3.6. Missing data, descriptive information and qualitative assessment of distribution form for continuous data in the analysis sample for the secondary outcome.

Variable	Missing	Mean (SD)	Median	Comment
KS2past writing paper (CalcTotal_Overall2)	1,784	16.24 (7.15)	16	slightly right-skewed
KS2 Writing assessment outcome (KS2_WRITTAOUTCOME_Code)	60	5.81 (.91)	6	strongly left-skewed
KS2 Grammar, Punctuation and Spelling assessment outcome (KS2_GPSMRK)	223	45.65 (14.33)	48	left-skewed
KS2 Reading assessment outcome (KS2_READMRK)	242	29.95 (10.06)	31	left-skewed
KS1 Writing Result (KS1_WRITPOINTS)	374	14.54 (3.79)	15	symmetric

Note. N = 7200; 5% of n = 360

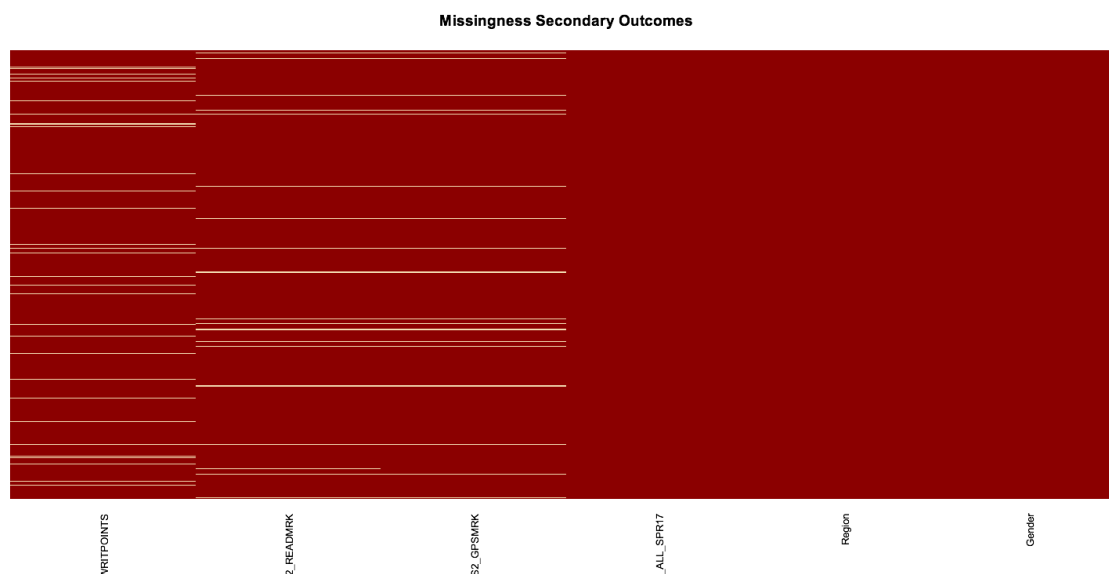
Table D.3.7. Frequency distributions for categorical data incl. missing values

	Observed	Missing
Treatment	Control=3,424 Intervention=3,776	0
Gender	1=3,631 2=3,569	0
EVERFSM_ALL_SPR17	0=3,921 1=3,275	4
Region	1=2,486 2=4,714	
Any baseline data missing	0=6,826 1=374	--
Any secondary outcome missing	0=6,938 1=262	--

Note. N = 7200; 5% of n = 360

Figure D.3.2 presents the missingness map for the variables that at some stage of the above process of sample selection showed missing values. The variables are ordered from the one with the highest percentage of missing values to the lowest one. The plot presents in principle horizontal lines for each student in the sample, which are red when data are observed and light-yellow if data are missing.

Figure D.3.2: Plot illustrating the patterns of missingness in the analysis sample for the secondary outcome analysis; each red line represents a single case and where depicted in light red a missing value on the specific variable is observed.



Although the number of available predictors is very small, we checked whether not reporting any of the three secondary outcomes depended on gender, KS1 writing result or FSM status. Being female ($b = -.42$, $p = .02$) and a higher KS1 writing score ($-.39$, $p < .001$) increased the probability that pupils returned any of the three outcomes.

Percentage of missing data per school Table D.3.8 presents the share of missing data patterns by school. The schools with a share of 1 (i.e. 100%) in the primary outcome (fifth column of Table D.3.8) are highlighted in grey since they are not part of the secondary analysis set. On multiple of the criteria it is also visible that our pre-set threshold of >15% missing values within a school has been reached.

Table D.3.8. Share of missing data per school.

SchoolID 2	Treatment (5=Control; 6=Intervention)	Any baseline data missing?	Any secondary outcomes missing?	Primary outcome missing?
151	6	0.04	0.04	0.15
152	6	0.07	0.07	0.20
153	6	0.10	0.10	1.00

154	5	0.07	0.00	0.07
155	6	0.06	0.02	0.10
157	6	0.04	0.00	0.13
158	5	0.14	0.00	0.16
159	6	0.07	0.00	0.05
160	6	0.24	0.10	0.20
161	5	0.12	0.06	0.12
163	6	0.02	0.00	0.04
164	5	0.00	0.04	0.04
165	6	0.07	0.01	0.04
166	6	0.02	0.11	0.14
167	6	0.15	0.00	0.24
168	5	0.12	0.02	1.00
169	5	0.06	0.06	0.17
170	6	0.04	0.00	0.18
172	6	0.04	0.02	0.13
173	5	0.09	0.04	0.16
174	5	0.00	0.12	0.19
175	5	0.00	0.00	0.10
177	5	0.07	0.00	0.04
178	6	0.03	0.02	0.14
179	6	0.04	0.00	0.04
180	5	0.00	0.00	0.10
181	5	0.00	0.04	0.11

183	6	0.01	0.03	0.11
185	6	0.04	0.08	0.08
186	6	0.00	0.00	0.20
187	6	0.00	0.00	0.06
189	5	0.07	0.00	0.07
192	6	0.09	0.05	0.14
193	5	0.04	0.04	0.13
194	5	0.00	0.00	0.11
195	5	0.00	0.00	0.05
196	5	0.04	0.00	0.24
197	6	0.03	0.06	1.00
198	5	0.02	0.02	1.00
199	6	0.00	0.03	0.11
200	5	0.00	0.00	0.31
202	5	0.00	0.04	0.36
203	5	0.07	0.00	0.30
204	6	0.02	0.02	0.08
205	5	0.05	0.05	0.09
206	6	0.00	0.00	0.07
207	6	0.00	0.14	0.18
208	5	0.00	0.00	0.21
209	6	0.18	0.15	0.25
210	6	0.00	0.00	0.17
211	5	0.00	0.00	0.00

212	5	0.17	0.17	1.00
213	5	0.00	0.02	0.13
214	5	0.10	0.00	0.03
215	6	0.00	0.00	0.20
217	6	0.02	0.00	1.00
218	5	0.00	0.00	0.04
219	6	0.00	0.00	0.21
220	6	0.02	0.03	0.15
224	5	0.08	0.04	0.20
227	5	0.00	0.02	0.08
228	6	0.00	0.04	0.07
229	6	0.10	0.05	0.15
233	6	0.06	0.02	0.09
234	6	0.16	0.19	0.45
235	6	0.03	0.00	0.09
236	5	0.05	0.00	1.00
240	6	0.06	0.04	0.15
243	5	0.00	0.00	0.17
244	6	0.03	0.03	1.00
245	6	0.10	0.00	1.00
247	5	0.00	0.17	0.13
249	6	0.00	0.08	0.19
250	5	0.00	0.03	0.07
251	5	0.08	0.07	0.07

252	5	0.03	0.03	0.20
253	6	0.03	0.00	1.00
255	6	0.00	0.00	1.00
256	5	0.00	0.03	0.03
259	5	0.18	0.05	1.00
260	6	0.29	0.11	0.20
261	6	0.03	0.03	0.20
262	6	0.19	0.09	0.21
263	6	0.13	0.07	0.05
265	5	0.03	0.00	0.03
266	5	0.10	0.01	0.12
267	6	0.02	0.02	0.10
268	6	0.05	0.03	0.03
269	6	0.16	0.08	1.00
270	5	0.06	0.11	0.20
271	5	0.00	0.08	0.18
272	5	0.02	0.05	0.49
273	5	0.02	0.16	0.25
274	5	0.09	0.05	0.16
275	6	0.09	0.02	0.09
277	5	0.00	0.03	0.61
278	6	0.00	0.02	0.02
280	5	0.00	0.03	0.10
281	6	0.00	0.11	0.32

282	5	0.00	0.00	1.00
283	5	0.04	0.07	1.00
284	6	0.00	0.02	0.28
285	6	0.05	0.02	0.04
286	5	0.00	0.00	0.06
287	6	0.04	0.04	0.17
290	5	0.06	0.02	0.10
291	6	0.03	0.00	0.03
292	5	0.00	0.00	0.19
293	6	0.00	0.00	0.04
294	6	0.02	0.00	0.05
295	6	0.00	0.00	0.08
298	6	0.02	0.02	0.14
301	5	0.23	0.00	0.09
302	5	0.05	0.03	0.09
303	6	0.17	0.00	1.00
304	6	0.05	0.12	0.20
305	5	0.00	0.00	0.00
306	5	0.02	0.07	0.10
307	6	0.00	0.02	0.14
308	6	0.12	0.01	1.00
310	5	0.13	0.03	0.53
311	5	0.05	0.00	0.32
312	6	0.00	0.07	0.07

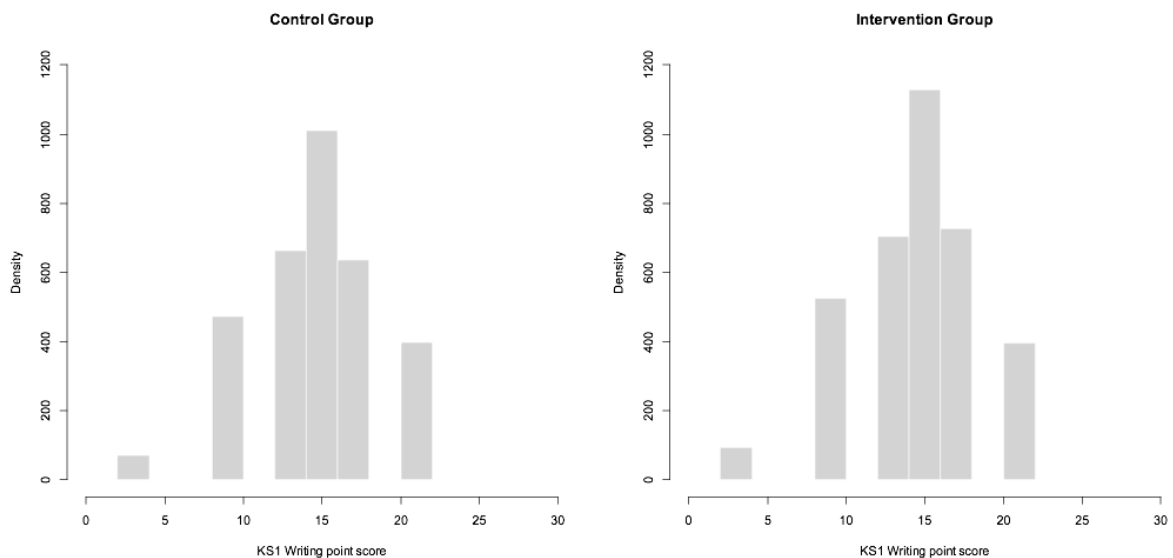
313	5	0.02	0.01	0.07
315	5	0.00	0.00	1.00
316	5	0.06	0.10	0.35
317	5	0.07	0.11	0.13
318	5	0.04	0.00	0.17
319	5	0.00	0.00	0.30
320	5	0.00	0.02	0.08
321	6	0.01	0.05	0.13
322	6	0.02	0.07	0.05
323	5	0.05	0.02	0.12
324	6	0.00	0.00	1.00
325	5	0.03	0.05	0.10
326	5	0.00	0.07	0.10
327	5	0.14	0.02	0.34
328	5	0.00	0.00	0.08
329	6	0.00	0.05	0.10
330	5	0.15	0.00	0.09
331	5	0.11	0.04	0.15
333	5	0.00	0.14	1.00
334	6	0.00	0.05	0.23
335	6	0.00	0.07	0.19
337	6	0.00	0.00	0.05
339	6	0.00	0.03	0.10
341	6	0.03	0.00	0.03

342	6	0.00	0.05	0.05
343	6	0.02	0.00	0.06
344	5	0.07	0.03	0.03
345	5	0.04	0.07	0.36
346	6	0.04	0.04	0.35
347	5	0.00	0.03	0.10
348	5	0.05	0.00	0.08
349	5	0.00	0.06	0.13

D.4 Distribution of Pre-Test Results

Table 7 in the main document reports averages and SDs for the pre-test (KS_WRITPOINTS). The estimated effect size for a between group difference is small (Hedges $g = -.03$) and according to our pre-defined criteria likely to be negligible. Figure D.4.1 further displays the group-wise histograms with the distribution of the pre-test. The distribution is symmetrical and nearly identical across both groups.

Figure D.4.1. Distribution of pre-test results by group (control in left panel; intervention in right panel).



D.5 Non-compliance analysis

The following presents more detail on the additional analysis in which the intervention allocation was replaced by the percentage of CPD training days attended by teachers as an approach to assess whether compliance with the training had an effect on the outcome. The GfW study was planned for a single primary outcome, the writing assessment developed by the team from previously used Key Stage 2 assessments ($KS2_{past}$) to answer the question 'how effective Grammar for Writing is in improving writing skills in Year 6 pupils?' In accordance with the power analysis, pre-test data from the Key Stage 1 (KS1) writing results were used as a student-level covariate ($KS1$) without random variation across schools. An individual student i 's $KS2_{past}$ result in a specific school was modelled as depending on school j 's average $KS2_{past}$ attainment (random school-level intercept; μ_{0j}) and a random error term (ε_{ij}). Each school's average $KS2_{past}$ performance (μ_{0j}) was predicted by an overall intercept (average performance; γ_{00}); each school's level on the stratification variable which controls for geographical region (North East/ not-North East; REG); and the intervention to which the school was randomised (GfW):

$$KS2past_{ij} = \mu_{0j} + \mu_{1j}KS1_{ij} + \varepsilon_{ij} \quad (1)$$

$$\text{with } \varepsilon_{ij} \sim N(0, \sigma^2)$$

$$\mu_{0j} = \gamma_{00} + \gamma_{01}REG_{0j} + \gamma_{02}GfW_{0j} + u_{00} \quad (2)$$

$$\mu_{1j} = \gamma_{10} \quad (3)$$

$$\text{with } u_{00} \sim N(0, \tau_1^2)$$

The analysis was performed in the R environment (R Core Team, 2016); specifically the R-package lme4 (Bates, Mächler, Bolker, & Walker, 2015). The corresponding formula expression in the command lmer():

$$KS2past \sim KS1 + REG + GfW + (1|school)$$

The analysis was cluster-bootstrapped as applied in previous projects (Hanley, Böhnke, Slavin, Elliott, & Croudace, 2016; Huang, 2016: From each school a random sample of the same size as its actual sample was drawn (with replacement) and across these school-wise bootstrap samples, the mixed model was then estimated.¹⁸ This process is repeated $b = 1000$ times and for a 95%-confidence interval the statistical estimates (here the γ_{03} values) were saved and their top and bottom 2.5%-quantiles were identified.

For this specific analysis the variable coding the intervention (GfW) was replaced by the number of CPD training days that teachers attended (as a percentage of possible days; see main report for more detail). The data were aggregated on school level, i.e. if several teachers attended for a given school, their number of days were averaged. These days were converted into a percentage, representing the percentage of CPD training days attended. Two different versions of this measure were available; one

¹⁸ E.g. if there were observations 1,2,3,4,5 in a school, one resample could be [1,2,2,5,4] and another [1,5,1,1,3].

for the first 3 days of CPD training and one for all 4 days of CPD training. Across the 77 intervention schools the average compliance score across the three days was 79.7% (SD = 30.71) and 69.9% (SD = 29.89) for the four days measure, respectively. The compliance score is by definition set to 0% for all schools in the control group.

Table D.5.1 presents the estimated coefficients for the percentage out of three days attended, the associated effect size and their respective confidence intervals. For the primary outcome the observed and imputed data analyses corroborate the finding from the main analysis: No treatment effect is also found with this proxy. The estimated effect sizes are small and all confidence intervals include "0".

The picture is different for the secondary outcomes. The main analysis showed already a potentially negative effect of the intervention on the KS2 GPS assessment outcome (see Table 9). This trend is found for all secondary outcomes when the degree of participation in CPD training is used as a proxy for training compliance: In all three measures small effects are found in the observed and imputed data analyses whose confidence intervals all consistently do not include "0". In all three measures the pupils in schools whose teachers received the intervention do on average slightly worse than those pupils in schools whose teachers did not take part or went to a smaller percentage of CPD training days. Since the variable for compliance was rescaled for this analysis to a range of 0-1 the results in Table D.5.1 can be read as:

- The average difference between a school with teachers who attended all CPD training days (1 / 100%) and a school where teachers did not attend/ did not take part/ were a control school (0 / 0%) was .05 score points in the writing task.
- The difference as an effect size between a school with teachers who attended all CPD training days (1 / 100%) and a school where teachers did not attend/ did not take part/ were a control school (0 / 0%) was .08 SDs.

Table D.5.1. Summary of results obtained for the non-compliance analysis

	Coefficient	Effect size	N
Primary Outcome	-.09 (-.41, .22)	-.02 (-.07, .04)	5,182.13 (29.19)
Primary Outcome imputed	-.13 (-.45, .20)	-.02 (-.08, .03)	6,306
Primary Outcome FSM	.36 (-.02, .79)	.06 (-.004, .13)	4,014.78 (26.38)
Primary Outcome FSM imputed	.16 (-.24, .57)	.03 (-.04, .10)	4,984.14 (1.78)
Secondary WRITTAMRK	-.04 (-.07, -.01)	-.06 (-.11, -.01)	6,787.19 (18.50)
Secondary WRITTAMRK, imputed	-.05 (-.09, -.02)	-.08 (-.12, -.02)	7,200°
Secondary READ	-.81 (-1.21, -.41)	-.10 (-.15, -.05)	6,646.99 (22.05)

Secondary READ, imputed	-0.91 (-1.33, -.50)	-.11 (-.16, -.06)	7,200
Secondary GPS	-1.49 (-2.00, -.96)	-.14 (-.19, -.09)	6,660.89 (21.51)
Secondary GPS, imputed	-1.55 (-2.08, -1.05)	-.14 (-.19, -.10)	7,200

Note. Compliance is set to 0% for all schools in the control group. For the analysis of the secondary outcome the compliance is also set to 0% for all schools withdrawn from the study. The compliance was converted on a 0 to 1 scale for the analysis so that coefficient sizes are comparable to those of the dichotomous treatment variable in the main analysis.

Table D.5.2 presents the estimated coefficients for the compliance variable (based on four days of CPD training), the associated effect size and their respective confidence intervals. For the primary outcome the observed and imputed data analyses corroborate the finding from the main analysis: No treatment effect is also found with this proxy. The estimated effect sizes are small and all confidence intervals include "0". Nevertheless, for the observed data looking at FSM pupils only, a small positive potential effect is found, which is nevertheless non-significant in the sensitivity analysis with the imputed data. It repeats the findings from the main analysis which indicates that FSM-pupils potentially did better (albeit with a small effect size).

Looking at the secondary outcomes the known pattern is repeated. All secondary outcomes show potentially small effects whose confidence intervals all consistently do not include "0". In all three measures the pupils in schools whose teachers went to more training days do on average slightly worse than those pupils in schools whose teachers did not take part or went to a smaller percentage of CPD training days.

These analyses were planned, but not controlled for the family-error rate, i.e. statistically they can only be seen as a potential indication for a negative effect of the treatment, they could be false positive findings. Nevertheless, the consistency of this finding across analytic strategies and outcome measures needs further scrutiny. It is important to remember, though, that the proxy is not assessing the intervention itself, but participation of the teachers in CPD training days.

Table D.5.2. Summary of results obtained for the non-compliance analysis (criterion: attending four days of CPD training)

	Coefficient	Effect size	N
Primary Outcome	.03 (-.33, .42)	.01 (-.05, .07)	5,184.10 (29.14)
Primary Outcome imputed	-.02 (-.40, .37)	-.003 (-.07, .06)	6,306
Primary Outcome FSM	.52 (.05, 1.00)	.09 (.01, .17)	4,015.49 (26.32)
Primary Outcome FSM imputed	.31 (-.14, .75)	.05 (-.02, .13)	4,984.14 (1.78)

Secondary WRITTAMRK	-0.04 (-.08, -.01)	-0.07 (-.12, -.01)	6,786.64 (19.38)
Secondary WRITTAMRK, imputed	-0.06 (-.10, -.02)	-0.08 (-.14, -.03)	7,200
Secondary READ	-0.75 (-1.18, -.33)	-0.09 (-.15, -.04)	6,645.62 (21.78)
Secondary READ, imputed	-0.88 (-1.31, -.40)	-0.11 (-.16, -.05)	7,200
Secondary GPS	-1.56 (-2.13, -.99)	-0.15 (-.20, -.09)	6,661.80 (21.17)
Secondary GPS, imputed	-1.61 (-2.19, -1.03)	-0.15 (-.20, -.10)	7,200

Note. Compliance is set to 0% for all schools in the control group. For the analysis of the secondary outcome the compliance is also set to 0% for all schools withdrawn from the study. The compliance was converted on a 0 to 1 scale for the analysis so that coefficient sizes are comparable to those of the dichotomous treatment variable in the main analysis.

D.6 Analysis of the Grammar Quiz scores

The results of the assessment of the grammar quiz are presented in the main report. In the following additional detail regarding the analyses and results obtained for the Grammar Quiz are presented.

Descriptive Analysis of the raw teacher-level data

Overall $N = 312$ teachers were eligible for participation in the Grammar Quiz. Of these $n = 297$ responded at T1; $n = 222$ had responses at T2 and $n = 222$ had responses to both assessments ($n = 15$ to none and $n = 75$ only to T1). The distribution across the intervention groups as well as means and SDs can be found in Table 7 in the main report.¹⁹ The individual items' descriptive statistics can be found in Table E.6.4 in this appendix.

Reliability estimates:

As an estimate of reliability Cronbach- α (more precisely: Kuder-Richardson-21 for dichotomous items) was determined. The estimated reliability was $Rel = .55$ (95% confidence interval: .48, .62; $N = 297$) and the item analysis showed that some items are negatively correlated with the overall score. Negatively correlated items were also found for the post-test and the Cronbach alpha was even lower, $Rel = .43$ (.34, .52; $N = 222$).

Pre-post correlation in the control group on observed scores at both assessments was also used as an indicator of the Grammar Quiz' reliability, but a $r = .35$ (.17, .51; $N = 222$) pointed to a correlation different from zero, although nevertheless not a very high one.

Tests of potential effect on teachers' knowledge:

Two tests looked more closely at whether the intervention had a potential effect on teachers' grammar knowledge. The first and planned analysis consisted of a bootstrapped t -Test testing whether teachers'

¹⁹ The median for both group scores at both pre and post the intervention is $Med = 21$ score points.

scores in the Grammar Quiz differed after the intervention. The average estimate across $b = 1000$ bootstrap runs (with each $N=222$) pointed to teachers in the intervention group scoring $-.27$ ($-.93, .32$) points lower than the teachers in the control condition. Since this bootstrapped confidence interval included 0 (as well as for the t -statistic: $-1.00, 2.74$; average: $.80$), no difference in scores in the Grammar Quiz between the two groups was found after the intervention.

One additional test evaluated whether differential gains may have happened in the intervention group. To this end, the post scores in the Grammar Quiz were regressed on pre-scores, including an interaction effect between pre-scores and the intervention. The results are presented in Table E.6.1. Teachers who performed better in the Grammar Quiz before the intervention also performed slightly better after the intervention ($+.30$ points per point in the pre-test). But neither the intervention nor the interaction between the intervention and the pre-test showed a statistically significant effect.

Table D.6.1. Linear regression model testing for a differential effect of the intervention on teachers' grammar knowledge (N = 222)

Variable	Average Coefficient (SE)	95% Confidence Interval
Intercept	14.82 (1.78)	11.31, 18.32
Pre-score	.30 (.08)	.13, .46
Intervention	-.58 (2.36)	-5.23, 4.08
Pre-score X intervention	.01 (.11)	-.20, .23

Note. $R^2_{adj} = .11$; $F_{df1=3, df2=218} = 10.45$

Factor analyses

The goal of this analyses was to determine how many factors are needed to represent the responses to the Grammar Quiz. Traditional rules of thumb like the Kaiser criterion (eigenvalues > 1) have been shown to be unreliable in their assessment of dimensionality and instead the size of the relevant eigenvalue is determined through newer methods.

Since the goal of the Grammar Quiz was to provide a single score proxy for grammar knowledge and no claim about the existence of a single trait is made, principal component analyses are used in the following (instead of factor analytic approaches; e.g., Costa, 2015).

Due to the small sample size the use of resampling techniques and categorical data methods was not possible (Edwards & Wirth, 2007), so the correlation matrix between the quiz responses was determined via full information maximum likelihood (i) across all 59 questions (pre and post), (ii) only for the 30 pre-intervention questions and (iii) only for the 29 post-intervention questions. $B = 500$ data sets were then simulated assuming no correlations between the items. The eigenvalues from these simulated principal component analyses were saved and the 95%-quantile of these used to establish a cut off for eigenvalues typically to be expected with this sample size and number of items. The empirical eigenvalues from the three observed matrices were then compared against their respective cut offs from the simulated data sets. Principal components were seen as relevant if their empirical eigenvalue was larger than the cut off.

The results were as follows:

- The parallel analysis resulted in the suggestion of 13 components across both assessments; eight components for the pre-test; and 4 components for the post-test.

- Extracting one component resulted in (i) 7% explained variance for across both pre and post questions; (ii) 9% for the pre-intervention Grammar Quiz; and (iii) 10% for the post-intervention Grammar Quiz.
- Extracting four components from the post-intervention quiz resulted in 29% explained variance for all components. Table E.6.2 presents the extracted component loadings to illustrate that at least two of the components (components 1 and 2) have several and often substantial loadings.
- Overall the Grammar Quiz was able to assess individual differences, but the reliability of the quiz' scores was low, it did not seem to respond to the treatment on teacher level, and the principal component analysis showed that it likely assessed multiple components instead of a single, strong score.

Table D.6.2. Varimax-rotated component loadings from principal component analysis of the post-intervention Gramma Quiz (N = 222); loadings < .20 blanked

	Component 1	Component 2	Component 3	Component 4	Explained Variance
P2_Q5_2		-0.28		0.26	0.168
P2_Q5_3	0.23	0.34			0.191
P2_Q5_4	0.61				0.375
P2_Q5_5				0.79	0.662
P2_Q5_6	0.42				0.19
P2_Q5_7	0.49		0.21		0.309
P2_Q5_8	0.51				0.276
P2_Q5_9	0.58				0.366
P2_Q5_10	0.53				0.315
P2_Q5_11	0.67				0.464
P2_Q5_13_1		-0.25			0.126
P2_Q5_13_2			-0.5		0.27
P2_Q5_13_3			0.47		0.23
P2_Q5_13_4		-0.23	0.31	0.41	0.315
P2_Q5_13_5			-0.24		0.093
P2_Q5_14_1				0.68	0.546
P2_Q5_14_2		0.37			0.156

P2_Q5_14_3				0.39	0.174
P2_Q5_14_4					0.091
P2_Q5_14_5	0.53				0.289
P2_Q5_16_1a					0.077
P2_Q5_16_2		-0.51			0.298
P2_Q5_16_3		-0.45			0.221
P2_Q5_16_4	0.38				0.15
P2_Q5_16_5		0.55		0.29	0.402
P2_Q5_16_6		0.66	0.24		0.502
P2_Q5_17			0.6		0.421
P2_Q5_18	0.24		0.59		0.407
P2_Q5_19		-0.24	-0.32		0.185

Mediation effect of grammar knowledge

To gauge the potential for a mediation effect of higher grammar knowledge on the side of the teachers the model used in the analysis of the primary outcome was extended by incorporating the teacher's grammar quiz performance (GQ) as a predictor on student level (for all other variables compare formulae 1-3 above).

$$KS2past_{ij} = \mu_{0j} + \mu_{1j}KS1_{ij} + \mu_{2j}GQ_{ij} + \varepsilon_{ij} \quad (4)$$

$$\text{with } \varepsilon_{ij} \sim N(0, \sigma^2)$$

$$\mu_{0j} = \gamma_{00} + \gamma_{01}REG_{0j} + \gamma_{02}GfW_{0j} + u_{00} \quad (5)$$

$$\mu_{1j} = \gamma_{10} \quad (6)$$

$$\mu_{2j} = \gamma_{20} + u_{20} \quad (7)$$

$$\text{with } u_{00} \sim N(0, \tau_1^2)$$

$$\text{and } u_{20} \sim N(0, \tau_3^2)$$

A potential mediation effect would be detected if the bootstrapped 95%-confidence interval of the product of the coefficients μ_{2j} and γ_{20} does not include 0 (details for the test can be found here: Pituch, Murphy, & Tate, 2009). As above, this analysis is purely exploratory and does not estimate the efficacy of the intervention itself.

Due to missing data in the post-intervention Grammar Quiz the sample size is reduced for this analysis to $N = 4981$ ($N_{\text{control}} = 2395$; $N_{\text{intervention}} = 2586$) attending 115 schools (58 control; 59 intervention). As with other analyses for the primary outcome the KS1 pretest does have a significant effect on the KS2past results, but neither the Region nor the Grammar Quiz Score do. The average multiplied coefficient to test the mediation hypothesis was .02 with a confidence interval $LB_{ES} = -.02$ and $UB_{ES} = .08$, i.e. no significant mediation was observed.

It has to be noted that in this analysis that the GfW intervention had a small, but significant effect on the outcome: pupils in intervention schools scored on average .46 points less than the pupils in the control schools. Several caveats apply to this finding:

- In models that contain the treatment as well as a proxy for the intermediate treatment outcome (here: teachers' grammar quiz scores) it is not always clear what the direct effects of these two variables mean: while the mediation effect is often well-defined, it is not clear what it means that pupils in the treatment group did worse after controlling for the Grammar Quiz – or the other way round, that they did worse when learning with teachers who had higher scores in the Grammar Quiz after controlling for the intervention. The intervention and the Grammar Quiz are part of a package and cannot have independent effects on the outcome (at least in our study design). Therefore the main focus should be the interpretation of the mediation result.
- The missing Grammar Quiz scores at the teacher level led to a further substantial loss of sample size, which would need (a) a definition of a new outcome set with its own CONSORT logic as well as (b) it's own imputation of the missing data (at least on student level). Since the analysis was only predefined as an additional analysis and no provision of a selection logic for the test (i.e. CONSORT criteria) were provided at the SAP stage at this stage no additional post-hoc analyses was conducted. This means that the finding of a potential treatment effect could also be due to selection effects leading to this specific sample of schools and/or pupils.

Table D.6.3. Bootstrapped coefficients for the estimated model to test a potential mediation effect of the intervention on the Grammar Quiz scores on pupils' KS2past writing paper (KS2past) attainment.

Variable	Average Coefficient (SD)	95% Confidence Interval
Intercept	17.40 (1.04)	15.40, 19.49
KS1 Writing Result	1.15 (.03)	1.10, 1.20
Region	-.08 (.19)	-.47, .30
Treatment	-.46 (.18)	-.83, -.12
Grammar Quiz	-.05 (.05)	-.14, .05

Note. The analysis is based on $b = 1000$ bootstrap samples. $N=4109.41$ ($SD=25.95$); Average N in control = 1953.03 ($SD=18.17$); Average N in intervention = 2156.37 ($SD=18.05$)

Conclusion

Overall the Grammar quiz was a potentially multidimensional variable and it is not entirely clear that it is well-suited for representing teachers' individual differences in grammar knowledge.

Based on the simple assumption that teachers with a more in-depth understanding of grammar should reach higher scores in the quiz composed of highly face-valid questions on grammar, it can be concluded that (a) the quiz did not respond to the intervention on teacher level and (b) that it was also not confirmed that the intervention led to increases in grammar knowledge which in turn led to increases in student performance.

Table D.6.4: Descriptive statistics for the individual items of the Grammar Quiz

	n	Mean	sd	Median	Trimmed	mad	Min	Max	range	skew	kurto	se
P1_Q5_2	297	0.95	0.21	1	1	0	0	1	1	-4.25	16.13	0.01
P1_Q5_3	297	0.84	0.37	1	0.92	0	0	1	1	-1.8	1.23	0.02
P1_Q5_4	297	0.98	0.14	1	1	0	0	1	1	-6.79	44.2	0.01
P1_Q5_5	297	0.96	0.2	1	1	0	0	1	1	-4.44	17.75	0.01
P1_Q5_6	297	0.98	0.13	1	1	0	0	1	1	-7.47	54.03	0.01
P1_Q5_7	297	0.95	0.22	1	1	0	0	1	1	-4.08	14.73	0.01
P1_Q5_8	297	0.71	0.45	1	0.77	0	0	1	1	-0.94	-1.12	0.03
P1_Q5_9	297	0.89	0.31	1	0.98	0	0	1	1	-2.46	4.08	0.02
P1_Q5_10	297	0.57	0.5	1	0.58	0	0	1	1	-0.26	-1.94	0.03
P1_Q5_11	297	0.95	0.22	1	1	0	0	1	1	-4.08	14.73	0.01
P1_Q5_13_1	297	0.83	0.38	1	0.91	0	0	1	1	-1.73	1	0.02
P1_Q5_13_2	297	0.72	0.45	1	0.77	0	0	1	1	-0.98	-1.05	0.03
P1_Q5_13_3	297	0.79	0.4	1	0.87	0	0	1	1	-1.45	0.11	0.02
P1_Q5_13_4	297	0.59	0.49	1	0.61	0	0	1	1	-0.35	-1.89	0.03

P1_Q5_1 3_5	297	0.7	0.46	1	0.75	0	0	1	1	-0.87	-1.25	0.03
P1_Q5_1 4_1	297	0.72	0.45	1	0.78	0	0	1	1	-1	-1.01	0.03
P1_Q5_1 4_2	297	0.97	0.17	1	1	0	0	1	1	-5.45	27.82	0.01
P1_Q5_1 4_3	297	0.99	0.12	1	1	0	0	1	1	-8.4	68.78	0.01
P1_Q5_1 4_4	297	0.98	0.14	1	1	0	0	1	1	-6.79	44.2	0.01
P1_Q5_1 4_5	297	0.49	0.5	0	0.49	0	0	1	1	0.03	-2.01	0.03
P1_Q5_1 6_1	297	0.68	0.47	1	0.72	0	0	1	1	-0.77	-1.41	0.03
P1_Q5_1 6_2	297	0.79	0.41	1	0.86	0	0	1	1	-1.43	0.03	0.02
P1_Q5_1 6_3	297	0.06	0.25	0	0	0	0	1	1	3.55	10.61	0.01
P1_Q5_1 6_4	297	0.66	0.47	1	0.7	0	0	1	1	-0.69	-1.53	0.03
P1_Q5_1 6_5	297	0.35	0.48	0	0.31	0	0	1	1	0.64	-1.6	0.03
P1_Q5_1 6_6	297	0.17	0.38	0	0.09	0	0	1	1	1.73	1	0.02
P1_Q5_1 6_7	297	0.08	0.27	0	0	0	0	1	1	3.15	7.92	0.02
P1_Q5_1 7	297	0.79	0.41	1	0.86	0	0	1	1	-1.4	-0.04	0.02
P1_Q5_1 8	297	0.68	0.47	1	0.72	0	0	1	1	-0.75	-1.44	0.03
P1_Q5_1 9	297	0.25	0.43	0	0.18	0	0	1	1	1.17	-0.62	0.03
P2_Q5_2	222	0.65	0.48	1	0.69	0	0	1	1	-0.64	-1.6	0.03
P2_Q5_3	222	0.78	0.42	1	0.85	0	0	1	1	-1.34	-0.21	0.03

P2_Q5_4	222	0.99	0.12	1	1	0	0	1	1	-8.37	68.37	0.01
P2_Q5_5	222	0.72	0.45	1	0.77	0	0	1	1	-0.95	-1.1	0.03
P2_Q5_6	222	0.95	0.21	1	1	0	0	1	1	-4.36	17.07	0.01
P2_Q5_7	222	0.96	0.2	1	1	0	0	1	1	-4.63	19.5	0.01
P2_Q5_8	222	0.98	0.15	1	1	0	0	1	1	-6.39	39.04	0.01
P2_Q5_9	222	0.98	0.13	1	1	0	0	1	1	-7.2	50.04	0.01
P2_Q5_10	222	0.98	0.13	1	1	0	0	1	1	-7.2	50.04	0.01
P2_Q5_11	222	0.99	0.12	1	1	0	0	1	1	-8.37	68.37	0.01
P2_Q5_13_1	222	0.32	0.47	0	0.28	0	0	1	1	0.77	-1.42	0.03
P2_Q5_13_2	222	0.65	0.48	1	0.69	0	0	1	1	-0.62	-1.62	0.03
P2_Q5_13_3	222	0.65	0.48	1	0.69	0	0	1	1	-0.64	-1.6	0.03
P2_Q5_13_4	222	0.33	0.47	0	0.29	0	0	1	1	0.7	-1.51	0.03
P2_Q5_13_5	222	0.7	0.46	1	0.75	0	0	1	1	-0.88	-1.23	0.03
P2_Q5_14_1	222	0.85	0.36	1	0.93	0	0	1	1	-1.91	1.67	0.02
P2_Q5_14_2	222	0.96	0.19	1	1	0	0	1	1	-4.95	22.56	0.01
P2_Q5_14_3	222	0.77	0.42	1	0.84	0	0	1	1	-1.28	-0.37	0.03
P2_Q5_14_4	222	0.95	0.21	1	1	0	0	1	1	-4.36	17.07	0.01
P2_Q5_14_5	222	0.98	0.13	1	1	0	0	1	1	-7.2	50.04	0.01
P2_Q5_16_1a	222	0.27	0.44	0	0.21	0	0	1	1	1.05	-0.89	0.03

P2_Q5_1 6_2	222	0.19	0.39	0	0.11	0	0	1	1	1.58	0.49	0.03
P2_Q5_1 6_3	222	0.73	0.45	1	0.79	0	0	1	1	-1.03	-0.95	0.03
P2_Q5_1 6_4	222	0.91	0.29	1	1	0	0	1	1	-2.75	5.6	0.02
P2_Q5_1 6_5	222	0.55	0.5	1	0.56	0	0	1	1	-0.18	-1.98	0.03
P2_Q5_1 6_6	222	0.47	0.5	0	0.47	0	0	1	1	0.11	-2	0.03
P2_Q5_1 7	222	0.74	0.44	1	0.8	0	0	1	1	-1.11	-0.78	0.03
P2_Q5_1 8	222	0.86	0.34	1	0.96	0	0	1	1	-2.12	2.51	0.02
P2_Q5_1 9	222	0.05	0.21	0	0	0	0	1	1	4.36	17.07	0.01

D.7. ANALYSIS CODE

This document contains the R code necessary to replicate the analyses presented in the report for the "Grammar for Writing" project. For the analysis R 3.4.2 (Platform: x86_64-w64-mingw32/x64 (64-bit)) was used. A full list of used packages follows below.

Since many of the analyses build on resampling and imputation processes, a one-to-one replication will not always be possible although seed values were used. Because of this, the full set of results was saved in a workspace, which will be uploaded for documentation with the EEF.

Used packages:

```
Amelia
lme4
multilevel
psych
Rcmdr
```

The software GPower (Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior research methods, 39(2), 175-191.) was used for the calculation of the statistic w for the evaluation of balance at baseline.

#Start of analyses

```
####Added on 8th of July
#School-level missing data analysis

library(Rcmdr)

#Read in school level external data
SchoolData <-
  readXL("L:/Jan/000_ORIGINAL/School Level Characteristics_Finalfor
Jan.xlsx",
  rownames=FALSE, header=TRUE, na="", sheet="Sheet1",
  stringsAsFactors=TRUE)

#generate variable that dummy-codes participating and non participating
schools
SchoolData$particip <- ifelse(SchoolData$Status==1, 1, 0)
table(SchoolData$particip)

#define function to report chi-square and cross-tab by participation variable
#sequence of commends taken from Rcmdr
printtable <- function(formuldat) {
  local({
    .Table <- xtabs(formuldat, data=SchoolData)
    cat("\nFrequency table:\n")
    print(.Table)
    cat("\nColumn percentages:\n")
```

```

    print(colPercents(.Table))
    .Test <- chisq.test(.Table, correct=FALSE)
    print(.Test)
  })
} #end of function

names(SchoolData)

printtable(formula(~School.Type+particip))
printtable(formula(~Ofsted.rating+particip))
printtable(formula(~Treatment+particip))

library(psych)
describeBy(SchoolData, group=SchoolData$particip)

hedges <- function(m1, m2, s1, s2, n1, n2) {
s.aster <- sqrt(
  (((n1-1)*(s1^2)) + ((n2-1)*(s2^2))) /
  (n1+n2-2)
) #end of sqrt
return((m1-m2)/s.aster)
} #end of function

#School Size
hedges(377.34, 334.32, 233.01, 202.33, 133, 19)

#Perc FSM
hedges(26.20, 23.98, 13.4, 14.23, 133, 19)

#GRammar
hedges(.67, .71, .18, .18, 134, 19)

#Writing
hedges(.77, .80, .18, .15, 134, 19)

#Reading
hedges(.67, .74, .17, .16, 134, 19)

#Maths
hedges(.72, .79, .17, .10, 134, 19)

#EAL
hedges(21.24, 29.75, 25.61, 32.26, 133, 19)

#SEN
hedges(1.43, .88, 1.61, .76, 133, 19)

#####
#End of school level analyses
#####

#Main analysis
#Loading R Commander to read in the data
library(Rcmdr)

#loading core data set from the Excel file
#Analysis Dataset - Pupil v4 20180423 with scaled scores.xlsx
GfWdata <-
readXL("L:/Jan/000_ORIGINAL/Analysis Dataset - Pupil v4 20180423 with scaled
scores.xlsx",
  rownames=FALSE, header=TRUE, na="", sheet="Blindeddataset",

```



```

stringsAsFactors=TRUE)

#checking properties
head(GfWdata)
names(GfWdata)

#Merge with region data from Excel file with region allocation
#Dataset_Region 20180411.xlsx
GfWregions <-
  readXL("L:/Jan/000_ORIGINAL/Dataset_Region 20180411.xlsx",
         rownames=FALSE, header=TRUE, na="",
sheet="qry_Dataset_Region_AnalysisScho",
stringsAsFactors=TRUE)

#checking properties and merging the two datasets
head(GfWregions)
names(GfWregions)
names(GfWdata)
names(GfWregions)[1] <- "SchoolID2"
names(GfWregions)
GfWdata2 <- merge(GfWdata, GfWregions, by="SchoolID2")
names(GfWdata2)

#merge with numeric writing outcome which was supplied in
# Dataset_WRITTAOUTCOME_Coded 20180411.xlsx
GfWwrit <-
  readXL("L:/Jan/000_ORIGINAL/Dataset_WRITTAOUTCOME_Coded 20180411.xlsx",
         rownames=FALSE, header=TRUE, na="", sheet="Blindeddataset",
stringsAsFactors=TRUE)

#checking properties and merging the two dataframes
head(GfWwrit)
names(GfWwrit)
names(GfWwrit)[2] <- "KS2_WRITTAOUTCOMEcopy"
names(GfWwrit)
names(GfWdata2)
GfWdata3 <- merge(GfWdata2, GfWwrit, by="UOYSTID2")
names(GfWdata3)
head(GfWdata3)

#GfWdata3 contains all available data (apart from teacher outcomes)
#The other two dataframes are deleted
rm(GfWdata, GfWdata2)

#recode missing and non-applicable data to R internal values
GfWdata3[GfWdata3=="-88"] <- NA
GfWdata3[GfWdata3=="-99"] <- NA
nrow(GfWdata3)
#This data set has all N=7239 rows from allocation

####
#Table 7 - baseline comparison of pupil level data:

#Number of Y6 students
table(GfWdata3$BlindTreatment2)

#average KS1 result
#identification of coding error in data set
#zeros need to be recoded as missing values
table(GfWdata3$KS1_WRITPOINTS)
GfWdata3$origKS1_WRITPOINTS <- GfWdata3$KS1_WRITPOINTS

```

```
GfWdata3$KS1_WRITPOINTS[GfWdata3$KS1_WRITPOINTS==0] <- NA
table(GfWdata3$KS1_WRITPOINTS)

#check whether variables have been correctly created
table(is.na(GfWdata3$KS1_WRITPOINTS), GfWdata3$BlindTreatment2)
table(GfWdata3$origKS1_WRITPOINTS, GfWdata3$BlindTreatment2)

#overlap of baseline KS1 with primary outcome
table(GfWdata3$origKS1_WRITPOINTS, is.na(GfWdata3$CalcTotal_Overall2))

#Students with at least one missing value
head(GfWdata3[, c(3, 5, 6, 10)])
mis.sum <- GfWdata3[, c(3, 5, 6, 10)]
mis.sum$missing <- rowSums(is.na(mis.sum))
mis.sum$any mis <- ifelse(mis.sum$missing>0,1,0)
table(mis.sum$any mis, mis.sum$BlindTreatment2)
sum(table(mis.sum$any mis, mis.sum$BlindTreatment2))
rm(mis.sum)

#Students eligible for FSM
table(is.na(GfWdata3$EVERFSM_ALL_SPR17), GfWdata3$BlindTreatment2)
table(GfWdata3$EVERFSM_ALL_SPR17, GfWdata3$BlindTreatment2)

#Balance checked with GPower calculation
      5      6
0 1882 2039
1 1541 1734
> 1882/(1882+1541)
[1] 0.5498101
> 1-(1882/(1882+1541))
[1] 0.4501899
> 2039/(2039+1734)
[1] 0.5404188
> 1-(2039/(2039+1734))
[1] 0.4595812
```

The screenshot shows the G*Power software interface. On the left, the 'Input Parameters' section is active, with 'Determine =>' selected. The parameters are: Effect size w (0.0188939), alpha error probability (0.05), Power (1-beta error probability) (0.95), and Df (5). The 'Output Parameters' section shows Noncentrality parameter lambda, Critical chi-square, Total sample size, and Actual power, all with question marks. On the right, a table displays p(H0) and p(H1) values for two cells. The 'Calculate and transfer to main window' button is highlighted.

Cell	p(H0)	p(H1)
1	0.5498	0.5404
2	0.4502	0.4596

```
#Gender
table(is.na(GfWdata3$Gender), GfWdata3$BlindTreatment2)
table(GfWdata3$Gender, GfWdata3$BlindTreatment2)

#KS2 WRITTAOUTCOME in raw format for table 7
table(GfWdata3$KS2_WRITTAOUTCOME)
sum(table(GfWdata3$KS2_WRITTAOUTCOME))
```

```

table(is.na(GfWdata3$KS2_WRITTAOUTCOME), GfWdata3$BlindTreatment2)
table(GfWdata3$KS2_WRITTAOUTCOME, GfWdata3$BlindTreatment2)

#percentages
#control
(round(table(GfWdata3$KS2_WRITTAOUTCOME,
GfWdata3$BlindTreatment2)[,1]/3406,3))*100

#intervention
(round(table(GfWdata3$KS2_WRITTAOUTCOME,
GfWdata3$BlindTreatment2)[,2]/3757,3))*100

#KS1 pretest at baseline
#check of baseline data in line with EEF Analysis Guidance
table(is.na(GfWdata3$KS1_WRITPOINTS), GfWdata3$BlindTreatment2)
#control
mean(GfWdata3$KS1_WRITPOINTS[GfWdata3$BlindTreatment2==5], na.rm=T)
sd(GfWdata3$KS1_WRITPOINTS[GfWdata3$BlindTreatment2==5], na.rm=T)
#intervention
mean(GfWdata3$KS1_WRITPOINTS[GfWdata3$BlindTreatment2==6], na.rm=T)
sd(GfWdata3$KS1_WRITPOINTS[GfWdata3$BlindTreatment2==6], na.rm=T)

#calculate pooled SD for baseline pre-test
table(is.na(GfWdata3$KS1_WRITPOINTS), GfWdata3$BlindTreatment2)
pool.d.var <- (
(3572* var(GfWdata3$KS1_WRITPOINTS[GfWdata3$BlindTreatment2==6], na.rm=T))
+
(3254* var(GfWdata3$KS1_WRITPOINTS[GfWdata3$BlindTreatment2==5], na.rm=T)))
/ #end numerator
(3254+3572-2)
#effect size Hedges g, not corrected for clustering
(mean(GfWdata3$KS1_WRITPOINTS[GfWdata3$BlindTreatment2==6], na.rm=T) -
mean(GfWdata3$KS1_WRITPOINTS[GfWdata3$BlindTreatment2==5], na.rm=T)) /
sqrt(pool.d.var)

min(GfWdata3$KS1_WRITPOINTS[GfWdata3$BlindTreatment2==5], na.rm=T)
max(GfWdata3$KS1_WRITPOINTS[GfWdata3$BlindTreatment2==5], na.rm=T)
min(GfWdata3$KS1_WRITPOINTS[GfWdata3$BlindTreatment2==6], na.rm=T)
max(GfWdata3$KS1_WRITPOINTS[GfWdata3$BlindTreatment2==6], na.rm=T)

#pre-test data distribution for all pupils
#as requested in guidance, p. 2
#histograms by group
par(mfrow=c(1,2))
hist(GfWdata3$KS1_WRITPOINTS[GfWdata3$BlindTreatment2==5], main="Control
Group", ylab="Density", xlab="KS1 Writing point score", xlim=c(0,30),
ylim=c(0,1200),
border="white", , col="lightgrey")

hist(GfWdata3$KS1_WRITPOINTS[GfWdata3$BlindTreatment2==6],
main="Intervention Group", ylab="Density", xlab="KS1 Writing point score",
xlim=c(0,30) , ylim=c(0,1200),
border="white", , col="lightgrey")

#####
#####
#####

#For primary outcome analyses we agreed via email on the 23.04.2018:
#Only the n=135 schools, N=5182 for PRIMARY
#All schools for SECONDARY analysis

```

```

#We predict RAW scores in secondary individual analyses

#check availability of demographic data
sum(is.na(GfWdata3$EVERFSM_ALL_SPR17))
sum(is.na(GfWdata3$Gender))
table(is.na(GfWdata3$EVERFSM_ALL_SPR17), is.na(GfWdata3$Gender))
#code exclusion vector
mis.gender <- is.na(GfWdata3$Gender)
mis.ever <- is.na(GfWdata3$EVERFSM_ALL_SPR17)
misdat <- data.frame(cbind(mis.gender, mis.ever))
misdat$exclvect <- rowSums(misdat)
table(misdat$exclvect)

#secondary analysis data frame
#here the analysis frame with N=7200 observations is defined.
GfWdata3.sec <- GfWdata3[misdat$exclvect<2, ]
length(GfWdata3.sec$Gender)
#this results in 7200 analysable cases

#remove these objects
rm(misdat, GfWdata3)

#Descriptive analyses of school level
attach(GfWdata3.sec)
#schools
table(SchoolID2)
length(table(SchoolID2))
mean(table(SchoolID2))
sd(table(SchoolID2))
length(SchoolID2)
detach(GfWdata3.sec)

#MISSING DATA APPENDIX
#Define function to read out descriptives and missing data
descr.data <- function(x) {
hist(x, main=names(x))
cat("Mean: ", round(mean(x, na.rm=T), 4), "\n")
cat("Median: ", round(median(x, na.rm=T), 4), "\n")
cat("SD: ", round(sd(x, na.rm=T), 4), "\n")
cat("N missing values: ", sum(is.na(x)), "\n")
} #end of descriptive function

descr.data(GfWdata3.sec$CalcTotal_Overall2)
table(GfWdata3.sec$CalcTotal_Overall2)

descr.data(GfWdata3.sec$KS2_WRITTAOUTCOME_Code)
table(GfWdata3.sec$KS2_WRITTAOUTCOME_Code)

descr.data(GfWdata3.sec$KS2_GPSMRK)
table(GfWdata3.sec$KS2_GPSMRK)

descr.data(GfWdata3.sec$KS2_READMRK)
table(GfWdata3.sec$KS2_READMRK)

descr.data(GfWdata3.sec$KS1_WRITPOINTS)
table(GfWdata3.sec$KS1_WRITPOINTS)

table(GfWdata3.sec$BlindTreatment2)
sum(is.na(GfWdata3.sec$BlindTreatment2))

table(GfWdata3.sec$Gender)

```

```

sum(is.na(GfWdata3.sec$Gender))

table(GfWdata3.sec$Region)
sum(is.na(GfWdata3.sec$Region))

table(GfWdata3.sec$EVERFSM_ALL_SPR17)
sum(is.na(GfWdata3.sec$EVERFSM_ALL_SPR17))
table(GfWdata3.sec$BlindTreatment2, GfWdata3.sec$EVERFSM_ALL_SPR17)

#this is basically the same sample as above, but nevertheless:
#calculated w with GPower:
      0      1
      5 1882 1541
      6 2039 1734
> 1882/(1882+1541)
[1] 0.5498101
> 1-(1882/(1882+1541))
[1] 0.4501899
> 2039/(2039+1734)
[1] 0.5404188
> 1-(2039/(2039+1734))
[1] 0.4595812
>

```

Cell	p(H0)	p(H1)
1	0.5498	0.5404
2	0.4502	0.4596

```

#Coding missing data
library(Amelia)
missmap(GfWdata3.sec[,c(5, 6, 10,15,16,19)], legend=F, main="Missingness
Secondary Outcomes", y.cex=0)
detach(package:Amelia)

#any pre values missing?
miss.pre <- data.frame(cbind(
is.na(GfWdata3.sec$Gender), is.na(GfWdata3.sec$EVERFSM_ALL_SPR17),
is.na(GfWdata3.sec$KS1_WRITPOINTS))
)
miss.pre$miss1 <- rowSums(miss.pre)
GfWdata3.sec$miss1 <- ifelse(miss.pre$miss1>0, 1,0)
table(GfWdata3.sec$miss1)

#checking bottom part of CONSORT fpr secondary outcomes
consort.sec <- subset(GfWdata3.sec, is.na(KS2_WRITTAOUTCOME_Code)==F)
table(consort.sec$BlindTreatment2, consort.sec$miss1)

consort.sec <- subset(GfWdata3.sec, is.na(KS2_READMRK)==F)
table(consort.sec$BlindTreatment2, consort.sec$miss1)

```

```

consort.sec <- subset(GfWdata3.sec, is.na(KS2_GPSMRK)==F)
table(consort.sec$BlindTreatment2, consort.sec$miss1)

#without primary // only secondary outcome
miss.post <- data.frame(cbind(
is.na(GfWdata3.sec$KS2_WRITTAOUTCOME_Code), is.na(GfWdata3.sec$KS2_GPSMRK),
is.na(GfWdata3.sec$KS2_READMRK)
))

miss.post$miss2 <- rowSums(miss.post)
GfWdata3.sec$miss2 <- ifelse(miss.post$miss2>0, 1,0)
table(GfWdata3.sec$miss2)

#primary outcome (already documented in table)
GfWdata3.sec$miss.prim <- is.na(GfWdata3.sec$CalcTotal_Overall2)
table(GfWdata3.sec$miss.prim)
rm(miss.pre, miss.post)

#school-level dataset
GfWdata.schools <- aggregate(GfWdata3.sec, by=list(GfWdata3.sec$SchoolID2),
FUN=mean, na.rm=T)

fix(GfWdata.schools)
max(GfWdata.schools$miss1)
max(GfWdata.schools$miss2)
#observed maximum within a school is 19%
#but not in the primary outcome alone, which was condition in SAP

#save for school-level missing data presentation:
#write.table(GfWdata.schools,
"L:\\Jan\\001_Analysis\\Aggregate_Secondary.txt", sep="\t", row.names=F)

#check of missing and analysable cases for primary outcome
GfWdata.schools$miss.prim
#several schools have more than the pre-specified threshold

#double check schools with specific missing on primary outcome
missing.primary <- subset(GfWdata.schools, select=c("SchoolID2",
"miss.prim"))

#dataset for primary analyses: how many schools reported at least one?
gfwdata3.test <- GfWdata3.sec
gfwdata3.test <- merge(gfwdata3.test, missing.primary, by="SchoolID2")
#now only schools that reported at least one case
gfwdata3.test <- subset(gfwdata3.test, gfwdata3.test$miss.prim.y!=1)
length(unique(gfwdata3.test$SchoolID2))

#primary analysis data frame
#start of primary outcome data set
nrow(gfwdata3.test)
#this dataset has N=6306 rows
#which was confirmed by Louise E
table(gfwdata3.test$BlindTreatment2)

#here the dataframe for the primary outcome analyses is produced
#it is a sub-sample of the dataframe for the secondary analysis
GfWdata3.prim <- gfwdata3.test
length(GfWdata3.prim$Gender)

#Descriptive analyses of school level
attach(GfWdata3.prim)

```

```
#schools
table(SchoolID2)
length(table(SchoolID2))
mean(table(SchoolID2))
sd(table(SchoolID2))
length(SchoolID2)
detach(GfWdata3.prim)

descr.data(GfWdata3.prim$CalcTotal_Overall2)
table(GfWdata3.prim$CalcTotal_Overall2)
#perc missing
890/6306

descr.data(GfWdata3.prim$KS2_WRITTAOUTCOME_Code)
table(GfWdata3.prim$KS2_WRITTAOUTCOME_Code)

descr.data(GfWdata3.prim$KS2_GPSMRK)
table(GfWdata3.prim$KS2_GPSMRK)

descr.data(GfWdata3.prim$KS2_READMRK)
table(GfWdata3.prim$KS2_READMRK)

descr.data(GfWdata3.prim$KS1_WRITPOINTS)
table(GfWdata3.prim$KS1_WRITPOINTS)

table(GfWdata3.prim$BlindTreatment2)
sum(is.na(GfWdata3.prim$BlindTreatment2))

table(GfWdata3.prim$Gender)
sum(is.na(GfWdata3.prim$Gender))

table(GfWdata3.prim$Region)
sum(is.na(GfWdata3.prim$Region))

table(GfWdata3.prim$EVERFSM_ALL_SPR17)
sum(is.na(GfWdata3.prim$EVERFSM_ALL_SPR17))
table(GfWdata3.prim$BlindTreatment2, GfWdata3.prim$EVERFSM_ALL_SPR17)

#again calculation of w with GPower:
> 1601/(1601+1358)
[1] 0.5410612
> 1-(1601/(1601+1358))
[1] 0.4589388
> 1777/(1777+1566)
[1] 0.5315585
> 1-(1777/(1777+1566))
[1] 0.4684415

#Result then again calculated in GPower
```

Cell	p(H0)	p(H1)
1	0.5411	0.5316
2	0.4589	0.4684

Test family		Statistical test	
χ ² tests		Goodness-of-fit tests: Contingency tables	
Type of power analysis			
A priori: Compute required sample size - given α, power, and effect size			
Input Parameters		Output Parameters	
Determine =>		Noncentrality parameter λ	?
Effect size w	0.0190645	Critical χ ²	?
α err prob	0.05	Total sample size	?
Power (1-β err prob)	0.95	Actual power	?
Df	5		

0.5	0.5
Equal p(H0)	Equal p(H1)
Normalize p(H0)	Normalize p(H1)
Auto calc last cell	Auto calc last cell
Calculate	Effect size w 0.01906452

```
#Several of the EEF outcome tables need report of sample mean and CI
#summary stats for outcome tables
#definition of function to produce confidence interval
mean.ci <- function(x) {
  holdmean <- mean(x, na.rm=T)
  holdse <- sd(x, na.rm=T)/sqrt(sum(!is.na(x)))
  lower <- holdmean - (qnorm(.975)*holdse)
  upper <- holdmean + (qnorm(.975)*holdse)
  return(round(c(lower, upper),3))
} #end of function

nrow(GfWdata3.prim)

#TABLE 8
#intervention
psych::describe(subset(GfWdata3.prim,
BlindTreatment2==6)$CalcTotal_Overall2)
mean.ci(subset(GfWdata3.prim, BlindTreatment2==6)$CalcTotal_Overall2)

#control
psych::describe(subset(GfWdata3.prim,
BlindTreatment2==5)$CalcTotal_Overall2)
mean.ci(subset(GfWdata3.prim, BlindTreatment2==5)$CalcTotal_Overall2)

#TABLE 9
#WRITTAOUT intervention
psych::describe(subset(GfWdata3.prim,
BlindTreatment2==6)$KS2_WRITTAOUTCOME_Code)
mean.ci(subset(GfWdata3.prim, BlindTreatment2==6)$KS2_WRITTAOUTCOME_Code)

#control
psych::describe(subset(GfWdata3.prim,
BlindTreatment2==5)$KS2_WRITTAOUTCOME_Code)
mean.ci(subset(GfWdata3.prim, BlindTreatment2==5)$KS2_WRITTAOUTCOME_Code)

#TABLE 9
#READMRK intervention
psych::describe(subset(GfWdata3.prim, BlindTreatment2==6)$KS2_READMRK)
mean.ci(subset(GfWdata3.prim, BlindTreatment2==6)$KS2_READMRK)

#READMRK control
psych::describe(subset(GfWdata3.prim, BlindTreatment2==5)$KS2_READMRK)
mean.ci(subset(GfWdata3.prim, BlindTreatment2==5)$KS2_READMRK)

#TABLE 9
```



```

#GPSMRK intervention
psych::describe(subset(GfWdata3.prim, BlindTreatment2==6)$KS2_GPSMRK)
mean.ci(subset(GfWdata3.prim, BlindTreatment2==6)$KS2_GPSMRK)

#GPSMRK control
psych::describe(subset(GfWdata3.prim, BlindTreatment2==5)$KS2_GPSMRK)
mean.ci(subset(GfWdata3.prim, BlindTreatment2==5)$KS2_GPSMRK)

#
table(GfWdata3.prim$BlindTreatment2, GfWdata3.prim$EVERFSM_ALL_SPR17)

#intervention
psych::describe(subset(GfWdata3.prim, ((BlindTreatment2==6) &
(EVERFSM_ALL_SPR17==1)))$CalcTotal_Overall2)
mean.ci(subset(GfWdata3.prim, ((BlindTreatment2==6) &
(EVERFSM_ALL_SPR17==1)))$CalcTotal_Overall2)

#control
psych::describe(subset(GfWdata3.prim, ((BlindTreatment2==5) &
(EVERFSM_ALL_SPR17==1)))$CalcTotal_Overall2)
mean.ci(subset(GfWdata3.prim, ((BlindTreatment2==5) &
(EVERFSM_ALL_SPR17==1)))$CalcTotal_Overall2)

#Coding missing data in the primary analysis file
library(Amelia)
missmap(GfWdata3.prim[,c(5, 6, 10, 13, 15,16)], legend=F, main="Missingness
Primary Outcomes", y.cex=0)
detach(package:Amelia)

#checking missingness patterns in primary outcome data set
names(GfWdata3.prim)
#missing at baseline
table(GfWdata3.prim$miss1)
table(GfWdata3.prim$miss2)
table((GfWdata3.prim$miss.prim.x + GfWdata3.prim$miss2))

#school level data and missings
GfWdata.schools.prim <- aggregate(GfWdata3.prim,
by=list(GfWdata3.prim$SchoolID2), FUN=mean, na.rm=T)
fix(GfWdata.schools.prim)
#save for school-level missing data presentation:
#write.table(GfWdata.schools,
"L:\\Jan\\001_Analysis\\Aggregate_Primary.txt", sep="\t", row.names=F)

#####
#code for Table 6: Minimum detectable effect size at different stages

library(psych)
names(GfWdata3.prim)

#Determine within and between correlations
#as well as ICCs
calc.dat <- subset(GfWdata3.prim, miss1==0)
calc.dat <- subset(calc.dat, miss.prim.x==0)
table(calc.dat$BlindTreatment2)

#average cluster size
mean(table(calc.dat$SchoolID))

#Full sample
deaggr.corr <- statsBy(calc.dat[, c(1,10,13)], "SchoolID2")

```

```

print(deaggr.corr, short=F)
deaggr.corr$n
deaggr.corr$nG

#N used for calculation
sum(deaggr.corr$n[,3])

#FSM only
calc.dat <- subset(calc.dat, EVERFSM_ALL_SPR17==1)
fsm.only.deaggrcorr <- statsBy(calc.dat[, c(1,10,13)], "SchoolID2")
print(fsm.only.deaggrcorr, short=F)

#N used for calculation
sum(fsm.only.deaggrcorr$n[,3])
table(calc.dat$BlindTreatment2)
mean(table(calc.dat$SchoolID))

rm(calc.dat)

#####
#MDES calculations

$$MDES = M_{J-K} \sqrt{\frac{\rho(1-R_2^2)}{P(1-P)J} + \frac{(1-\rho)(1-R_1^2)}{P(1-P)nJ}}$$

#Formula for original power analysis
#For the columns "Protocol"
2.85* sqrt(
  ((.15*(1-(0)))/(.5*.5*150)) + #we assumed that stratification vars would
  explain 10%, not-pre-post test correlation between schools!
  ((.85*(1-(.53)))/(.5*.5*7500))
)
75*50
75*16
sqrt(.1)
sqrt(.53)

#Protocol - FSM
2.85* sqrt(
  ((.15*(1-(0)))/(.5*.5*150)) + #we assumed that stratification variables
  would explain 10%, not-pre-post test correlation between schools!
  ((.85*(1-(.53)))/(.5*.5*150*16))
)
16*150

#last two columns!
#Analysis - overall
2.85* sqrt(
  ((.12*(1-(0)))/(.5*.5*135)) + #pre-test not used as a predictor
  ((.88*(1-(.60^2)))/(.5*.5*5182))
)

#Analysis - FSM
2.85* sqrt(
  ((.10*(1-(0)))/(.5*.5*134)) + ##not used as a predictor
  ((.90*(1-(.60^2)))/(.5*.5*2362))
)

#evaluating impact of ICC/ why are MDES so good?
#last two columns!
#Analysis - overall
2.85* sqrt(

```

```

((.15*(1-(0)))/(.5*.5*135)) + #pre-test not used as a predictor
((.85*(1-(.60^2)))/(.5*.5*5182))
)

#Analysis - FSM
2.85* sqrt(
((.15*(1-(0)))/(.5*.5*134)) + ##not used as a predictor
((.85*(1-(.60^2)))/(.5*.5*2362))
)

#####
#for the randomisation the original data set needs to be read in again:
#these must be all pupils for whom in principle data would have been available

#loading core data set
GfWdata <-
  readXL("L:/Jan/000_ORIGINAL/Analysis Dataset - Pupil v4 20180423 with
scaled scores.xlsx",
  rownames=FALSE, header=TRUE, na="", sheet="Blindeddataset",
  stringsAsFactors=TRUE)

#checking properties
head(GfWdata)
names(GfWdata)

table(GfWdata$BlindTreatment2)

table(GfWdata$EVERFSM_ALL_SPR17)
table(GfWdata$BlindTreatment2[GfWdata$EVERFSM_ALL_SPR17==1])

#average cluster sizes secondary/as randomised
table(GfWdata$SchoolID)
mean(table(GfWdata$SchoolID))
#FSM-only
table(GfWdata$SchoolID[GfWdata$EVERFSM_ALL_SPR17==1])
mean(table(GfWdata$SchoolID[GfWdata$EVERFSM_ALL_SPR17==1]))

#Randomisation - overall
2.85* sqrt(
((.15*(1-(0)))/(.5*.5*155)) +
((.85*(1-(.73^2)))/(.5*.5*7239))
)

#Randomisation - FSM
2.85* sqrt(
((.15*(1-(0)))/(.5*.5*155)) +
((.85*(1-(.73^2)))/(.5*.5*3275))
)

rm(GfWdata)
detach(package:psych)

#####
#Preparing Outcome analyses

#coding treatment variable correctly in both datasets
# (6=1) Intervention
# (5=0) Control

GfWdata3.prim$treat <- ifelse(GfWdata3.prim$BlindTreatment2==5, 0, 1)
table(GfWdata3.prim$BlindTreatment2, GfWdata3.prim$treat)

```

```

GfWdata3.sec$treat <- ifelse(GfWdata3.sec$BlindTreatment2==5, 0, 1)
table(GfWdata3.sec$BlindTreatment2, GfWdata3.sec$treat)

#Dry-run of model
#first hlm, test-run
library(lme4)
#repeat primary outcome estimation
prim.model.test <- lmer(CalcTotal_Overall2 ~ KS1_WRITPOINTS + Region + treat
+ (1| SchoolID2), data=GfWdata3.prim)
summary(prim.model.test)

#test object slots to read:
#for N in contr + intervention
length(prim.model.test@frame$treat)

rm(prim.model.test)

#test-run in fsm-only for analysable N
table(GfWdata3.prim$EVERFSM_ALL_SPR17)
#N=2924 observed cases
check.fsm <- subset(GfWdata3.prim, EVERFSM_ALL_SPR17==1)
prim.model.fsm <- lmer(CalcTotal_Overall2 ~ KS1_WRITPOINTS + Region + treat
+ (1| SchoolID2), data=check.fsm)
summary(prim.model.fsm)
rm(check.fsm, prim.model.fsm)

#Grand mean centring of continuous predictor in both datasets:
mean(GfWdata3.prim$KS1_WRITPOINTS, na.rm=T)
sd(GfWdata3.prim$KS1_WRITPOINTS , na.rm=T)
GfWdata3.prim$origKS1_WRITPOINTS <- GfWdata3.prim$KS1_WRITPOINTS
GfWdata3.prim$KS1_WRITPOINTS <- scale(GfWdata3.prim$KS1_WRITPOINTS,
center=T, scale=F)
round(mean(GfWdata3.prim$KS1_WRITPOINTS , na.rm=T), 3)
sd(GfWdata3.prim$KS1_WRITPOINTS , na.rm=T)

mean(GfWdata3.sec$KS1_WRITPOINTS, na.rm=T)
sd(GfWdata3.sec$KS1_WRITPOINTS , na.rm=T)
GfWdata3.sec$origKS1_WRITPOINTS <- GfWdata3.sec$KS1_WRITPOINTS
GfWdata3.sec$KS1_WRITPOINTS <- scale(GfWdata3.sec$KS1_WRITPOINTS, center=T,
scale=F)
round(mean(GfWdata3.sec$KS1_WRITPOINTS , na.rm=T), 3)
sd(GfWdata3.sec$KS1_WRITPOINTS , na.rm=T)

#LIBRARIES
library(lme4) #should already be loaded
library(multilevel)

#bootstrap of PRIMARY OUTCOME
#full sample, not imputed
#define variables capturing estimates:
ICC.KS1 <- NULL
ICC.KS2 <- NULL
primary.coeffs <- NULL
effect.size <- NULL
table.7 <- NULL
primary.n <- NULL

t1 <- Sys.time()
#define consecutive numbers for bootstrap selection within schools
GfWdata3.prim$number <- 1:nrow(GfWdata3.prim)

```

```

#Actual bootstrap
for (boot in 1:1000) {
#generate list of student codes to select from
set.seed<-boot
id.list <- tapply(GfWdata3.prim$number, INDEX=GfWdata3.prim$SchoolID2,
sample, replace=T)
id.list <- unlist(id.list)
boot.data <- GfWdata3.prim[id.list, ]

#average in each group, n in each group with KS2past
table.7 <- rbind(table.7, c(
mean(boot.data$CalcTotal_Overall2[boot.data$treat==0], na.rm=T),
mean(boot.data$CalcTotal_Overall2[boot.data$treat==1], na.rm=T),
var(boot.data$CalcTotal_Overall2[boot.data$treat==0], na.rm=T),
var(boot.data$CalcTotal_Overall2[boot.data$treat==1], na.rm=T),
sum(!is.na(boot.data$CalcTotal_Overall2[boot.data$treat==0])),
sum(!is.na(boot.data$CalcTotal_Overall2[boot.data$treat==1]))
)) #end of table collector

#ICC estimation
ks1.anova <- aov(KS1_WRITPOINTS ~ as.factor(SchoolID2), data=boot.data)
ks2.anova <- aov(CalcTotal_Overall2 ~ as.factor(SchoolID2), data=boot.data)
ICC.KS1[[boot]] <- ICC1(ks1.anova)
ICC.KS2[[boot]] <- ICC1(ks2.anova)
rm(ks1.anova, ks2.anova)

#Formula in SAP
#KS2past ~ KS1 + REG + GfW + (1|school)
prim.model <- lmer(CalcTotal_Overall2 ~ KS1_WRITPOINTS + Region + treat +
(1| SchoolID2), data=boot.data)
primary.coeffs <- rbind(primary.coeffs, fixef(prim.model))
primary.n <- rbind(primary.n, c(length(prim.model@frame$treat),
sum(prim.model@frame$treat)))

varcomp <- data.frame(VarCorr(prim.model))
#collecting coefficient separately and the two variances
effect.size <- rbind(effect.size, c(
summary(prim.model)$coefficients[4,1], varcomp[1,4], varcomp[2,4]
))

rm(prim.model, boot.data, varcomp)

#plot as progress bar
if (boot%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap primary outcome
Sys.time() - t1 #~7mins

#determine N per run in groups
primary.n <- data.frame(primary.n)
names(primary.n) <- c("total", "intervention")
primary.n$control <- primary.n$total - primary.n$intervention
apply(primary.n, MARGIN=2, mean)
apply(primary.n, MARGIN=2, sd)

#
#Fist the intra-class correlations
mean(ICC.KS1)

```

```

sd(ICC.KS1)
mean(ICC.KS2)
sd(ICC.KS2)

#Coefficients
head(primary.coeffs)
apply(primary.coeffs, MARGIN=2, mean)
apply(primary.coeffs, MARGIN=2, sd)
apply(primary.coeffs, MARGIN=2, quantile, probs=c(.025, .975))
hist(primary.coeffs[,4])

#HLM effect size estimate
head(effect.size)
effect.size <- data.frame(effect.size)
names(effect.size) <- c("treat", "var.l2", "var.l1")
#variance for effect size denominator
effect.size$variance <- effect.size$var.l2 + effect.size$var.l1
effect.size$estimate <- effect.size$treat/sqrt(effect.size$variance)
quantile(effect.size$estimate, probs=c(.025, .975))
mean(effect.size$estimate)

#Variance components
apply(effect.size, MARGIN=2, mean)

#For table 7 in main report
#it was orinally table 7, it is now 8 and following
#this is mainly to look at the raw data without cluster corrections
#as requested by template
head(table.7)
table.7 <- data.frame(table.7)
names(table.7) <- c("mean.contr", "mean.inter", "var.contr", "var.inter",
"n.contr", "n.inter")

mean(table.7$n.inter)
mean(table.7$mean.inter)
quantile(table.7$mean.inter, probs=c(.025, .975))
mean(table.7$n.contr)
mean(table.7$mean.contr)
quantile(table.7$mean.contr, probs=c(.025, .975))
mean(table.7$n.contr + table.7$n.inter)
#unadjusted Hedges g
var.asterisk <- (((table.7$n.contr-1) * table.7$var.contr) +
                ((table.7$n.inter-1) * table.7$var.inter)) /
                (table.7$n.inter + table.7$n.contr - 2)

g.test <- ((table.7$mean.inter - table.7$mean.contr) /
           sqrt(var.asterisk))

mean(g.test)
quantile(g.test, probs=c(.025, .975))
rm(g.test, var.asterisk)

#define function to read and produce data for table 7 from results
#put in a result matrix
#For table 7 (now 8 and ff) in main report
#this is raw data without cluster or missingness taken into account
read.reportdat <- function(x) {

table7 <- x
table7 <- data.frame(table7)

```

```

names(table7) <- c("mean.contr", "mean.inter", "var.contr", "var.inter",
"n.contr", "n.inter")

cat("Ave N intervention: ", mean(table7$n.inter), "\n", "\n")

cat("Ave dep var intervention: ", mean(table7$mean.inter), "\n", "\n")

cat("Quantiles intervention", quantile(table7$mean.inter, probs=c(.025,
.975)), "\n", "\n")
cat("Ave N contr: ", mean(table7$n.contr), "\n", "\n")
cat("Ave dep var contr: ", mean(table7$mean.contr), "\n", "\n")
cat("Quantiles control", quantile(table7$mean.contr, probs=c(.025, .975)),
"\n", "\n")
cat("N overall: ", mean(table7$n.contr + table7$n.inter), "\n", "\n")
#unadjusted Hedges g
var.asterisk <- (((table7$n.contr-1) * table7$var.contr) +
((table7$n.inter-1) * table7$var.inter)) /
(table7$n.inter + table7$n.contr - 2)

g.test <- ((table7$mean.inter - table7$mean.contr) /
sqrt(var.asterisk))

cat("Hedges G: ", mean(g.test), "\n", "\n")
cat("Quantile Hedges G", quantile(g.test, probs=c(.025, .975)), "\n", "\n")
rm(g.test, var.asterisk)

} #end of function

#bootstrap of PRIMARY OUTCOME ONLY IN FSM
#Not imputed
#define variables capturing estimates:
ICC.KS1.fsm <- NULL
ICC.KS2.fsm <- NULL
primary.coeffs.fsm <- NULL
effect.size.fsm <- NULL
primary.n.fsm <- NULL
table.7fsm <- NULL

#define consecutive numbers for bootstrap selection within schools
GfWdata3.prim$number <- 1:nrow(GfWdata3.prim)

t1 <- Sys.time()
#Actual bootstrap
for (boot in 1:1000) {
#generate list of student codes to select from
set.seed<-boot
id.list <- tapply(GfWdata3.prim$number, INDEX=GfWdata3.prim$SchoolID2,
sample, replace=T)
id.list <- unlist(id.list)
boot.data <- GfWdata3.prim[id.list, ]
boot.data <- boot.data
boot.data <- subset(boot.data, EVERFSM_ALL_SPR17==1)

#average in each group, n in each group with KS2past
table.7fsm <- rbind(table.7fsm,
c(
mean(boot.data$CalcTotal_Overall12[boot.data$treat==0], na.rm=T),
mean(boot.data$CalcTotal_Overall12[boot.data$treat==1], na.rm=T),
var(boot.data$CalcTotal_Overall12[boot.data$treat==0], na.rm=T),
var(boot.data$CalcTotal_Overall12[boot.data$treat==1], na.rm=T),
sum(!is.na(boot.data$CalcTotal_Overall12[boot.data$treat==0])),
sum(!is.na(boot.data$CalcTotal_Overall12[boot.data$treat==1]))))

```

```

)) #end of table collector

ks1.anova <- aov(KS1_WRITPOINTS ~ as.factor(SchoolID2), data=boot.data)
ks2.anova <- aov(CalcTotal_Overall2 ~ as.factor(SchoolID2), data=boot.data)
ICC.KS1.fsm[[boot]] <- ICC1(ks1.anova)
ICC.KS2.fsm[[boot]] <- ICC1(ks2.anova)
rm(ks1.anova, ks2.anova)

#Formula in SAP
#KS2past ~ KS1 + REG + GfW + (1|school)
prim.model.fsm <- lmer(CalcTotal_Overall2 ~ KS1_WRITPOINTS + Region + treat
+ (1| SchoolID2), data=boot.data)
primary.coeffs.fsm <- rbind(primary.coeffs.fsm, fixef(prim.model.fsm))

primary.n.fsm <- rbind(primary.n.fsm, c(length(prim.model.fsm@frame$treat),
sum(prim.model.fsm@frame$treat)))

varcomp <- data.frame(VarCorr(prim.model.fsm))
#collecting coefficient separately and the two variances
effect.size.fsm <- rbind(effect.size.fsm, c(
summary(prim.model.fsm)$coefficients[4,1], varcomp[1,4], varcomp[2,4]
))
rm(prim.model.fsm, boot.data, varcomp)

if (boot%%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap primary outcome in FSM
Sys.time() - t1 #about 4min

primary.n.fsm <- data.frame(primary.n.fsm)
names(primary.n.fsm) <- c("total", "intervention")
primary.n.fsm$control <- primary.n.fsm$total - primary.n.fsm$intervention
apply(primary.n.fsm, MARGIN=2, mean)
apply(primary.n.fsm, MARGIN=2, sd)

#Fist the intra-class correlations
mean(ICC.KS1.fsm)
sd(ICC.KS1.fsm)
mean(ICC.KS2.fsm)
sd(ICC.KS2.fsm)

#coefficient data
head(primary.coeffs.fsm)
apply(primary.coeffs.fsm, MARGIN=2, mean)
apply(primary.coeffs.fsm, MARGIN=2, sd)
apply(primary.coeffs.fsm, MARGIN=2, quantile, probs=c(.025, .975))
hist(primary.coeffs.fsm[,4])

head(effect.size.fsm)
effect.size.fsm <- data.frame(effect.size.fsm)
names(effect.size.fsm) <- c("treat", "var.l2", "var.l1")
#variance for effect size denominator
effect.size.fsm$variance <- effect.size.fsm$var.l2 + effect.size.fsm$var.l1
effect.size.fsm$estimate <-
effect.size.fsm$treat/sqrt(effect.size.fsm$variance)
quantile(effect.size.fsm$estimate, probs=c(.025, .975))
mean(effect.size.fsm$estimate)

apply(effect.size.fsm, MARGIN=2, mean)

```



```

#For table 7 in main report
head(table.7fsm)
table.7fsm <- data.frame(table.7fsm)
names(table.7fsm) <- c("mean.contr", "mean.inter", "var.contr", "var.inter",
"n.contr", "n.inter")

mean(table.7fsm$n.inter)
mean(table.7fsm$mean.inter)
quantile(table.7fsm$mean.inter, probs=c(.025, .975))
mean(table.7fsm$n.contr)
mean(table.7fsm$mean.contr)
quantile(table.7fsm$mean.contr, probs=c(.025, .975))
mean(table.7fsm$n.contr + table.7fsm$n.inter)
#unadjusted Hedges g
var.asterisk <- (((table.7fsm$n.contr-1) * table.7fsm$var.contr) +
((table.7fsm$n.inter-1) * table.7fsm$var.inter)) /
(table.7fsm$n.inter + table.7fsm$n.contr - 2)

g.test <- ((table.7fsm$mean.inter - table.7fsm$mean.contr) /
sqrt(var.asterisk))

mean(g.test)
quantile(g.test, probs=c(.025, .975))
rm(g.test, var.asterisk)

read.reportdat(table.7fsm)

#bootstrap of IMPUTED PRIMARY OUTCOME
#Variables and procedures according to SAP
#Gender, EverFSM, KS1
#The primary and secondary outcome variables
#n-1 dummy variables for the schools to approximate the multilevel structure
of the data as well as the described analytic approach with school-level
intercepts (no missing data, since known for every student)
#one dummy coding whether baseline data is missing (yes/ no)
#one dummy coding whether only follow-up data (yes/ no) missing

#code variable for post-data missing
miss.post2 <- data.frame(cbind(
is.na(GfWdata3.prim$KS2_WRITTAOUTCOME_Code),
is.na(GfWdata3.prim$KS2_GPSMRK), is.na(GfWdata3.prim$KS2_READMRK),
is.na(GfWdata3.prim$CalcTotal_Overall2)
))
miss.post2$miss3 <- rowSums(miss.post2)
GfWdata3.prim$miss.p <- ifelse(miss.post2$miss3>0, 1,0)
table(GfWdata3.prim$miss.p)
#N=964 miss.p

names(GfWdata3.prim)
#reduced dataset containing only these variables
GfWdata3.prim.i <- subset(GfWdata3.prim,
select=c("SchoolID2", "Gender", "EVERFSM_ALL_SPR17", "KS1_WRITPOINTS",
"CalcTotal_Overall2", "KS2_GPSMRK", "KS2_READMRK",
"KS2_WRITTAOUTCOME_Code", "miss1", "miss.p", "treat", "Region"))
names(GfWdata3.prim.i)

#define matrix with range restrictions for imputation
bound.mat.primary <- rbind(
c(4, min(GfWdata3.prim.i$KS1_WRITPOINTS, na.rm=T),
max(GfWdata3.prim.i$KS1_WRITPOINTS, na.rm=T)), #KS1

```

```

c(5,          min(GfWdata3.prim.i$CalcTotal_Overall2,          na.rm=T),
max(GfWdata3.prim.i$CalcTotal_Overall2, na.rm=T)), #KS2old
c(6,          min(GfWdata3.prim.i$KS2_GPSMRK,          na.rm=T),
max(GfWdata3.prim.i$KS2_GPSMRK, na.rm=T)), #GPS
c(7,          min(GfWdata3.prim.i$KS2_READMRK,          na.rm=T),
max(GfWdata3.prim.i$KS2_READMRK, na.rm=T)), #READ
c(8,          min(GfWdata3.prim.i$KS2_WRITTAOUTCOME_Code, na.rm=T),
max(GfWdata3.prim.i$KS2_WRITTAOUTCOME_Code, na.rm=T)) #WRIT
) #end of bound matrix
names(GfWdata3.prim.i)[c(4:8)]

```

#impute primary outcome data frame

```

t1 <- Sys.time()
library(Amelia)
#impute 1000 data sets that are used in the following
#treat and region are treated as IDvariables, i.e. not used
#imputed.data <- amelia(GfWdata3.prim.i, bounds=bound.mat.primary,
  m=1000, p2s=2, idvars=c(11,12),
  noms=c(1), ords=c(2,3,9,10),
  emburn=c(100, 250)
) #end of imputation
Sys.time() - t1 #4.7hrs

round(apply(imputed.data$imputations$imp765, MARGIN=2, FUN=min),2)
round(apply(imputed.data$imputations$imp765, MARGIN=2, FUN=max), 2)

```

#Imputed data, primary outcome in full sample

```

#define variables capturing estimates:
ICC.KS1.i <- NULL
ICC.KS2.i <- NULL
primary.coeffs.i <- NULL
effect.size.i <- NULL
capt.conv <- NULL
table.7i <- NULL
primary.n.i <- NULL

t1 <- Sys.time()
#Actual bootstrap
for (boot in 1:1000) {

if (is.data.frame(imputed.data$imputations[boot][[1]])==F) {
  capt.conv <- c(capt.conv, boot)
} #end of if-skip

if (is.data.frame(imputed.data$imputations[boot][[1]])==T) {

boot.dat <- imputed.data$imputations[boot][[1]]

boot.dat$number <- 1:nrow(boot.dat)

#generate list of student codes to select from
set.seed<-boot
id.list <- tapply(boot.dat$number, INDEX=boot.dat$SchoolID2, sample,
replace=T)
id.list <- unlist(id.list)
boot.data <- boot.dat[id.list, ]

table.7i <- rbind(table.7i,
  c(
mean(boot.data$CalcTotal_Overall2[boot.data$treat==0], na.rm=T),
mean(boot.data$CalcTotal_Overall2[boot.data$treat==1], na.rm=T),
var(boot.data$CalcTotal_Overall2[boot.data$treat==0], na.rm=T),

```

```

var(boot.data$CalcTotal_Overall2[boot.data$treat==1], na.rm=T),
sum(!is.na(boot.data$CalcTotal_Overall2[boot.data$treat==0])),
sum(!is.na(boot.data$CalcTotal_Overall2[boot.data$treat==1]))
)) #end of table collector

ks1.anova <- aov(KS1_WRITPOINTS ~ as.factor(SchoolID2), data=boot.data)
ks2.anova <- aov(CalcTotal_Overall2 ~ as.factor(SchoolID2), data=boot.data)
ICC.KS1.i[[boot]] <- ICC1(ks1.anova)
ICC.KS2.i[[boot]] <- ICC1(ks2.anova)
rm(ks1.anova, ks2.anova)

#Formula in SAP
#KS2past ~ KS1 + REG + GfW + (1|school)
prim.model <- lmer(CalcTotal_Overall2 ~ KS1_WRITPOINTS + Region + treat +
(1| SchoolID2), data=boot.data)
primary.coeffs.i <- rbind(primary.coeffs.i, fixef(prim.model))

primary.n.i <- rbind(primary.n.i, c(length(prim.model@frame$treat),
sum(prim.model@frame$treat)))

varcomp <- data.frame(VarCorr(prim.model))
#collecting coefficient separately and the two variances
effect.size.i <- rbind(effect.size.i, c(
summary(prim.model)$coefficients[4,1], varcomp[1,4], varcomp[2,4]
))

rm(prim.model, boot.dat, boot.data, varcomp)

} #end of if-compute

if (boot%%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap primary outcome
Sys.time() - t1 #7mins

primary.n.i <- data.frame(primary.n.i)
names(primary.n.i) <- c("total", "intervention")
primary.n.i$control <- primary.n.i$total - primary.n.i$intervention
apply(primary.n.i, MARGIN=2, mean)
apply(primary.n.i, MARGIN=2, sd)

#
#Fist the intra-class correlations
mean(ICC.KS1.i, na.rm=T)
sd(ICC.KS1.i, na.rm=T)
mean(ICC.KS2.i, na.rm=T)
sd(ICC.KS2.i, na.rm=T)

effect.size.read <- function(x) {
effect.tab <- x
effect.tab <- data.frame(effect.tab)
names(effect.tab) <- c("treat", "var.l2", "var.l1")
#variance for effect size denominator
effect.tab$variance <- effect.tab$var.l2 + effect.tab$var.l1
effect.tab$estimate <- effect.tab$treat/sqrt(effect.tab$variance)
cat("Quantile Effect size estimate: ", quantile(effect.tab$estimate,
probs=c(.025, .975)), "\n", "\n")
cat("Ave effect size: ", mean(effect.tab$estimate), "\n")
} #end of effect size read function

```

```

effect.size.read(effect.size.i)

apply(effect.size.i, MARGIN=2, mean)

#coefficient data
head(primary.coeffs.i)
apply(primary.coeffs.i, MARGIN=2, mean, na.rm=T)
apply(primary.coeffs.i, MARGIN=2, sd, na.rm=T)
apply(primary.coeffs.i, MARGIN=2, quantile, probs=c(.025, .975) , na.rm=T)
hist(primary.coeffs.i[,4])

read.reportdat(table.7i)

###Primary outcome, FSM-only, imputed
t1 <- Sys.time()
#define variables capturing estimates:
ICC.KS1.i.fsm <- NULL
ICC.KS2.i.fsm <- NULL
primary.coeffs.i.fsm <- NULL
effect.size.i.fsm <- NULL
capt.conv.i.fsm <- NULL
primary.n.i.fsm <- NULL
table.7ifsm <- NULL

#Actual bootstrap
for (boot in 1:1000) {

if (is.data.frame(imputed.data$imputations[boot][[1]])==F) {
  capt.conv <- c(capt.conv, boot)
} #end of if-skip

if (is.data.frame(imputed.data$imputations[boot][[1]])==T) {

boot.dat <- imputed.data$imputations[boot][[1]]
boot.dat <- subset(boot.dat, EVERFSM_ALL_SPR17==1)
#school 280 has exactly n=1 FSM students
boot.dat <- subset(boot.dat, SchoolID2!=280)

boot.dat$number <- 1:nrow(boot.dat)

#generate list of student codes to select from
set.seed<-boot
id.list <- tapply(boot.dat$number, INDEX=boot.dat$SchoolID2, sample,
replace=T)
id.list <- unlist(id.list)
boot.data <- boot.dat[id.list, ]

table.7ifsm <- rbind(table.7ifsm,
c(
mean(boot.data$CalcTotal_Overall12[boot.data$treat==0], na.rm=T),
mean(boot.data$CalcTotal_Overall12[boot.data$treat==1], na.rm=T),
var(boot.data$CalcTotal_Overall12[boot.data$treat==0], na.rm=T),
var(boot.data$CalcTotal_Overall12[boot.data$treat==1], na.rm=T),
sum(!is.na(boot.data$CalcTotal_Overall12[boot.data$treat==0])),
sum(!is.na(boot.data$CalcTotal_Overall12[boot.data$treat==1]))
)) #end of table collector

ks1.anova <- aov(KS1_WRITPOINTS ~ as.factor(SchoolID2), data=boot.data)
ks2.anova <- aov(CalcTotal_Overall12 ~ as.factor(SchoolID2), data=boot.data)
ICC.KS1.i.fsm[[boot]] <- ICC1(ks1.anova)
ICC.KS2.i.fsm[[boot]] <- ICC1(ks2.anova)
rm(ks1.anova, ks2.anova)

```

```

#Formula in SAP
#KS2past ~ KS1 + REG + GfW + (1|school)
prim.model.i.fsm <- lmer(CalcTotal_Overall12 ~ KS1_WRITPOINTS + Region + treat
+ (1| SchoolID2), data=boot.data)
primary.coeffs.i.fsm <- rbind(primary.coeffs.i.fsm,
fixef(prim.model.i.fsm))

primary.n.i.fsm <- rbind(primary.n.i.fsm,
c(length(prim.model.i.fsm@frame$treat), sum(prim.model.i.fsm@frame$treat)))

varcomp <- data.frame(VarCorr(prim.model.i.fsm))
#collecting coefficient separately and the two variances
effect.size.i.fsm <- rbind(effect.size.i.fsm, c(
summary(prim.model.i.fsm)$coefficients[4,1], varcomp[1,4], varcomp[2,4]
))

rm(prim.model.i.fsm, boot.dat, boot.data, varcomp)

} #end of if-compute

if (boot%%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap primary outcome
Sys.time() - t1 #~4min

primary.n.i.fsm <- data.frame(primary.n.i.fsm)
names(primary.n.i.fsm) <- c("total", "intervention")
primary.n.i.fsm$control <- primary.n.i.fsm$total
primary.n.i.fsm$intervention
apply(primary.n.i.fsm, MARGIN=2, mean)
apply(primary.n.i.fsm, MARGIN=2, sd)

#
#Fist the intra-class correlations
mean(ICC.KS1.i.fsm, na.rm=T)
sd(ICC.KS1.i.fsm, na.rm=T)
mean(ICC.KS2.i.fsm, na.rm=T)
sd(ICC.KS2.i.fsm, na.rm=T)

#coefficient data
head(primary.coeffs.i.fsm)
apply(primary.coeffs.i.fsm, MARGIN=2, mean, na.rm=T)
apply(primary.coeffs.i.fsm, MARGIN=2, sd, na.rm=T)
apply(primary.coeffs.i.fsm, MARGIN=2, quantile, probs=c(.025, .975) ,
na.rm=T)
hist(primary.coeffs.i.fsm[,4])
effect.size.read(effect.size.i.fsm)

apply(effect.size.i.fsm, MARGIN=2, mean)

read.reportdat(table.7ifsm)

SECONDARY OUTCOMES
#check whether variables have been produced correctly:
table(GfWdata3.sec$BlindTreatment2, GfWdata3.sec$treat)
round(mean(GfWdata3.sec$KS1_WRITPOINTS , na.rm=T), 3)
sd(GfWdata3.sec$KS1_WRITPOINTS , na.rm=T)

```

```

#check availability of outcome
names(GfWdata3.sec)
table(GfWdata3.sec$KS2_WRITTAOUTCOME_Code)

#####
#bootstrap of SECONDARY OUTCOME 1 // WRIT
#not imputed
#define variables capturing estimates:
ICC.writ <- NULL
coeffs.writ <- NULL
effect.size.writ <- NULL
table.7writ <- NULL
secondary.n.writ <- NULL

#define consecutive numbers for bootstrap selection within schools
GfWdata3.sec$number <- 1:nrow(GfWdata3.sec)

t1 <- Sys.time()
#Actual bootstrap
for (boot in 1:1000) {
#generate list of student codes to select from
set.seed<-boot
id.list <- tapply(GfWdata3.sec$number, INDEX=GfWdata3.sec$SchoolID2, sample,
replace=T)
id.list <- unlist(id.list)
boot.data <- GfWdata3.sec[id.list, ]

#average in each group, n in each group with KS2past
table.7writ <- rbind(table.7writ, c(
mean(boot.data$KS2_WRITTAOUTCOME_Code[boot.data$treat==0], na.rm=T),
mean(boot.data$KS2_WRITTAOUTCOME_Code[boot.data$treat==1], na.rm=T),
var(boot.data$KS2_WRITTAOUTCOME_Code[boot.data$treat==0], na.rm=T),
var(boot.data$KS2_WRITTAOUTCOME_Code[boot.data$treat==1], na.rm=T),
sum(!is.na(boot.data$KS2_WRITTAOUTCOME_Code[boot.data$treat==0])),
sum(!is.na(boot.data$KS2_WRITTAOUTCOME_Code[boot.data$treat==1]))
)) #end of table collector

writ.anova <- aov(KS2_WRITTAOUTCOME_Code ~ as.factor(SchoolID2),
data=boot.data)
ICC.writ[[boot]] <- ICC1(writ.anova)
rm(writ.anova)

#Formula in SAP
#KS2past ~ KS1 + REG + GfW + (1|school)
writ.model <- lmer(KS2_WRITTAOUTCOME_Code ~ KS1_WRITPOINTS + Region + treat
+ (1| SchoolID2), data=boot.data)
coeffs.writ <- rbind(coeffs.writ, fixef(writ.model))

secondary.n.writ <- rbind(secondary.n.writ,
c(length(writ.model@frame$treat), sum(writ.model@frame$treat)))

varcomp <- data.frame(VarCorr(writ.model))
#collecting coefficient separately and the two variances
effect.size.writ <- rbind(effect.size.writ, c(
summary(writ.model)$coefficients[4,1], varcomp[1,4], varcomp[2,4]
))

rm(writ.model, boot.data, varcomp)

if (boot%%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
}

```

```

} #end of plot if-clause

} # end of bootstrap primary outcome

Sys.time() - t1 #6 minutes

#n for runs
secondary.n.writ <- data.frame(secondary.n.writ)
names(secondary.n.writ) <- c("total", "intervention")
secondary.n.writ$control <- secondary.n.writ$total
secondary.n.writ$intervention
apply(secondary.n.writ, MARGIN=2, mean)
apply(secondary.n.writ, MARGIN=2, sd)

#
#Fist the intra-class correlations
mean(ICC.writ)
sd(ICC.writ)

#coefficient data
head(coeffs.writ)
apply(coeffs.writ, MARGIN=2, mean)
apply(coeffs.writ, MARGIN=2, sd)
apply(coeffs.writ, MARGIN=2, quantile, probs=c(.025, .975))
hist(coeffs.writ[,4])

effect.size.read(effect.size.writ)

apply(effect.size.writ, MARGIN=2, mean)

read.reportdat(table.7writ)

#####
#bootstrap of SECONDARY OUTCOME 2 // GPS
#not imputed
#check variable
table(GfWdata3.sec$KS2_GPSMRK)

#define variables capturing estimates:
ICC.gps <- NULL
coeffs.gps <- NULL
effect.size.gps <- NULL
table.7gps <- NULL
secondary.n.gps <- NULL

#define consecutive numbers for bootstrap selection within schools
GfWdata3.sec$number <- 1:nrow(GfWdata3.sec)

t1 <- Sys.time()
#Actual bootstrap
for (boot in 1:1000) {
#generate list of student codes to select from
set.seed<-boot
id.list <- tapply(GfWdata3.sec$number, INDEX=GfWdata3.sec$SchoolID2, sample,
replace=T)
id.list <- unlist(id.list)
boot.data <- GfWdata3.sec[id.list, ]

table.7gps <- rbind(table.7gps,
c(
mean(boot.data$KS2_GPSMRK[boot.data$treat==0], na.rm=T),
mean(boot.data$KS2_GPSMRK[boot.data$treat==1], na.rm=T),

```

```

var(boot.data$KS2_GPSMRK[boot.data$treat==0], na.rm=T),
var(boot.data$KS2_GPSMRK[boot.data$treat==1], na.rm=T),
sum(!is.na(boot.data$KS2_GPSMRK[boot.data$treat==0])),
sum(!is.na(boot.data$KS2_GPSMRK[boot.data$treat==1]))
)) #end of table collector

gps.anova <- aov(KS2_GPSMRK ~ as.factor(SchoolID2), data=boot.data)
ICC.gps[[boot]] <- ICC1(gps.anova)
rm(gps.anova)

#Formula in SAP
#KS2past ~ KS1 + REG + GfW + (1|school)
gps.model <- lmer(KS2_GPSMRK ~ KS1_WRITPOINTS + Region + treat + (1|
SchoolID2), data=boot.data)
coeffs.gps <- rbind(coeffs.gps, fixef(gps.model))

secondary.n.gps <- rbind(secondary.n.gps, c(length(gps.model@frame$treat),
sum(gps.model@frame$treat)))

varcomp <- data.frame(VarCorr(gps.model))
#collecting coefficient separately and the two variances
effect.size.gps <- rbind(effect.size.gps, c(
summary(gps.model)$coefficients[4,1], varcomp[1,4], varcomp[2,4]
))

rm(gps.model, boot.data, varcomp)

if (boot%%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap primary outcome
Sys.time() - t1 #6min

secondary.n.gps <- data.frame(secondary.n.gps)
names(secondary.n.gps) <- c("total", "intervention")
secondary.n.gps$control <- secondary.n.gps$total -
secondary.n.gps$intervention
apply(secondary.n.gps, MARGIN=2, mean)
apply(secondary.n.gps, MARGIN=2, sd)

#
#Fist the intra-class correlations
mean(ICC.gps)
sd(ICC.gps)

#coefficient data
head(coeffs.gps)
apply(coeffs.gps, MARGIN=2, mean)
apply(coeffs.gps, MARGIN=2, sd)
apply(coeffs.gps, MARGIN=2, quantile, probs=c(.025, .975))
hist(coeffs.gps[,4])
sum(is.na(coeffs.gps[,4]))

#hlm effect size
effect.size.read(effect.size.gps)

#variance components
apply(effect.size.gps, MARGIN=2, mean)

#raw statistics and effect sizes

```



```

read.reportdat(table.7gps)

#####
#bootstrap of SECONDARY OUTCOME 3 // READ
#not imputed
#check variable
table(GfWdata3.sec$KS2_READMRK)

t1 <- Sys.time()
#define variables capturing estimates:
ICC.read <- NULL
coeffs.read <- NULL
effect.size.2read <- NULL
table.7read <- NULL
secondary.n.read <- NULL

#define consecutive numbers for bootstrap selection within schools
GfWdata3.sec$number <- 1:nrow(GfWdata3.sec)

#Actual bootstrap
for (boot in 1:1000) {
#generate list of student codes to select from
set.seed<-boot
id.list <- tapply(GfWdata3.sec$number, INDEX=GfWdata3.sec$SchoolID2, sample,
replace=T)
id.list <- unlist(id.list)
boot.data <- GfWdata3.sec[id.list, ]

table.7read <- rbind(table.7read,
mean(boot.data$KS2_READMRK[boot.data$treat==0], na.rm=T),
mean(boot.data$KS2_READMRK[boot.data$treat==1], na.rm=T),
var(boot.data$KS2_READMRK[boot.data$treat==0], na.rm=T),
var(boot.data$KS2_READMRK[boot.data$treat==1], na.rm=T),
sum(!is.na(boot.data$KS2_READMRK[boot.data$treat==0])),
sum(!is.na(boot.data$KS2_READMRK[boot.data$treat==1]))
)) #end of table collector

read.anova <- aov(KS2_READMRK ~ as.factor(SchoolID2), data=boot.data)
ICC.read[[boot]] <- ICC1(read.anova)
rm(read.anova)

#Formula in SAP
#KS2past ~ KS1 + REG + GfW + (1|school)
read.model <- lmer(KS2_READMRK ~ KS1_WRITPOINTS + Region + treat + (1|
SchoolID2), data=boot.data)
coeffs.read <- rbind(coeffs.read, fixef(read.model))

secondary.n.read <- rbind(secondary.n.read,
c(length(read.model@frame$treat), sum(read.model@frame$treat)))

varcomp <- data.frame(VarCorr(read.model))
#collecting coefficient separately and the two variances
effect.size.2read <- rbind(effect.size.2read, c(
summary(read.model)$coefficients[4,1], varcomp[1,4], varcomp[2,4]
))

rm(read.model, boot.data, varcomp)

if (boot%%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

```

```

} # end of bootstrap primary outcome
Sys.time() - t1 #6min

#
#Fist the intra-class correlations
mean(ICC.read)
sd(ICC.read)

secondary.n.read <- data.frame(secondary.n.read)
names(secondary.n.read) <- c("total", "intervention")
secondary.n.read$control <- secondary.n.read$total
secondary.n.read$intervention
apply(secondary.n.read, MARGIN=2, mean)
apply(secondary.n.read, MARGIN=2, sd)

#coefficient data
head(coeffs.read)
apply(coeffs.read, MARGIN=2, mean)
apply(coeffs.read, MARGIN=2, sd)
apply(coeffs.read, MARGIN=2, quantile, probs=c(.025, .975))
hist(coeffs.read[,4])

#remember: effect.size.read is already a function!
#hlm effect size
effect.size.read(effect.size.2read)

#variance components
apply(effect.size.2read, MARGIN=2, mean)

#raw data
read.reportdat(table.7read)

#Imputation of SECONDARY OUTCOME

#Variables and procedures according to SAP
#Gender, EverFSM, KS1
#The primary and secondary outcome variables
#n-1 dummy variables for the schools to approximate the multilevel structure
of the data as well as the described analytic approach with school-level
intercepts (no missing data, since known for every student)
#one dummy coding whether baseline data is missing (yes/ no)
#one dummy coding whether only follow-up data (yes/ no) missing

#code variable for post-data missing
miss.post2 <- data.frame(cbind(
is.na(GfWdata3.sec$KS2_WRITTAOUTCOME_Code), is.na(GfWdata3.sec$KS2_GPMSRK),
is.na(GfWdata3.sec$KS2_READMRK), is.na(GfWdata3.sec$CalcTotal_Overall2)
))
miss.post2$miss3 <- rowSums(miss.post2)
GfWdata3.sec$miss.p <- ifelse(miss.post2$miss3>0, 1,0)
table(GfWdata3.sec$miss.p)
nrow(GfWdata3.sec)
#N=1858 with at least one missing post

names(GfWdata3.sec)
#reduced dataset containing only these variables
#here without KS2old tasks since too rarely filled in
GfWdata3.sec.i <- subset(GfWdata3.sec,

```

```

select=c("SchoolID2", "Gender", "EVERFSM_ALL_SPR17", "KS1_WRITPOINTS",
"KS2_GPSMRK", "KS2_READMRK", "KS2_WRITTAOUTCOME_Code", "miss1", "miss.p",
"treat", "Region"))
names(GfWdata3.sec.i)
nrow(GfWdata3.sec.i)

#define matrix with range restrictions for imputation
bound.mat.secondary <- rbind(
c(4, min(GfWdata3.sec.i$KS1_WRITPOINTS, na.rm=T),
max(GfWdata3.sec.i$KS1_WRITPOINTS, na.rm=T)), #KS1
c(5, min(GfWdata3.sec.i$KS2_GPSMRK, na.rm=T),
max(GfWdata3.sec.i$KS2_GPSMRK, na.rm=T)), #GPS
c(6, min(GfWdata3.sec.i$KS2_READMRK, na.rm=T),
max(GfWdata3.sec.i$KS2_READMRK, na.rm=T)), #READ
c(7, min(GfWdata3.sec.i$KS2_WRITTAOUTCOME_Code, na.rm=T),
max(GfWdata3.sec.i$KS2_WRITTAOUTCOME_Code, na.rm=T)) #WRIT
) #end of bound matrix
names(GfWdata3.sec.i)[c(4:7)]

#imputation of secondary outcome
t1 <- Sys.time()
library(Amelia)
#impute 1000 data sets that are used in the following
#treat and region are treated as IDvariables, i.e. not used
imputed.data.secondary <- amelia(GfWdata3.sec.i,
bounds=bound.mat.secondary,
m=1000, p2s=2, idvars=c(10,11),
noms=c(1), ords=c(2,3,8,9),
emburn=c(100, 250)
) #end of imputation

Sys.time() - t1 #5.5 hours

round(apply(imputed.data.secondary$imputations$imp765,
FUN=min), 2), MARGIN=2,
round(apply(imputed.data.secondary$imputations$imp765, MARGIN=2, FUN=max),
2)

#check imputed data:
head(imputed.data.secondary$imputations$imp1)

#####
#bootstrap of SECONDARY OUTCOME 1 // WRIT // IMPUTED
#define variables capturing estimates:
ICC.writ.i <- NULL
coeffs.writ.i <- NULL
effect.size.writ.i <- NULL
table.7writ.i <- NULL
secondary.n.writ.i <- NULL

t1 <- Sys.time()
#Actual bootstrap
for (boot in 1:1000) {
boot.dat <- imputed.data.secondary$imputations[boot][[1]]
boot.dat$number <- 1:nrow(boot.dat)

#generate list of student codes to select from
set.seed<-boot
id.list <- tapply(boot.dat$number, INDEX=boot.dat$SchoolID2, sample,
replace=T)

```

```

id.list <- unlist(id.list)
boot.data <- boot.dat[id.list, ]

#average in each group, n in each group with KS2past
table.7writ.i <- rbind(table.7writ, c(
mean(boot.data$KS2_WRITTAOUTCOME_Code[boot.data$treat==0], na.rm=T),
mean(boot.data$KS2_WRITTAOUTCOME_Code[boot.data$treat==1], na.rm=T),
var(boot.data$KS2_WRITTAOUTCOME_Code[boot.data$treat==0], na.rm=T),
var(boot.data$KS2_WRITTAOUTCOME_Code[boot.data$treat==1], na.rm=T),
sum(!is.na(boot.data$KS2_WRITTAOUTCOME_Code[boot.data$treat==0])),
sum(!is.na(boot.data$KS2_WRITTAOUTCOME_Code[boot.data$treat==1]))
)) #end of table collector

writ.i.anova <- aov(KS2_WRITTAOUTCOME_Code ~ as.factor(SchoolID2),
data=boot.data)
ICC.writ.i[[boot]] <- ICC1(writ.i.anova)
rm(writ.i.anova)

#Formula in SAP
#KS2past ~ KS1 + REG + GfW + (1|school)
writ.i.model <- lmer(KS2_WRITTAOUTCOME_Code ~ KS1_WRITPOINTS + Region + treat
+ (1| SchoolID2), data=boot.data)
coeffs.writ.i <- rbind(coeffs.writ.i, fixef(writ.i.model))

secondary.n.writ.i <- rbind(secondary.n.writ.i,
c(length(writ.i.model@frame$treat), sum(writ.i.model@frame$treat)))

varcomp <- data.frame(VarCorr(writ.i.model))
#collecting coefficient separately and the two variances
effect.size.writ.i <- rbind(effect.size.writ.i, c(
summary(writ.i.model)$coefficients[4,1], varcomp[1,4], varcomp[2,4]
))

rm(writ.i.model, boot.dat, boot.data, varcomp)

if (boot%%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap primary outcome

Sys.time() - t1 #10 minutes

#n for runs
secondary.n.writ.i <- data.frame(secondary.n.writ.i)
names(secondary.n.writ.i) <- c("total", "intervention")
secondary.n.writ.i$control <- secondary.n.writ.i$total -
secondary.n.writ.i$intervention
apply(secondary.n.writ.i, MARGIN=2, mean)
apply(secondary.n.writ.i, MARGIN=2, sd)

#
#Fist the intra-class correlations
mean(ICC.writ.i)
sd(ICC.writ.i)

#coefficient data
head(coeffs.writ.i)
apply(coeffs.writ.i, MARGIN=2, mean)
apply(coeffs.writ.i, MARGIN=2, sd)
apply(coeffs.writ.i, MARGIN=2, quantile, probs=c(.025, .975))

```

```

hist(coeffs.writ.i[,4])

#hlm effect size
effect.size.read(effect.size.writ.i)

#variance components
apply(effect.size.writ.i, MARGIN=2, mean)

#raw data
read.reportdat(table.7writ.i)

#####
#bootstrap of SECONDARY OUTCOME 2 // GPS // IMPUTED

#check variable
table(imputed.data.secondary$imputations$impl$KS2_GPSMRK)
table(GfWdata3.sec$KS2_GPSMRK)

#define variables capturing estimates:
ICC.gps.i <- NULL
coeffs.gps.i <- NULL
effect.size.gps.i <- NULL
table.7gps.i <- NULL
secondary.n.gps.i <- NULL

t1 <- Sys.time()
#Actual bootstrap
for (boot in 1:1000) {

boot.dat <- imputed.data.secondary$imputations[boot][[1]]
boot.dat$number <- 1:nrow(boot.dat)
#generate list of student codes to select from
set.seed<-boot
id.list <- tapply(boot.dat$number, INDEX=boot.dat$SchoolID2, sample,
replace=T)
id.list <- unlist(id.list)
boot.data <- boot.dat[id.list, ]
rm(boot.dat)

table.7gps.i <- rbind(table.7gps.i, c(
mean(boot.data$KS2_GPSMRK[boot.data$treat==0], na.rm=T),
mean(boot.data$KS2_GPSMRK[boot.data$treat==1], na.rm=T),
var(boot.data$KS2_GPSMRK[boot.data$treat==0], na.rm=T),
var(boot.data$KS2_GPSMRK[boot.data$treat==1], na.rm=T),
sum(!is.na(boot.data$KS2_GPSMRK[boot.data$treat==0])),
sum(!is.na(boot.data$KS2_GPSMRK[boot.data$treat==1]))
)) #end of table collector

gps.i.anova <- aov(KS2_GPSMRK ~ as.factor(SchoolID2), data=boot.data)
ICC.gps.i[[boot]] <- ICC1(gps.i.anova)
rm(gps.i.anova)

#Formula in SAP
#KS2past ~ KS1 + REG + GfW + (1|school)
gps.i.model <- lmer(KS2_GPSMRK ~ KS1_WRITPOINTS + Region + treat + (1|
SchoolID2), data=boot.data)
coeffs.gps.i <- rbind(coeffs.gps.i, fixef(gps.i.model))

secondary.n.gps.i <- rbind(secondary.n.gps.i,
c(length(gps.i.model@frame$treat), sum(gps.i.model@frame$treat)))

```

```

varcomp <- data.frame(VarCorr(gps.i.model))
#collecting coefficient separately and the two variances
effect.size.gps.i <- rbind(effect.size.gps.i, c(
summary(gps.i.model)$coefficients[4,1], varcomp[1,4], varcomp[2,4]
))

rm(gps.i.model, boot.data, varcomp)

if (boot%%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap primary outcome
Sys.time() - t1 #8min

secondary.n.gps.i <- data.frame(secondary.n.gps.i)
names(secondary.n.gps.i) <- c("total", "intervention")
secondary.n.gps.i$control <- secondary.n.gps.i$total
secondary.n.gps.i$intervention
apply(secondary.n.gps.i, MARGIN=2, mean)
apply(secondary.n.gps.i, MARGIN=2, sd)

#
#Fist the intra-class correlations
mean(ICC.gps.i)
sd(ICC.gps.i)

#coefficient data
head(coeffs.gps.i)
apply(coeffs.gps.i, MARGIN=2, mean)
apply(coeffs.gps.i, MARGIN=2, sd)
apply(coeffs.gps.i, MARGIN=2, quantile, probs=c(.025, .975))
hist(coeffs.gps.i[,4])
sum(is.na(coeffs.gps.i[,4]))

effect.size.read(effect.size.gps.i)

apply(effect.size.gps.i, MARGIN=2, mean)

read.reportdat(table.7gps.i)

#####
#bootstrap of SECONDARY OUTCOME 3 // READ // IMPUTED

t1 <- Sys.time()
#define variables capturing estimates:
ICC.read.i <- NULL
coeffs.read.i <- NULL
effect.size.2read.i <- NULL
table.7read.i <- NULL
secondary.n.read.i <- NULL

#define consecutive numbers for bootstrap selection within schools
GfWdata3.sec$number <- 1:nrow(GfWdata3.sec)

#Actual bootstrap
for (boot in 1:1000) {

boot.dat <- imputed.data.secondary$imputations[boot][[1]]
boot.dat$number <- 1:nrow(boot.dat)
#generate list of student codes to select from

```

```

set.seed<-boot
id.list <- tapply(boot.dat$number, INDEX=boot.dat$SchoolID2, sample,
replace=T)
id.list <- unlist(id.list)
boot.data <- boot.dat[id.list, ]
rm(boot.dat)

table.7read.i <- rbind(table.7read.i, c(
mean(boot.data$KS2_READMRK[boot.data$treat==0], na.rm=T),
mean(boot.data$KS2_READMRK[boot.data$treat==1], na.rm=T),
var(boot.data$KS2_READMRK[boot.data$treat==0], na.rm=T),
var(boot.data$KS2_READMRK[boot.data$treat==1], na.rm=T),
sum(!is.na(boot.data$KS2_READMRK[boot.data$treat==0])),
sum(!is.na(boot.data$KS2_READMRK[boot.data$treat==1]))
)) #end of table collector

read.i.anova <- aov(KS2_READMRK ~ as.factor(SchoolID2), data=boot.data)
ICC.read.i[[boot]] <- ICC1(read.i.anova)
rm(read.i.anova)

#Formula in SAP
#KS2past ~ KS1 + REG + GfW + (1|school)
read.i.model <- lmer(KS2_READMRK ~ KS1_WRITPOINTS + Region + treat + (1|
SchoolID2), data=boot.data)
coeffs.read.i <- rbind(coeffs.read.i, fixef(read.i.model))

secondary.n.read.i <- rbind(secondary.n.read.i,
c(length(read.i.model@frame$treat), sum(read.i.model@frame$treat)))

varcomp <- data.frame(VarCorr(read.i.model))
#collecting coefficient separately and the two variances
effect.size.2read.i <- rbind(effect.size.2read.i, c(
summary(read.i.model)$coefficients[4,1], varcomp[1,4], varcomp[2,4]
))

rm(read.i.model, boot.data, varcomp)

if (boot%%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap primary outcome
Sys.time() - t1 #8min

#
#Fist the intra-class correlations
mean(ICC.read.i)
sd(ICC.read.i)

secondary.n.read.i <- data.frame(secondary.n.read.i)
names(secondary.n.read.i) <- c("total", "intervention")
secondary.n.read.i$control <- secondary.n.read.i$total -
secondary.n.read.i$intervention
apply(secondary.n.read.i, MARGIN=2, mean)
apply(secondary.n.read.i, MARGIN=2, sd)

#coefficient data
head(coeffs.read.i)
apply(coeffs.read.i, MARGIN=2, mean)
apply(coeffs.read.i, MARGIN=2, sd)
apply(coeffs.read.i, MARGIN=2, quantile, probs=c(.025, .975))

```

```

hist(coeffs.read.i[,4])

#remember: effect.size.read is already a function, therefor ethe change
effect.size.read(effect.size.2read.i)

apply(effect.size.2read.i, MARGIN=2, mean)

read.reportdat(table.7read.i)

#####
###GRAMMAR QUIZ
#####

#here only the HLM analysis is presented
#the teacher-level analysis is presented in a separate code file

#Load Teacher scores at t2
load("L:\\Jan\\001_Analysis\\Include_in_report_MAY2018\\2018_05_23_Teachers
cores_t2.RData")

#merge the scores by Teacher ID
names(Grammar.post)
names(GfWdata3.prim)
nrow(GfWdata3.prim)
names(Grammar.post)[1] <- "TeacherId_1"
names(Grammar.post)

#merging all over ID1
GfWdata3.prim.gq <- merge(x= GfWdata3.prim , y=Grammar.post,
by="TeacherId_1", all.x=T)
nrow(GfWdata3.prim.gq)

#add additional scores via ID2
names(Grammar.post) <- c("TeacherId_2", "score2.id2")
names(Grammar.post)
GfWdata3.prim.gq <- merge(x= GfWdata3.prim.gq, y=Grammar.post,
by="TeacherId_2", all.x=T)
nrow(GfWdata3.prim.gq)

#add additional scores via ID3
names(Grammar.post) <- c("TeacherId_3", "score2.id3")
names(Grammar.post)
GfWdata3.prim.gq <- merge(x= GfWdata3.prim.gq, y=Grammar.post,
by="TeacherId_3", all.x=T)
nrow(GfWdata3.prim.gq)

#calculate average score for the analysis
names(GfWdata3.prim.gq)
GfWdata3.prim.gq$avegscore <- rowMeans(GfWdata3.prim.gq[, 30:32], na.rm=T)

#check calculation
#GfWdata3.prim.gq[, 30:33]
GfWdata3.prim.gq[1:1000, 30:33]

#data set only with non-missing scores to determine frequencies
GfWdata3.prim.gq <- GfWdata3.prim.gq[!is.na(GfWdata3.prim.gq$avegscore),]
nrow(GfWdata3.prim.gq)
table(GfWdata3.prim.gq$treat)

#schools:
length(unique(GfWdata3.prim.gq$SchoolID))

```



```

#data frame with SchoolID and treat to identify group numbers
school.gq <- unique(GfWdata3.prim.gq[, c(4,27)])
table(school.gq$treat)

#Bootstrap of mediation parameter
#bootstrap of PRIMARY OUTCOME
#define variables capturing estimates:
GQ.coeffs <- NULL
table.7gq <- NULL
GQ.n.primary <- NULL

t1 <- Sys.time()
#define consecutive numbers for bootstrap selection within schools
GfWdata3.prim.gq$number <- 1:nrow(GfWdata3.prim.gq)

#Actual bootstrap
for (boot in 1:1000) {
#generate list of student codes to select from
set.seed<-boot
id.list <- tapply(GfWdata3.prim.gq$number,
INDEX=GfWdata3.prim.gq$SchoolID2, sample, replace=T)
id.list <- unlist(id.list)
boot.data <- GfWdata3.prim.gq[id.list, ]

#average in each group, n in each group with KS2past
table.7gq <- rbind(table.7gq, c(
mean(boot.data$CalcTotal_Overall2[boot.data$treat==0], na.rm=T),
mean(boot.data$CalcTotal_Overall2[boot.data$treat==1], na.rm=T),
var(boot.data$CalcTotal_Overall2[boot.data$treat==0], na.rm=T),
var(boot.data$CalcTotal_Overall2[boot.data$treat==1], na.rm=T),
sum(!is.na(boot.data$CalcTotal_Overall2[boot.data$treat==0])),
sum(!is.na(boot.data$CalcTotal_Overall2[boot.data$treat==1]))
)) #end of table collector

#Formula in SAP
#KS2past ~ KS1 + REG + GQ + GfW + (1+GQ|school)
#see main report: GQ not estimated as a random effect due to lack of
convergence
prim.model <- lmer(CalcTotal_Overall2 ~ KS1_WRITPOINTS + Region + treat +
avegscore + (1 | SchoolID2), data=boot.data)
GQ.coeffs <- rbind(GQ.coeffs, fixef(prim.model))

GQ.n.primary <- rbind(GQ.n.primary, c(length(prim.model@frame$treat),
sum(prim.model@frame$treat)))

rm(prim.model, boot.data)

if (boot%%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap primary outcome
Sys.time() - t1 #2mins

#Data for table
GQ.n.primary <- data.frame(GQ.n.primary)
names(GQ.n.primary) <- c("total", "intervention")
GQ.n.primary$control <- GQ.n.primary$total - GQ.n.primary$intervention
apply(GQ.n.primary, MARGIN=2, mean)
apply(GQ.n.primary, MARGIN=2, sd)

```

```

#coefficient data
head(GQ.coeffs)
apply(GQ.coeffs, MARGIN=2, mean)
apply(GQ.coeffs, MARGIN=2, sd)
apply(GQ.coeffs, MARGIN=2, quantile, probs=c(.025, .975))
hist(GQ.coeffs[,4])

#multiply coefficients for confidence interval
colnames(GQ.coeffs)
mediation.gq <- GQ.coeffs[,4] *GQ.coeffs[,5]
mean(mediation.gq)
sd(mediation.gq)
quantile(mediation.gq, probs=c(.025, .975))
hist(mediation.gq)

#
head(table.7gq)
table.7gq <- data.frame(table.7gq)
names(table.7gq) <- c("mean.contr", "mean.inter", "var.contr", "var.inter",
"n.contr", "n.inter")
head(table.7gq)

#####Grammar Quiz Analyses
#Here the analyses regarding the Grammar Quiz are presented
#They were performed in a different workspace

#First read_in of the teacher patterns from the main data file
#Loading relevant packages:
library(Rcmdr)

#loading core data set
GfWdata <-
  readXL("L:/Jan/000_ORIGINAL/Analysis Dataset - Pupil v4 20180423 with
scaled scores.xlsx",
  rownames=FALSE, header=TRUE, na="", sheet="Blindeddataset",
  stringsAsFactors=TRUE)

#read out of unique combinations of
#these teacher codes are needed to merge the grammar quiz scores into the
main file for analysis
teacher.codes <- data.frame(unique(GfWdata[, c(7,8,9)]))
fix(teacher.codes)

#remove main file, not further needed for this analysis
rm(GfWdata)

#####
#read in grammar quiz data
GrammarQuiz <-
  readXL("L:/Jan/000_ORIGINAL/Final Grammar Quiz dataset v0.7 20180419 for
Jan.xlsx",
  rownames=FALSE, header=TRUE, na="", sheet="Coded data",
  stringsAsFactors=TRUE)

#check content of data
#check item names across the different versions
fix(GrammarQuiz)
names(GrammarQuiz)

names(GrammarQuiz)[7:36]
names(GrammarQuiz)[37:65]

```

```

GrammarQuiz[GrammarQuiz==99] <- NA
fix(GrammarQuiz)

#table with item-level descriptive data
psych::describe(GrammarQuiz[,7:65])

#count missings for the two quizzes
miss.quiz <- data.frame(is.na(GrammarQuiz[7:65]))
fix(miss.quiz)
names(miss.quiz)
GrammarQuiz$mis1 <- rowSums(miss.quiz[, 1:30])
GrammarQuiz$mis2 <- rowSums(miss.quiz[, 31:59])
table(GrammarQuiz$mis1)
table(GrammarQuiz$mis2)
table(GrammarQuiz$mis1, GrammarQuiz$mis2)

#Distribution across groups
table(GrammarQuiz$mis1, GrammarQuiz$BlindTreatment2)
table(GrammarQuiz$mis2, GrammarQuiz$BlindTreatment2)
#This pre-scored variable provides the same information
table(GrammarQuiz$mis1, GrammarQuiz$CompletedPreQuiz)
table(GrammarQuiz$mis2, GrammarQuiz$CompletedPostQuiz)

#Since only full or empty responses are available per assessment, simple
score calculation works:
GrammarQuiz$score1 <- rowSums(GrammarQuiz[, 7:36])
GrammarQuiz$score2 <- rowSums(GrammarQuiz[, 37:65])
sum(!is.na(GrammarQuiz$score1))
sum(!is.na(GrammarQuiz$score2))

#####
#Descriptive data on score level as well as ranks

#T1 intervention
mean(GrammarQuiz$score1[GrammarQuiz$BlindTreatment2==6], na.rm=T)
median(GrammarQuiz$score1[GrammarQuiz$BlindTreatment2==6], na.rm=T)
sd(GrammarQuiz$score1[GrammarQuiz$BlindTreatment2==6], na.rm=T)
sum(!is.na(GrammarQuiz$score1[GrammarQuiz$BlindTreatment2==6]))
#T2 intervention
mean(GrammarQuiz$score2[GrammarQuiz$BlindTreatment2==6], na.rm=T)
median(GrammarQuiz$score2[GrammarQuiz$BlindTreatment2==6], na.rm=T)
sd(GrammarQuiz$score2[GrammarQuiz$BlindTreatment2==6], na.rm=T)
sum(!is.na(GrammarQuiz$score2[GrammarQuiz$BlindTreatment2==6]))

#T1 control
mean(GrammarQuiz$score1[GrammarQuiz$BlindTreatment2==5], na.rm=T)
median(GrammarQuiz$score1[GrammarQuiz$BlindTreatment2==5], na.rm=T)
sd(GrammarQuiz$score1[GrammarQuiz$BlindTreatment2==5], na.rm=T)
sum(!is.na(GrammarQuiz$score1[GrammarQuiz$BlindTreatment2==5]))
#T2 control
mean(GrammarQuiz$score2[GrammarQuiz$BlindTreatment2==5], na.rm=T)
median(GrammarQuiz$score2[GrammarQuiz$BlindTreatment2==5], na.rm=T)
sd(GrammarQuiz$score2[GrammarQuiz$BlindTreatment2==5], na.rm=T)
sum(!is.na(GrammarQuiz$score2[GrammarQuiz$BlindTreatment2==5]))

#determine the rank order
#highest rank = highest score!
GrammarQuiz$rank1 <- rank(GrammarQuiz$score1, na.last="keep",
ties.method="average")
#check result
plot(x= GrammarQuiz$rank1, GrammarQuiz$score1)

```

```

GrammarQuiz$rank2      <-      rank(GrammarQuiz$score2,      na.last="keep",
ties.method="average")

#T1 control
mean(GrammarQuiz$rank1[GrammarQuiz$BlindTreatment2==5], na.rm=T)
sd(GrammarQuiz$rank1[GrammarQuiz$BlindTreatment2==5], na.rm=T)
sum(!is.na(GrammarQuiz$rank1[GrammarQuiz$BlindTreatment2==5]))
#T2 control
mean(GrammarQuiz$rank2[GrammarQuiz$BlindTreatment2==5], na.rm=T)
sd(GrammarQuiz$rank2[GrammarQuiz$BlindTreatment2==5], na.rm=T)
sum(!is.na(GrammarQuiz$rank2[GrammarQuiz$BlindTreatment2==5]))

#T1 intervention
mean(GrammarQuiz$rank1[GrammarQuiz$BlindTreatment2==6], na.rm=T)
sd(GrammarQuiz$rank1[GrammarQuiz$BlindTreatment2==6], na.rm=T)
sum(!is.na(GrammarQuiz$rank1[GrammarQuiz$BlindTreatment2==6]))
#T2 intervention
mean(GrammarQuiz$rank2[GrammarQuiz$BlindTreatment2==6], na.rm=T)
sd(GrammarQuiz$rank2[GrammarQuiz$BlindTreatment2==6], na.rm=T)
sum(!is.na(GrammarQuiz$rank2[GrammarQuiz$BlindTreatment2==6]))

#####
#Difference of ranks T2-T1 (i.e. the better rank at second assessment
#the bigger positive change in rank)
#Determine on data set that has only pre- AND post values!
GrammarQuiz.prepost <- na.omit(data.frame(GrammarQuiz$BlindTreatment2,
GrammarQuiz$score1, GrammarQuiz$score2))
fix(GrammarQuiz.prepost)
names(GrammarQuiz.prepost)

GrammarQuiz.prepost$rank1.d <- rank(GrammarQuiz.prepost$GrammarQuiz.score1,
na.last="keep", ties.method="average")
plot(x=          GrammarQuiz.prepost$rank1.d,          y=
GrammarQuiz.prepost$GrammarQuiz.score1)
GrammarQuiz.prepost$rank2.d <- rank(GrammarQuiz.prepost$GrammarQuiz.score2,
na.last="keep", ties.method="average")
GrammarQuiz.prepost$rankdiff.d <- (GrammarQuiz.prepost$rank2.d -
GrammarQuiz.prepost$rank1.d)
hist(GrammarQuiz.prepost$rankdiff.d)
sum(!is.na(GrammarQuiz.prepost$rankdiff.d))

#T1 control
mean(GrammarQuiz.prepost$rankdiff.d[GrammarQuiz.prepost$GrammarQuiz.BlindTr
eatment2==5], na.rm=T)
sd(GrammarQuiz.prepost$rankdiff.d[GrammarQuiz.prepost$GrammarQuiz.BlindTrea
tment2==5], na.rm=T)
sum(!is.na(GrammarQuiz.prepost$rankdiff.d[GrammarQuiz.prepost$GrammarQuiz.B
lindTreatment2==5]))
#T1 intervention
mean(GrammarQuiz.prepost$rankdiff.d[GrammarQuiz.prepost$GrammarQuiz.BlindTr
eatment2==6], na.rm=T)
sd(GrammarQuiz.prepost$rankdiff.d[GrammarQuiz.prepost$GrammarQuiz.BlindTrea
tment2==6], na.rm=T)
sum(!is.na(GrammarQuiz.prepost$rankdiff.d[GrammarQuiz.prepost$GrammarQuiz.B
lindTreatment2==6]))
####

#Create an object that contains only post-scores and teacher IDs for export
#used for analysis of mediation hypothesis
Grammar.post <- GrammarQuiz[GrammarQuiz$mis2==0, ]
Grammar.post <- subset(Grammar.post, select=c("PersonId", "score2"))

```

```

nrow(Grammar.post)
names(Grammar.post)
#save(file="L:\\Jan\\001_Analysis\\Include_in_report_MAY2018\\2018_05_23_Te
acherScores_t2.RData", list="Grammar.post")
#This object is the basis for the Grammar Quiz HLM analysis presented a
couple of pages before

####Reliabilty analysis
#Cronbach
psych::alpha(GrammarQuiz[,7:36])
psych::alpha(GrammarQuiz[,37:65])

#control grp only
contr.cor <- na.omit(GrammarQuiz[GrammarQuiz$BlindTreatment2==5, ])
nrow(contr.cor)
cor.test(contr.cor$score1, contr.cor$score2)
rm(contr.cor)

#Regression for differential effects
pre.post.data <- na.omit(GrammarQuiz[ , c(4,68,69)])
nrow(pre.post.data)
names(pre.post.data)

difeffect.reg <- lm(score2~score1*as.factor(BlindTreatment2),
data=pre.post.data)
summary(difeffect.reg)
confint(difeffect.reg)

#bootstrapped T-test
#post data only
post.data <- na.omit(GrammarQuiz[ , c(4,69)])
names(post.data)
nrow(post.data)
post.data$treat <- ifelse(post.data$BlindTreatment2==5, 0, 1)

stat.cap <- NULL
for (i in 1:1000) {
select <- sample(1: nrow(post.data), replace=T)
t.res <- t.test(score2 ~ BlindTreatment2, alternative = c("two.sided"),
paired=F, data=post.data[select, ])
val <- c(t.res$estimate)
stat.cap <- rbind(stat.cap, c(t.res$statistic, t.res$p.value, val))
rm(t.res, select, val)
} #end of bootstrap
mean(stat.cap[,1])
quantile(stat.cap[,1], probs=c(.025, .975))
mean(stat.cap[,2])
quantile(stat.cap[,2], probs=c(.025, .975))
ave.boot <- (stat.cap[,4] - stat.cap[,3])
mean(ave.boot)
quantile(ave.boot, probs=c(.025, .975))

#Parallel analyses based on FIML correlation matrix
library(psych)
parallel.all <- fa.parallel(corFiml(GrammarQuiz[,7:65]), fm="minres",
fa="both", sim=T, n.iter=500, quant=.95, use="pairwise", n.obs=312)
parallel.pre <- fa.parallel(corFiml(GrammarQuiz[,7:36]), fm="minres",
fa="both", sim=T, n.iter=500, quant=.95, use="pairwise", n.obs=297)
parallel.post <- fa.parallel(corFiml(GrammarQuiz[,37:65]), fm="minres",
fa="both", sim=T, n.iter=500, quant=.95, use="pairwise", n.obs=222)

```

```

#estimating and inspecting key reference solutions
compl.all <- principal(corFiml(GrammarQuiz[,7:65]), n.obs=312, n.factors=1)
compl.pre <- principal(corFiml(GrammarQuiz[,7:36]), n.obs=297, n.factors=1)
compl.post <- principal(corFiml(GrammarQuiz[,37:65]), n.obs=222, n.factors=1)
comp4.post <- principal(corFiml(GrammarQuiz[,37:65]), n.obs=222, n.factors=4)

#for exchange with the EEF saved as:
#save.image("L:\\Jan\\001_Analysis\\Include_in_report\\2018_05_23_GrammarQuiz.RData")

#####
#Added on 7th July to investigate
#the correlation between the Grammar Quiz and CPD:
#Addition in July 2018, investigating correlation between CPD and Quiz scores

#Read in of 3-day CPD data
Fidelity <- readXL("L:/Jan/000_ORIGINAL/Implementation Fidelity v1 20180420 for Jan.xlsx", rownames=FALSE, header=TRUE, na="", sheet="Implementation Fidelity", stringsAsFactors=TRUE)

names(Fidelity)
Fidelity$Imp.Fid..
mean(Fidelity$Imp.Fid..)
sd(Fidelity$Imp.Fid..)
length(Fidelity$Imp.Fid..)

fidelity.merge <- subset(Fidelity, select=c("SchoolId2", "Imp.Fid.."))
names(fidelity.merge) <- c("SchoolID2", "Fidelity")

#merge with main data set
Fidelity.GQ1 <- merge(GrammarQuiz, fidelity.merge, by="SchoolID2", all.x=T)
summary(Fidelity.GQ1$Fidelity)
Fidelity.GQ1$Fidelity[is.na(Fidelity.GQ1$Fidelity)] <- 0
summary(Fidelity.GQ1$Fidelity)

head(Fidelity.GQ1)

#Read 4-day CPD data
Fidelity4CPD <- readXL("L:/Jan/000_ORIGINAL/Implementation Fidelity v3 20180420_4CPD.xlsx", rownames=FALSE, header=TRUE, na="", sheet="Implementation Fidelity", stringsAsFactors=TRUE)

names(Fidelity4CPD)
Fidelity4CPD$Imp.Fid...4
mean(Fidelity4CPD$Imp.Fid...4)
sd(Fidelity4CPD$Imp.Fid...4)
length(Fidelity4CPD$Imp.Fid...4)

fidelity4CPD.merge <- subset(Fidelity4CPD, select=c("SchoolId2", "Imp.Fid...4"))
names(fidelity4CPD.merge) <- c("SchoolID2", "Fidelity4CPD")

fidelity4CPD.merge

#merge with main data set
Fidelity.GQ2 <- merge(Fidelity.GQ1, fidelity4CPD.merge, by="SchoolID2", all.x=T)
summary(Fidelity.GQ2$Fidelity4CPD)

```

```

Fidelity.GQ2$Fidelity4CPD[is.na(Fidelity.GQ2$Fidelity4CPD)] <- 0
summary(Fidelity.GQ2$Fidelity4CPD)

#Determine Correlations

#Pre- and post-score with CPD 3
cor.test(Fidelity.GQ2$score1, Fidelity.GQ2$Fidelity)
cor.test(Fidelity.GQ2$score2, Fidelity.GQ2$Fidelity)

#Pre- and post-score with CPD 4
cor.test(Fidelity.GQ2$score1, Fidelity.GQ2$Fidelity4CPD)
cor.test(Fidelity.GQ2$score2, Fidelity.GQ2$Fidelity4CPD)

#Bootstrapping the subgroup tests
#The following is basically a repeat of the code for the primary outcome
#Here only the three specified sub-populations are evaluated
#in observed and imputed data

#library(lme4)

#bootstrap of PRIMARY OUTCOME - EVERFSM, UNIMPUTED
#define variables capturing estimates:
primary.coeffs.s fsm <- NULL
primary.n.s fsm <- NULL

t1 <- Sys.time()
#define consecutive numbers for bootstrap selection within schools
GfWdata3.prim$number <- 1:nrow(GfWdata3.prim)

#Actual bootstrap
for (boot in 1:1000) {
#generate list of student codes to select from
set.seed<-boot
id.list <- tapply(GfWdata3.prim$number, INDEX=GfWdata3.prim$SchoolID2,
sample, replace=T)
id.list <- unlist(id.list)
boot.data <- GfWdata3.prim[id.list, ]

#Formula in SAP
#KS2past ~ KS1 + Subgroup + REG + GfW + Subgroup:GfW + (1|school)
prim.model <- lmer(CalcTotal_Overall12 ~ KS1_WRITPOINTS + EVERFSM_ALL_SPR17
+ Region + treat + EVERFSM_ALL_SPR17:treat + (1+ EVERFSM_ALL_SPR17|
SchoolID2), data=boot.data)

primary.coeffs.s fsm <- rbind(primary.coeffs.s fsm, fixef(prim.model))

primary.n.s fsm <- rbind(primary.n.s fsm, c(length(prim.model@frame$treat),
sum(prim.model@frame$treat)))

rm(prim.model, boot.data)

if (boot%%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap primary outcome
Sys.time() - t1 #7.2mins

#n for runs

```

```

primary.n.s fsm <- data.frame(primary.n.s fsm)
names(primary.n.s fsm) <- c("total", "intervention")
primary.n.s fsm$control <- primary.n.s fsm$total - primary.n.s fsm$intervention
apply(primary.n.s fsm, MARGIN=2, mean)
apply(primary.n.s fsm, MARGIN=2, sd)

#coefficient data
head(primary.coeffs.s fsm)
apply(primary.coeffs.s fsm, MARGIN=2, mean)
apply(primary.coeffs.s fsm, MARGIN=2, sd)
apply(primary.coeffs.s fsm, MARGIN=2, quantile, probs=c(.025, .975))
hist(primary.coeffs.s fsm[,4])

#bootstrap of PRIMARY OUTCOME - GENDER, UNIMPUTED
#define variables capturing estimates:
primary.coeffs.gen <- NULL
effect.size.gen <- NULL
table.7.gen <- NULL
primary.n.gen <- NULL

t1 <- Sys.time()
#define consecutive numbers for bootstrap selection within schools
GfWdata3.prim$number <- 1:nrow(GfWdata3.prim)

#Actual bootstrap
for (boot in 1:1000) {
#generate list of student codes to select from
set.seed<-boot
id.list <- tapply(GfWdata3.prim$number, INDEX=GfWdata3.prim$SchoolID2,
sample, replace=T)
id.list <- unlist(id.list)
boot.data <- GfWdata3.prim[id.list, ]

#Formula in SAP
#KS2past ~ KS1 + Subgroup + REG + GfW + Subgroup:GfW + (1|school)
prim.model <- lmer(CalcTotal_Overall12 ~ KS1_WRITPOINTS + Gender + Region +
treat + Gender:treat + (1+Gender| SchoolID2), data=boot.data)

primary.coeffs.gen <- rbind(primary.coeffs.gen, fixef(prim.model))

primary.n.gen <- rbind(primary.n.gen, c(length(prim.model@frame$treat),
sum(prim.model@frame$treat)))

rm(prim.model, boot.data)

if (boot%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap primary outcome
Sys.time() - t1 #7.2mins

#n for runs
primary.n.gen <- data.frame(primary.n.gen)
names(primary.n.gen) <- c("total", "intervention")
primary.n.gen$control <- primary.n.gen$total - primary.n.gen$intervention
apply(primary.n.gen, MARGIN=2, mean)
apply(primary.n.gen, MARGIN=2, sd)

#coefficient data
head(primary.coeffs.gen)

```



```

apply(primary.coeffs.gen, MARGIN=2, mean)
apply(primary.coeffs.gen, MARGIN=2, sd)
apply(primary.coeffs.gen, MARGIN=2, quantile, probs=c(.025, .975))
hist(primary.coeffs.gen[,4])

#bootstrap of PRIMARY OUTCOME - PRETEST, UNIMPUTED

#Define new variable that contains the dichotomised KS1:
GfWdata3.prim$kslmed <-
ifelse(GfWdata3.prim$KS1_WRITPOINTS>=median(GfWdata3.prim$KS1_WRITPOINTS,
na.rm=T), 1, 0)
table(GfWdata3.prim$kslmed)

#save median for later reference:
median.KS1 <- median(GfWdata3.prim$KS1_WRITPOINTS, na.rm=T)

#define variables capturing estimates:
primary.coeffs.ks1 <- NULL
primary.n.ks1 <- NULL

t1 <- Sys.time()
#define consecutive numbers for bootstrap selection within schools
GfWdata3.prim$number <- 1:nrow(GfWdata3.prim)

#Actual bootstrap
for (boot in 1:1000) {
#generate list of student codes to select from
set.seed<-boot
id.list <- tapply(GfWdata3.prim$number, INDEX=GfWdata3.prim$SchoolID2,
sample, replace=T)
id.list <- unlist(id.list)
boot.data <- GfWdata3.prim[id.list, ]

#Formula in SAP
#KS2past ~ KS1 + Subgroup + REG + GfW + Subgroup:GfW + (1|school)
prim.model <- lmer(CalcTotal_Overall12 ~ kslmed + Region + treat +
kslmed:treat + (1+ kslmed| SchoolID2), data=boot.data)

primary.coeffs.ks1 <- rbind(primary.coeffs.ks1, fixef(prim.model))

primary.n.ks1 <- rbind(primary.n.ks1, c(length(prim.model@frame$treat),
sum(prim.model@frame$treat)))

rm(prim.model, boot.data)

if (boot%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap primary outcome
Sys.time() - t1 #7.2mins

#n for runs
primary.n.ks1 <- data.frame(primary.n.ks1)
names(primary.n.ks1) <- c("total", "intervention")
primary.n.ks1$control <- primary.n.ks1$total - primary.n.ks1$intervention
apply(primary.n.ks1, MARGIN=2, mean)
apply(primary.n.ks1, MARGIN=2, sd)

#coefficient data
head(primary.coeffs.ks1)

```

```

apply(primary.coeffs.ks1, MARGIN=2, mean)
apply(primary.coeffs.ks1, MARGIN=2, sd)
apply(primary.coeffs.ks1, MARGIN=2, quantile, probs=c(.025, .975))
hist(primary.coeffs.ks1[,4])

####
#analyses with imputed data follow for the three subgroup tests
####

#bootstrap of PRIMARY OUTCOME - EVERFSM, IMPUTED
#define variables capturing estimates:
primary.coeffs.s fsm.i <- NULL
primary.n.s fsm.i <- NULL

t1 <- Sys.time()
#Actual bootstrap
for (boot in 1:1000) {
#generate list of student codes to select from

boot.dat <- imputed.data$imputations[boot][[1]]
boot.dat$number <- 1:nrow(boot.dat)

set.seed<-boot
id.list <- tapply(boot.dat$number, INDEX=boot.dat$SchoolID2, sample,
replace=T)
id.list <- unlist(id.list)
boot.data <- boot.dat[id.list, ]
rm(boot.dat)

#Formula in SAP
#KS2past ~ KS1 + Subgroup + REG + GfW + Subgroup:GfW + (1|school)
prim.model <- lmer(CalcTotal_Overall12 ~ KS1_WRITPOINTS + EVERFSM_ALL_SPR17
+ Region + treat + EVERFSM_ALL_SPR17:treat + (1+ EVERFSM_ALL_SPR17|
SchoolID2), data=boot.data)

primary.coeffs.s fsm.i <- rbind(primary.coeffs.s fsm.i, fixef(prim.model))

primary.n.s fsm.i <- rbind(primary.n.s fsm.i,
c(length(prim.model@frame$treat), sum(prim.model@frame$treat)))

rm(prim.model, boot.data)

if (boot%%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap primary outcome
Sys.time() - t1 #7.2mins

#n for runs
primary.n.s fsm.i <- data.frame(primary.n.s fsm.i)
names(primary.n.s fsm.i) <- c("total", "intervention")
primary.n.s fsm.i$control <- primary.n.s fsm.i$total -
primary.n.s fsm.i$intervention
apply(primary.n.s fsm.i, MARGIN=2, mean)
apply(primary.n.s fsm.i, MARGIN=2, sd)

#coefficient data
head(primary.coeffs.s fsm.i)
apply(primary.coeffs.s fsm.i, MARGIN=2, mean)
apply(primary.coeffs.s fsm.i, MARGIN=2, sd)

```

```

apply(primary.coeffs.s fsm.i, MARGIN=2, quantile, probs=c(.025, .975))
hist(primary.coeffs.s fsm.i[,4])

#bootstrap of PRIMARY OUTCOME - GENDER, IMPUTED
#define variables capturing estimates:
primary.coeffs.gen.i <- NULL
primary.n.gen.i <- NULL

t1 <- Sys.time()
#Actual bootstrap
for (boot in 1:1000) {
#generate list of student codes to select from

boot.dat <- imputed.data$imputations[boot][[1]]
boot.dat$number <- 1:nrow(boot.dat)

set.seed<-boot
id.list <- tapply(boot.dat$number, INDEX=boot.dat$SchoolID2, sample,
replace=T)
id.list <- unlist(id.list)
boot.data <- boot.dat[id.list, ]

#Formula in SAP
#KS2past ~ KS1 + Subgroup + REG + GfW + Subgroup:GfW + (1|school)
prim.model <- lmer(CalcTotal_Overall2 ~ KS1_WRITPOINTS + Gender + Region +
treat + Gender:treat + (1+Gender| SchoolID2), data=boot.data)

primary.coeffs.gen.i <- rbind(primary.coeffs.gen.i, fixef(prim.model))

primary.n.gen.i <- rbind(primary.n.gen.i, c(length(prim.model@frame$treat),
sum(prim.model@frame$treat)))

rm(prim.model, boot.data)

if (boot%%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap primary outcome
Sys.time() - t1 #7.2mins

#n for runs
primary.n.gen.i <- data.frame(primary.n.gen.i)
names(primary.n.gen.i) <- c("total", "intervention")
primary.n.gen.i$control <- primary.n.gen.i$total
primary.n.gen.i$intervention
apply(primary.n.gen.i, MARGIN=2, mean)
apply(primary.n.gen.i, MARGIN=2, sd)

#coefficient data
head(primary.coeffs.gen.i)
apply(primary.coeffs.gen.i, MARGIN=2, mean)
apply(primary.coeffs.gen.i, MARGIN=2, sd)
apply(primary.coeffs.gen.i, MARGIN=2, quantile, probs=c(.025, .975))
hist(primary.coeffs.gen.i[,4])

#bootstrap of PRIMARY OUTCOME - PRETEST, IMPUTED

#from above:
#save median for later reference:
#median.KS1 <- median(GfWdata3.prim$KS1_WRITPOINTS, na.rm=T)

```

```

#define variables capturing estimates:
primary.coeffs.ks1.i <- NULL
primary.n.ks1.i <- NULL

t1 <- Sys.time()
#Actual bootstrap
for (boot in 1:1000) {

boot.dat <- imputed.data$imputations[boot][[1]]
boot.dat$number <- 1:nrow(boot.dat)

#generate list of student codes to select from
set.seed<-boot
id.list <- tapply(boot.dat$number, INDEX=boot.dat$SchoolID2, sample,
replace=T)
id.list <- unlist(id.list)
boot.data <- boot.dat[id.list, ]
rm(boot.dat)

#Define new variable that contains the dichotomised KS1:
boot.data$kslmed <- ifelse(boot.data$KS1_WRITPOINTS>=median.KS1, 1, 0)

#Formula in SAP
#KS2past ~ KS1 + Subgroup + REG + GfW + Subgroup:GfW + (1|school)
prim.model <- lmer(CalcTotal_Overall2 ~ kslmed + Region + treat +
kslmed:treat + (1+ kslmed| SchoolID2), data=boot.data)

primary.coeffs.ks1.i <- rbind(primary.coeffs.ks1.i, fixef(prim.model))

primary.n.ks1.i <- rbind(primary.n.ks1.i, c(length(prim.model@frame$treat),
sum(prim.model@frame$treat)))

rm(prim.model, boot.data)

if (boot%%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap primary outcome
Sys.time() - t1 #7.2mins

#n for runs
primary.n.ks1.i <- data.frame(primary.n.ks1.i)
names(primary.n.ks1.i) <- c("total", "intervention")
primary.n.ks1.i$control <- primary.n.ks1.i$total
primary.n.ks1.i$intervention
apply(primary.n.ks1.i, MARGIN=2, mean)
apply(primary.n.ks1.i, MARGIN=2, sd)

#coefficient data
head(primary.coeffs.ks1.i)
apply(primary.coeffs.ks1.i, MARGIN=2, mean)
apply(primary.coeffs.ks1.i, MARGIN=2, sd)
apply(primary.coeffs.ks1.i, MARGIN=2, quantile, probs=c(.025, .975))
hist(primary.coeffs.ks1.i[,4])

#####
###Results for KS1 in detail

#bootstrap of PRIMARY OUTCOME

```

#NONIMPUTED, TOP50 KS1 pretest

```

#check variable
table(GfWdata3.prim$ks1med)

#define variables capturing estimates:
primary.coeffs.topks1 <- NULL
effect.size.topks1 <- NULL
table.7.topks1 <- NULL
primary.n.topks1 <- NULL

t1 <- Sys.time()
#Actual bootstrap
for (boot in 1:1000) {

boot.dat <- subset(GfWdata3.prim, ks1med==1)
#define consecutive numbers for bootstrap selection within schools
boot.dat$number <- 1:nrow(boot.dat)

#generate list of student codes to select from
set.seed<-boot
id.list <- tapply(boot.dat$number, INDEX=boot.dat$SchoolID2, sample,
replace=T)
id.list <- unlist(id.list)
boot.data <- boot.dat[id.list, ]
rm(boot.dat)

#average in each group, n in each group with KS2past
table.7.topks1 <- rbind(table.7.topks1, c(
mean(boot.data$CalcTotal_Overall2[boot.data$treat==0], na.rm=T),
mean(boot.data$CalcTotal_Overall2[boot.data$treat==1], na.rm=T),
var(boot.data$CalcTotal_Overall2[boot.data$treat==0], na.rm=T),
var(boot.data$CalcTotal_Overall2[boot.data$treat==1], na.rm=T),
sum(!is.na(boot.data$CalcTotal_Overall2[boot.data$treat==0])),
sum(!is.na(boot.data$CalcTotal_Overall2[boot.data$treat==1]))
)) #end of table collector

#not controlled for KS1
#since this would be doubly controlling for it
prim.model <- lmer(CalcTotal_Overall2 ~ Region + treat + (1| SchoolID2),
data=boot.data)
primary.coeffs.topks1 <- rbind(primary.coeffs.topks1, fixef(prim.model))

primary.n.topks1 <- rbind(primary.n.topks1,
c(length(prim.model@frame$treat), sum(prim.model@frame$treat)))

varcomp <- data.frame(VarCorr(prim.model))
#collecting coefficient separately and the two variances
effect.size.topks1 <- rbind(effect.size.topks1, c(
summary(prim.model)$coefficients[3,1], varcomp[1,4], varcomp[2,4]
))

rm(prim.model, boot.data, varcomp)

if (boot%%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap primary outcome
Sys.time() - t1 #7.2mins

```

```

#n for runs
primary.n.topks1 <- data.frame(primary.n.topks1)
names(primary.n.topks1) <- c("total", "intervention")
primary.n.topks1$control <- primary.n.topks1$total -
primary.n.topks1$intervention
apply(primary.n.topks1, MARGIN=2, mean)
apply(primary.n.topks1, MARGIN=2, sd)

#coefficient data
head(primary.coeffs.topks1)
apply(primary.coeffs.topks1, MARGIN=2, mean)
apply(primary.coeffs.topks1, MARGIN=2, sd)
apply(primary.coeffs.topks1, MARGIN=2, quantile, probs=c(.025, .975))
hist(primary.coeffs.topks1 [,3])

effect.size.read(effect.size.topks1)
read.reportdat(table.7.topks1)

#bootstrap of PRIMARY OUTCOME
#NONIMPUTED, Lower50 KS1 pretest

#check variable
table(GfWdata3.prim$ks1med)

#define variables capturing estimates:
primary.coeffs.loks1 <- NULL
effect.size.loks1 <- NULL
table.7.loks1 <- NULL
primary.n.loks1 <- NULL

t1 <- Sys.time()
#Actual bootstrap
for (boot in 1:1000) {

boot.dat <- subset(GfWdata3.prim, ks1med==0)
#define consecutive numbers for bootstrap selection within schools
boot.dat$number <- 1:nrow(boot.dat)

#generate list of student codes to select from
set.seed<-boot
id.list <- tapply(boot.dat$number, INDEX=boot.dat$SchoolID2, sample,
replace=T)
id.list <- unlist(id.list)
boot.data <- boot.dat[id.list, ]
rm(boot.dat)

#average in each group, n in each group with KS2past
table.7.loks1 <- rbind(table.7.loks1,
mean(boot.data$CalcTotal_Overall12[boot.data$treat==0], na.rm=T),
mean(boot.data$CalcTotal_Overall12[boot.data$treat==1], na.rm=T),
var(boot.data$CalcTotal_Overall12[boot.data$treat==0], na.rm=T),
var(boot.data$CalcTotal_Overall12[boot.data$treat==1], na.rm=T),
sum(!is.na(boot.data$CalcTotal_Overall12[boot.data$treat==0])),
sum(!is.na(boot.data$CalcTotal_Overall12[boot.data$treat==1]))
)) #end of table collector

#not controlled for KS1
#since this would be doubly controlling for it
prim.model <- lmer(CalcTotal_Overall12 ~ Region + treat + (1| SchoolID2),
data=boot.data)
primary.coeffs.loks1 <- rbind(primary.coeffs.loks1, fixef(prim.model))

```

```

primary.n.loks1 <- rbind(primary.n.loks1, c(length(prim.model@frame$treat),
sum(prim.model@frame$treat)))

varcomp <- data.frame(VarCorr(prim.model))
#collecting coefficient separately and the two variances
effect.size.loks1 <- rbind(effect.size.loks1, c(
summary(prim.model)$coefficients[3,1], varcomp[1,4], varcomp[2,4]
))

rm(prim.model, boot.data, varcomp)

if (boot%%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap primary outcome
Sys.time() - t1 #7.2mins

#n for runs
primary.n.loks1 <- data.frame(primary.n.loks1)
names(primary.n.loks1) <- c("total", "intervention")
primary.n.loks1$control <- primary.n.loks1$total -
primary.n.loks1$intervention
apply(primary.n.loks1, MARGIN=2, mean)
apply(primary.n.loks1, MARGIN=2, sd)

#coefficient data
head(primary.coeffs.loks1)
apply(primary.coeffs.loks1, MARGIN=2, mean)
apply(primary.coeffs.loks1, MARGIN=2, sd)
apply(primary.coeffs.loks1, MARGIN=2, quantile, probs=c(.025, .975))
hist(primary.coeffs.loks1 [,3])

effect.size.read(effect.size.loks1)
read.reportdat(table.7.loks1)

#bootstrap of PRIMARY OUTCOME
#IMPUTED, TOP50 KS1 pretest

#define variables capturing estimates:
primary.coeffs.topks1.i <- NULL
effect.size.topks1.i <- NULL
table.7.topks1.i <- NULL
primary.n.topks1.i <- NULL

t1 <- Sys.time()
#Actual bootstrap
for (boot in 1:1000) {

boot.dat <- imputed.data$imputations[boot][[1]]
#Define new variable that contains the dichotomised KS1:
boot.dat$ks1med <- ifelse(boot.dat$KS1_WRITPOINTS>=median.KS1, 1, 0)
boot.dat <- subset(boot.dat, ks1med==1)
boot.dat$number <- 1:nrow(boot.dat)

#generate list of student codes to select from
set.seed<-boot
id.list <- tapply(boot.dat$number, INDEX=boot.dat$SchoolID2, sample,
replace=T)
id.list <- unlist(id.list)

```

```

boot.data <- boot.dat[id.list, ]
rm(boot.dat)

#average in each group, n in each group with KS2past
table.7.topks1.i <- rbind(table.7.topks1.i, c(
mean(boot.data$CalcTotal_Overall2[boot.data$treat==0], na.rm=T),
mean(boot.data$CalcTotal_Overall2[boot.data$treat==1], na.rm=T),
var(boot.data$CalcTotal_Overall2[boot.data$treat==0], na.rm=T),
var(boot.data$CalcTotal_Overall2[boot.data$treat==1], na.rm=T),
sum(!is.na(boot.data$CalcTotal_Overall2[boot.data$treat==0])),
sum(!is.na(boot.data$CalcTotal_Overall2[boot.data$treat==1]))
)) #end of table collector

#not controlled for KS1
#since this would be doubly controlling for it
prim.model <- lmer(CalcTotal_Overall2 ~ Region + treat + (1| SchoolID2),
data=boot.data)
primary.coeffs.topks1.i <- rbind(primary.coeffs.topks1.i,
fixef(prim.model))

primary.n.topks1.i <- rbind(primary.n.topks1.i,
c(length(prim.model@frame$treat), sum(prim.model@frame$treat)))

varcomp <- data.frame(VarCorr(prim.model))
#collecting coefficient separately and the two variances
effect.size.topks1.i <- rbind(effect.size.topks1.i, c(
summary(prim.model)$coefficients[3,1], varcomp[1,4], varcomp[2,4]
))

rm(prim.model, boot.data, varcomp)

if (boot%%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap primary outcome
Sys.time() - t1 #7.2mins

#n for runs
primary.n.topks1.i <- data.frame(primary.n.topks1.i)
names(primary.n.topks1.i) <- c("total", "intervention")
primary.n.topks1.i$control <- primary.n.topks1.i$total -
primary.n.topks1.i$intervention
apply(primary.n.topks1.i, MARGIN=2, mean)
apply(primary.n.topks1.i, MARGIN=2, sd)

#coefficient data
head(primary.coeffs.topks1.i)
apply(primary.coeffs.topks1.i, MARGIN=2, mean)
apply(primary.coeffs.topks1.i, MARGIN=2, sd)
apply(primary.coeffs.topks1.i, MARGIN=2, quantile, probs=c(.025, .975))
hist(primary.coeffs.topks1 [,3])

effect.size.read(effect.size.topks1.i)
read.reportdat(table.7.topks1.i)

#bootstrap of PRIMARY OUTCOME
#IMPUTED, LOWER50 KS1 pretest

#define variables capturing estimates:
primary.coeffs.loks1.i <- NULL

```



```

effect.size.loks1.i <- NULL
table.7.loks1.i <- NULL
primary.n.loks1.i <- NULL

t1 <- Sys.time()
#Actual bootstrap
for (boot in 1:1000) {

boot.dat <- imputed.data$imputations[boot][[1]]
#Define new variable that contains the dichotomised KS1:
boot.dat$ks1med <- ifelse(boot.dat$KS1_WRITPOINTS>=median.KS1, 1, 0)
boot.dat <- subset(boot.dat, ks1med==0)
boot.dat$number <- 1:nrow(boot.dat)

#generate list of student codes to select from
set.seed<-boot
id.list <- tapply(boot.dat$number, INDEX=boot.dat$SchoolID2, sample,
replace=T)
id.list <- unlist(id.list)
boot.data <- boot.dat[id.list, ]
rm(boot.dat)

#average in each group, n in each group with KS2past
table.7.loks1.i <- rbind(table.7.loks1.i, c(
mean(boot.data$CalcTotal_Overall2[boot.data$treat==0], na.rm=T),
mean(boot.data$CalcTotal_Overall2[boot.data$treat==1], na.rm=T),
var(boot.data$CalcTotal_Overall2[boot.data$treat==0], na.rm=T),
var(boot.data$CalcTotal_Overall2[boot.data$treat==1], na.rm=T),
sum(!is.na(boot.data$CalcTotal_Overall2[boot.data$treat==0])),
sum(!is.na(boot.data$CalcTotal_Overall2[boot.data$treat==1]))
)) #end of table collector

#not controlled for KS1
#since this would be doubly controlling for it
prim.model <- lmer(CalcTotal_Overall2 ~ Region + treat + (1| SchoolID2),
data=boot.data)
primary.coeffs.loks1.i <- rbind(primary.coeffs.loks1.i, fixef(prim.model))

primary.n.loks1.i <- rbind(primary.n.loks1.i,
c(length(prim.model@frame$treat), sum(prim.model@frame$treat)))

varcomp <- data.frame(VarCorr(prim.model))
#collecting coefficient separately and the two variances
effect.size.loks1.i <- rbind(effect.size.loks1.i, c(
summary(prim.model)$coefficients[3,1], varcomp[1,4], varcomp[2,4]
))

rm(prim.model, boot.data, varcomp)

if (boot%%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap primary outcome
Sys.time() - t1 #7.2mins

#n for runs
primary.n.loks1.i <- data.frame(primary.n.loks1.i)
names(primary.n.loks1.i) <- c("total", "intervention")
primary.n.loks1.i$control <- primary.n.loks1.i$total -
primary.n.loks1.i$intervention

```

```

apply(primary.n.loks1.i, MARGIN=2, mean)
apply(primary.n.loks1.i, MARGIN=2, sd)

#coefficient data
head(primary.coeffs.loks1.i)
apply(primary.coeffs.loks1.i, MARGIN=2, mean)
apply(primary.coeffs.loks1.i, MARGIN=2, sd)
apply(primary.coeffs.loks1.i, MARGIN=2, quantile, probs=c(.025, .975))
hist(primary.coeffs.loks1.i [,3])

effect.size.read(effect.size.loks1.i)
read.reportdat(table.7.loks1.i)

#Descriptive results for table 10

#TABLE 10
#Higher performance
#intervention
psych::describe(subset(GfWdata3.prim, ((BlindTreatment2==6) &
(ks1med==1)))$CalcTotal_Overall2)
mean.ci(subset(GfWdata3.prim, ((BlindTreatment2==6) &
(ks1med==1)))$CalcTotal_Overall2)

table(GfWdata3.prim$BlindTreatment2, GfWdata3.prim$ks1med)

#control
psych::describe(subset(GfWdata3.prim, ((BlindTreatment2==5) &
(ks1med==1)))$CalcTotal_Overall2)
mean.ci(subset(GfWdata3.prim, ((BlindTreatment2==5) &
(ks1med==1)))$CalcTotal_Overall2)

#TABLE 10
#Lower performance
#intervention
psych::describe(subset(GfWdata3.prim, ((BlindTreatment2==6) &
(ks1med==0)))$CalcTotal_Overall2)
mean.ci(subset(GfWdata3.prim, ((BlindTreatment2==6) &
(ks1med==0)))$CalcTotal_Overall2)

table(GfWdata3.prim$BlindTreatment2, GfWdata3.prim$ks1med)

#control
psych::describe(subset(GfWdata3.prim, ((BlindTreatment2==5) &
(ks1med==0)))$CalcTotal_Overall2)
mean.ci(subset(GfWdata3.prim, ((BlindTreatment2==5) &
(ks1med==0)))$CalcTotal_Overall2)

#These are the codes and objects for the analysis that uses the fidelity
assessment as a stand-in for the actual treatment variable.
#Language corrected in July 2018: this pertains to COMPLIANCE

#This is again a repeat in form and structure of the analyses
#The main change is that instead of "treat", "Fidelity" is used
#As well as that the fidelity ratings are brought in from outside

#Read in fidelity data set, which was provided as an Excel file:
#Implementation Fidelity v1 20180420 for Jan.xlsx
Fidelity <- readXL("L:/Jan/000_ORIGINAL/Implementation Fidelity v1 20180420
for Jan.xlsx", rownames=FALSE, header=TRUE, na="", sheet="Implementation
Fidelity", stringsAsFactors=TRUE)

```

```

#check data
names(Fidelity)
Fidelity$Imp.Fid..
mean(Fidelity$Imp.Fid..)
sd(Fidelity$Imp.Fid..)
length(Fidelity$Imp.Fid..)

#use only SchoolIDs and Fidelity scores
fidelity.merge <- subset(Fidelity, select=c("SchoolID2", "Imp.Fid.."))
names(fidelity.merge) <- c("SchoolID2", "Fidelity")
fidelity.merge

#merge with main data set
#and include empty rows in the full data set
#only the intervention schools are coded in "fidelity.merge"!
Fidelity.prim <- merge(GfWdata3.prim, fidelity.merge, by="SchoolID2",
all.x=T)
summary(Fidelity.prim$Fidelity)
#assign zero to control/non-compliant schools
Fidelity.prim$Fidelity[is.na(Fidelity.prim$Fidelity)] <- 0
summary(Fidelity.prim$Fidelity)

#merge with secondary data set
#and same recoding as before
Fidelity.sec <- merge(GfWdata3.sec, fidelity.merge, by="SchoolID2", all.x=T)
summary(Fidelity.sec$Fidelity)
Fidelity.sec$Fidelity[is.na(Fidelity.sec$Fidelity)] <- 0
summary(Fidelity.sec$Fidelity)

#####

#FIDELITY ANALYSIS, FULL - NON-IMPURED

model.coeffs.fid <- NULL
effect.size.fid <- NULL
FID.n.primary <- NULL

#define consecutive numbers for bootstrap selection within schools
Fidelity.prim$number <- 1:nrow(Fidelity.prim)

#Actual bootstrap
for (boot in 1:1000) {

#generate list of student codes to select from
set.seed<-boot
id.list <- tapply(Fidelity.prim$number, INDEX=Fidelity.prim$SchoolID2,
sample, replace=T)
id.list <- unlist(id.list)
boot.data <- Fidelity.prim[id.list, ]

#Formula in SAP
#KS2past ~ KS1 + REG + GfW + (1|school)
prim.model <- lmer(CalcTotal_Overall2 ~ KS1_WRITPOINTS + Region +
I(Fidelity/100) + (1 | SchoolID2), data=boot.data)
model.coeffs.fid <- rbind(model.coeffs.fid, fixef(prim.model))

FID.n.primary <- rbind(FID.n.primary,
length(prim.model@frame$KS1_WRITPOINTS))

varcomp <- data.frame(VarCorr(prim.model))
#collecting coefficient separately and the two variances

```

```

effect.size.fid <- rbind(effect.size.fid, c(
summary(prim.model)$coefficients[4,1], varcomp[1,4], varcomp[2,4]
))

rm(prim.model, boot.data, varcomp)

if (boot%%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap

mean(FID.n.primary)
sd(FID.n.primary)

#Data for table 1
#coefficient data
apply(model.coeffs.fid, MARGIN=2, mean)
apply(model.coeffs.fid, MARGIN=2, sd)
apply(model.coeffs.fid, MARGIN=2, quantile, probs=c(.025, .975))
hist(model.coeffs.fid[,4])

effect.size.read(effect.size.fid)

#FIDELITY ANALYSIS, FSM-ONLY - NON-IMPUTED

model.coeffs.fid.fsm <- NULL
effect.size.fid.fsm <- NULL
FID.n.primary.fsm <- NULL

#Actual bootstrap
for (boot in 1:1000) {

boot.dat <- subset(Fidelity.prim, EVERFSM_ALL_SPR17==1)
boot.dat$number <- 1:nrow(boot.dat)

#generate list of student codes to select from
set.seed<-boot
id.list <- tapply(boot.dat$number, INDEX=boot.dat$SchoolID2, sample,
replace=T)
id.list <- unlist(id.list)
boot.data <- boot.dat[id.list, ]
rm(boot.dat)

#Formula in SAP
#KS2past ~ KS1 + REG + GfW + (1|school)
prim.model <- lmer(CalcTotal_Overall12 ~ KS1_WRITPOINTS + Region +
I(Fidelity/100) + (1 | SchoolID2), data=boot.data)
model.coeffs.fid.fsm <- rbind(model.coeffs.fid.fsm, fixef(prim.model))

FID.n.primary.fsm <- rbind(FID.n.primary.fsm,
length(prim.model@frame$KS1_WRITPOINTS))

varcomp <- data.frame(VarCorr(prim.model))
#collecting coefficient separately and the two variances
effect.size.fid.fsm <- rbind(effect.size.fid.fsm, c(
summary(prim.model)$coefficients[4,1], varcomp[1,4], varcomp[2,4]
))

rm(prim.model, boot.data, varcomp)

```

```

if (boot%%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap

mean(FID.n.primary.fsm)
sd(FID.n.primary.fsm)

#Data for table 1
#coefficient data
apply(model.coeffs.fid.fsm, MARGIN=2, mean)
apply(model.coeffs.fid.fsm, MARGIN=2, sd)
apply(model.coeffs.fid.fsm, MARGIN=2, quantile, probs=c(.025, .975))
hist(model.coeffs.fid.fsm[,4])

effect.size.read(effect.size.fid.fsm)

# FIDELITY ANALYSIS, IMPUTED PRIMARY OUTCOME FULL SAMPLE

primary.coeffs.i.fid <- NULL
effect.size.i.fid <- NULL
FID.n.primary.i <- NULL

t1 <- Sys.time()
#Actual bootstrap
for (boot in 1:1000) {

if (is.data.frame(imputed.data$imputations[boot][[1]])==F) {
  capt.conv <- c(capt.conv, boot)
} #end of if-skip

if (is.data.frame(imputed.data$imputations[boot][[1]])==T) {

boot.dat <- imputed.data$imputations[boot][[1]]
boot.dat <- merge(boot.dat, fidelity.merge, by="SchoolID2", all.x=T)
boot.dat$Fidelity[is.na(boot.dat$Fidelity)] <- 0
boot.dat$number <- 1:nrow(boot.dat)

#generate list of student codes to select from
set.seed<-boot
id.list <- tapply(boot.dat$number, INDEX=boot.dat$SchoolID2, sample,
replace=T)
id.list <- unlist(id.list)
boot.data <- boot.dat[id.list, ]

#Formula in SAP
#KS2past ~ KS1 + REG + GfW + (1|school)
prim.model <- lmer(CalcTotal_Overall12 ~ KS1_WRITPOINTS + Region +
I(Fidelity/100) + (1| SchoolID2), data=boot.data)
primary.coeffs.i.fid <- rbind(primary.coeffs.i.fid, fixef(prim.model))

FID.n.primary.i <- rbind(FID.n.primary.i,
length(prim.model@frame$KS1_WRITPOINTS))

varcomp <- data.frame(VarCorr(prim.model))
#collecting coefficient separately and the two variances
effect.size.i.fid <- rbind(effect.size.i.fid, c(
summary(prim.model)$coefficients[4,1], varcomp[1,4], varcomp[2,4]
))

```

```

rm(prim.model, boot.dat, boot.data, varcomp)

} #end of if-compute

if (boot%%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap primary outcome
Sys.time() - t1

#
mean(FID.n.primary.i)
sd(FID.n.primary.i)

#coefficient data
head(primary.coeffs.i.fid)
apply(primary.coeffs.i.fid, MARGIN=2, mean, na.rm=T)
apply(primary.coeffs.i.fid, MARGIN=2, sd, na.rm=T)
apply(primary.coeffs.i.fid, MARGIN=2, quantile, probs=c(.025, .975) ,
na.rm=T)

effect.size.read(effect.size.i.fid)

### FIDELITY ANALYSIS, Only FSM, imputed

primary.coeffs.i.fid.fsm <- NULL
effect.size.i.fid.fsm <- NULL
FID.n.primary.i.fsm <- NULL

t1 <- Sys.time()
#Actual bootstrap
for (boot in 1:1000) {

if (is.data.frame(imputed.data$imputations[boot][[1]])==F) {
  capt.conv <- c(capt.conv, boot)
} #end of if-skip

if (is.data.frame(imputed.data$imputations[boot][[1]])==T) {

boot.dat <- imputed.data$imputations[boot][[1]]
boot.dat <- merge(boot.dat, fidelity.merge, by="SchoolID2", all.x=T)
boot.dat$Fidelity[is.na(boot.dat$Fidelity)] <- 0
boot.dat <- subset(boot.dat, EVERFSM_ALL_SPR17==1)
boot.dat$number <- 1:nrow(boot.dat)

#generate list of student codes to select from
set.seed<-boot
id.list <- tapply(boot.dat$number, INDEX=boot.dat$SchoolID2, sample,
replace=T)
id.list <- unlist(id.list)
boot.data <- boot.dat[id.list, ]

#Formula in SAP
#KS2past ~ KS1 + REG + GfW + (1|school)
prim.model <- lmer(CalcTotal_Overall12 ~ KS1_WRITPOINTS + Region +
I(Fidelity/100) + (1| SchoolID2), data=boot.data)
primary.coeffs.i.fid.fsm <- rbind(primary.coeffs.i.fid.fsm,
fixef(prim.model))

```

```

FID.n.primary.i.fsm          <-          rbind(FID.n.primary.i.fsm,
length(prim.model@frame$KS1_WRITPOINTS))

varcomp <- data.frame(VarCorr(prim.model))
#collecting coefficient separately and the two variances
effect.size.i.fid.fsm <- rbind(effect.size.i.fid.fsm, c(
summary(prim.model)$coefficients[4,1], varcomp[1,4], varcomp[2,4]
))

rm(prim.model, boot.dat, boot.data, varcomp)

} #end of if-compute

if (boot%%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap primary outcome
Sys.time() - t1

#

mean(FID.n.primary.i.fsm)
sd(FID.n.primary.i.fsm)

#coefficient data
head(primary.coeffs.i.fid.fsm)
apply(primary.coeffs.i.fid.fsm, MARGIN=2, mean, na.rm=T)
apply(primary.coeffs.i.fid.fsm, MARGIN=2, sd, na.rm=T)
apply(primary.coeffs.i.fid.fsm, MARGIN=2, quantile, probs=c(.025, .975) ,
na.rm=T)

effect.size.read(effect.size.i.fid.fsm)

# FIDELITY ANALYSIS, SECONDARY OUTCOMES

#prepare data
#check availability of outcome
names(Fidelity.sec)
table(Fidelity.sec$KS2_WRITTAOUTCOME_Code)

t1 <- Sys.time()
#####
# FIDELITY ANALYSIS, bootstrap of SECONDARY OUTCOME 1 // WRIT
#define variables capturing estimates:
secondary.coeffs.writ.fid <- NULL
effect.size.writ.fid <- NULL
FID.n.secondary.writ <- NULL

#define consecutive numbers for bootstrap selection within schools
Fidelity.sec$number <- 1:nrow(Fidelity.sec)

#Actual bootstrap
for (boot in 1:1000) {

set.seed<-boot
id.list <- tapply(Fidelity.sec$number, INDEX=Fidelity.sec$SchoolID2, sample,
replace=T)
id.list <- unlist(id.list)
boot.data <- Fidelity.sec[id.list, ]

```

```

#Formula in SAP
#KS2past ~ KS1 + REG + GfW + (1|school)
writ.model <- lmer(KS2_WRITTAOUTCOME_Code ~ KS1_WRITPOINTS + Region +
I(Fidelity/100) + (1| SchoolID2), data=boot.data)
secondary.coeffs.writ.fid <- rbind(secondary.coeffs.writ.fid,
fixef(writ.model))

FID.n.secondary.writ <- rbind(FID.n.secondary.writ,
length(writ.model@frame$KS1_WRITPOINTS))

varcomp <- data.frame(VarCorr(writ.model))
#collecting coefficient separately and the two variances
effect.size.writ.fid <- rbind(effect.size.writ.fid, c(
summary(writ.model)$coefficients[4,1], varcomp[1,4], varcomp[2,4]
))

rm(writ.model, boot.data, varcomp)

if (boot%%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap primary outcome

Sys.time() - t1

#

mean(FID.n.secondary.writ)
sd(FID.n.secondary.writ)

#coefficient data
head(secondary.coeffs.writ.fid)
apply(secondary.coeffs.writ.fid, MARGIN=2, mean)
apply(secondary.coeffs.writ.fid, MARGIN=2, sd)
apply(secondary.coeffs.writ.fid, MARGIN=2, quantile, probs=c(.025, .975))
hist(secondary.coeffs.writ.fid[,4])

effect.size.read(effect.size.writ.fid)

#####
# FIDELITY ANALYSIS, SECONDARY OUTCOME 2 // GPS

#check variable
table(Fidelity.sec$KS2_GPSMRK)

secondary.coeffs.gps.fid <- NULL
effect.size.gps.fid <- NULL
FID.n.secondary.gps <- NULL

#define consecutive numbers for bootstrap selection within schools
Fidelity.sec$number <- 1:nrow(Fidelity.sec)

#Actual bootstrap
for (boot in 1:1000) {

set.seed<-boot
id.list <- tapply(Fidelity.sec$number, INDEX=Fidelity.sec$SchoolID2, sample,
replace=T)
id.list <- unlist(id.list)
boot.data <- Fidelity.sec[id.list, ]

```



```

#Formula in SAP
#KS2past ~ KS1 + REG + GfW + (1|school)
gps.model <- lmer(KS2_GPSMRK ~ KS1_WRITPOINTS + Region + I(Fidelity/100) +
(1| SchoolID2), data=boot.data)
secondary.coeffs.gps.fid <- rbind(secondary.coeffs.gps.fid,
fixef(gps.model))

FID.n.secondary.gps <- rbind(FID.n.secondary.gps,
length(gps.model@frame$KS1_WRITPOINTS))

varcomp <- data.frame(VarCorr(gps.model))
#collecting coefficient separately and the two variances
effect.size.gps.fid <- rbind(effect.size.gps.fid, c(
summary(gps.model)$coefficients[4,1], varcomp[1,4], varcomp[2,4]
))

rm(gps.model, boot.data, varcomp)

if (boot%%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap primary outcome

Sys.time() - t1

#
mean(FID.n.secondary.gps)
sd(FID.n.secondary.gps)

#coefficient data
head(secondary.coeffs.gps.fid)
apply(secondary.coeffs.gps.fid, MARGIN=2, mean)
apply(secondary.coeffs.gps.fid, MARGIN=2, sd)
apply(secondary.coeffs.gps.fid, MARGIN=2, quantile, probs=c(.025, .975))
hist(secondary.coeffs.gps.fid[,4])

effect.size.read(effect.size.gps.fid)

#####
# FIDELITY ANALYSIS, SECONDARY OUTCOME 3 // READ

#check variable
table(Fidelity.sec$KS2_READMRK)

secondary.coeffs.read.fid <- NULL
effect.size.read.fid <- NULL
FID.n.secondary.read <- NULL

#define consecutive numbers for bootstrap selection within schools
Fidelity.sec$number <- 1:nrow(Fidelity.sec)

#Actual bootstrap
for (boot in 1:1000) {

set.seed<-boot
id.list <- tapply(Fidelity.sec$number, INDEX=Fidelity.sec$SchoolID2, sample,
replace=T)
id.list <- unlist(id.list)
boot.data <- Fidelity.sec[id.list, ]

```

```

#Formula in SAP
#KS2past ~ KS1 + REG + GfW + (1|school)
read.model <- lmer(KS2_READMRK ~ KS1_WRITPOINTS + Region + I(Fidelity/100)
+ (1| SchoolID2), data=boot.data)
secondary.coeffs.read.fid <- rbind(secondary.coeffs.read.fid,
fixef(read.model))

FID.n.secondary.read <- rbind(FID.n.secondary.read,
length(read.model@frame$KS1_WRITPOINTS))

varcomp <- data.frame(VarCorr(read.model))
#collecting coefficient separately and the two variances
effect.size.read.fid <- rbind(effect.size.read.fid, c(
summary(read.model)$coefficients[4,1], varcomp[1,4], varcomp[2,4]
))

rm(read.model, boot.data, varcomp)

if (boot%%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap primary outcome

Sys.time() - t1

#
mean(FID.n.secondary.read)
sd(FID.n.secondary.read)

#coefficient data
head(secondary.coeffs.read.fid)
apply(secondary.coeffs.read.fid, MARGIN=2, mean)
apply(secondary.coeffs.read.fid, MARGIN=2, sd)
apply(secondary.coeffs.read.fid, MARGIN=2, quantile, probs=c(.025, .975))
hist(secondary.coeffs.read.fid[,4])

effect.size.read(effect.size.read.fid)

#####
### FIDELITY ANALYSIS, secondary imputed to follow

#FIDELITY ANALYSIS, bootstrap of SECONDARY OUTCOME 1 // WRIT
t1 <- Sys.time()
#####
#define variables capturing estimates:
secondary.coeffs.writ.fid.i <- NULL
effect.size.writ.fid.i <- NULL
FID.n.secondary.writ.i <- NULL

#Actual bootstrap
for (boot in 1:1000) {

boot.dat <- imputed.data.secondary$imputations[boot][[1]]
boot.dat <- merge(boot.dat, fidelity.merge, by="SchoolID2", all.x=T)
boot.dat$Fidelity[is.na(boot.dat$Fidelity)] <- 0
boot.dat$number <- 1:nrow(boot.dat)

#generate list of student codes to select from
set.seed<-boot

```

```

id.list <- tapply(boot.dat$number, INDEX=boot.dat$SchoolID2, sample,
replace=T)
id.list <- unlist(id.list)
boot.data <- boot.dat[id.list, ]
rm(boot.dat)

#Formula in SAP
#KS2past ~ KS1 + REG + GfW + (1|school)
writ.model.i <- lmer(KS2_WRITTAOUTCOME_Code ~ KS1_WRITPOINTS + Region +
I(Fidelity/100) + (1| SchoolID2), data=boot.data)
secondary.coeffs.writ.fid.i <- rbind(secondary.coeffs.writ.fid.i,
fixef(writ.model.i))

FID.n.secondary.writ.i <- rbind(FID.n.secondary.writ.i,
length(writ.model.i@frame$KS1_WRITPOINTS))

varcomp <- data.frame(VarCorr(writ.model.i))
#collecting coefficient separately and the two variances
effect.size.writ.fid.i <- rbind(effect.size.writ.fid.i, c(
summary(writ.model.i)$coefficients[4,1], varcomp[1,4], varcomp[2,4]
))

rm(writ.model.i, boot.data, varcomp)

if (boot%%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap primary outcome

Sys.time() - t1

#
mean(FID.n.secondary.writ.i)
sd(FID.n.secondary.writ.i)

#coefficient data
head(secondary.coeffs.writ.fid.i)
apply(secondary.coeffs.writ.fid.i, MARGIN=2, mean)
apply(secondary.coeffs.writ.fid.i, MARGIN=2, sd)
apply(secondary.coeffs.writ.fid.i, MARGIN=2, quantile, probs=c(.025, .975))
hist(secondary.coeffs.writ.fid.i[,4])

effect.size.read(effect.size.writ.fid.i)

#####
# FIDELITY ANALYSIS, SECONDARY OUTCOME 2 // GPS

t1 <- Sys.time()
secondary.coeffs.gps.fid.i <- NULL
effect.size.gps.fid.i <- NULL
FID.n.secondary.gps.i <- NULL

#Actual bootstrap
for (boot in 1:1000) {

boot.dat <- imputed.data.secondary$imputations[boot][[1]]
boot.dat <- merge(boot.dat, fidelity.merge, by="SchoolID2", all.x=T)
boot.dat$Fidelity[is.na(boot.dat$Fidelity)] <- 0
boot.dat$number <- 1:nrow(boot.dat)

```

```

#generate list of student codes to select from
set.seed<-boot
id.list <- tapply(boot.dat$number, INDEX=boot.dat$SchoolID2, sample,
replace=T)
id.list <- unlist(id.list)
boot.data <- boot.dat[id.list, ]
rm(boot.dat)

#Formula in SAP
#KS2past ~ KS1 + REG + GfW + (1|school)
gps.model.i <- lmer(KS2_GPSMRK ~ KS1_WRITPOINTS + Region + I(Fidelity/100)
+ (1| SchoolID2), data=boot.data)
secondary.coeffs.gps.fid.i <- rbind(secondary.coeffs.gps.fid.i,
fixef(gps.model.i))

FID.n.secondary.gps.i <- rbind(FID.n.secondary.gps.i,
length(gps.model.i@frame$KS1_WRITPOINTS))

varcomp <- data.frame(VarCorr(gps.model.i))
#collecting coefficient separately and the two variances
effect.size.gps.fid.i <- rbind(effect.size.gps.fid.i, c(
summary(gps.model.i)$coefficients[4,1], varcomp[1,4], varcomp[2,4]
))

rm(gps.model.i, boot.data, varcomp)

if (boot%%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap primary outcome

Sys.time() - t1

#
mean(FID.n.secondary.gps.i)
sd(FID.n.secondary.gps.i)

#coefficient data
head(secondary.coeffs.gps.fid.i)
apply(secondary.coeffs.gps.fid.i, MARGIN=2, mean)
apply(secondary.coeffs.gps.fid.i, MARGIN=2, sd)
apply(secondary.coeffs.gps.fid.i, MARGIN=2, quantile, probs=c(.025, .975))
hist(secondary.coeffs.gps.fid.i[,4])

effect.size.read(effect.size.gps.fid.i)

#####
# FIDELITY ANALYSIS, SECONDARY OUTCOME 3 // READ

t1 <- Sys.time()
secondary.coeffs.read.fid.i <- NULL
effect.size.read.fid.i <- NULL
FID.n.secondary.read.i <- NULL

#Actual bootstrap
for (boot in 1:1000) {

boot.dat <- imputed.data.secondary$imputations[boot][[1]]
boot.dat <- merge(boot.dat, fidelity.merge, by="SchoolID2", all.x=T)
boot.dat$Fidelity[is.na(boot.dat$Fidelity)] <- 0

```

```

boot.dat$number <- 1:nrow(boot.dat)

#generate list of student codes to select from
set.seed<-boot
id.list <- tapply(boot.dat$number, INDEX=boot.dat$SchoolID2, sample,
replace=T)
id.list <- unlist(id.list)
boot.data <- boot.dat[id.list, ]
rm(boot.dat)

#Formula in SAP
#KS2past ~ KS1 + REG + GfW + (1|school)
read.model.i <- lmer(KS2_READMRK ~ KS1_WRITPOINTS + Region + I(Fidelity/100)
+ (1| SchoolID2), data=boot.data)
secondary.coeffs.read.fid.i <- rbind(secondary.coeffs.read.fid.i,
fixef(read.model.i))

FID.n.secondary.read.i <- rbind(FID.n.secondary.read.i,
length(read.model.i@frame$KS1_WRITPOINTS))

varcomp <- data.frame(VarCorr(read.model.i))
#collecting coefficient separately and the two variances
effect.size.read.fid.i <- rbind(effect.size.read.fid.i, c(
summary(read.model.i)$coefficients[4,1], varcomp[1,4], varcomp[2,4]
))

rm(read.model.i, boot.data, varcomp)

if (boot%%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap primary outcome

Sys.time() - t1

#
mean(FID.n.secondary.read.i)
sd(FID.n.secondary.read.i)

#coefficient data
head(secondary.coeffs.read.fid.i)
apply(secondary.coeffs.read.fid.i, MARGIN=2, mean)
apply(secondary.coeffs.read.fid.i, MARGIN=2, sd)
apply(secondary.coeffs.read.fid.i, MARGIN=2, quantile, probs=c(.025, .975))
hist(secondary.coeffs.read.fid.i[,4])

effect.size.read(effect.size.read.fid.i)

#saved all the above for exchange with EEF:
save.image("L:\\Jan\\001 Analysis\\Include in report MAY2018\\2018 05 24 a1
12.RData")

#Added in July 2018:
#####Analysis of CPD4 data as a compliance indicator

#Read 4day CPD data
Fidelity4CPD <- readXL("L:/Jan/000_ORIGINAL/Implementation Fidelity v3
20180420_4CPD.xlsx", rownames=FALSE, header=TRUE, na="",
sheet="Implementation Fidelity", stringsAsFactors=TRUE)

```

```

names(Fidelity4CPD)
Fidelity4CPD$Imp.Fid...4
mean(Fidelity4CPD$Imp.Fid...4)
sd(Fidelity4CPD$Imp.Fid...4)
length(Fidelity4CPD$Imp.Fid...4)

fidelity4CPD.merge <- subset(Fidelity4CPD, select=c("SchoolID2",
"Imp.Fid...4"))
names(fidelity4CPD.merge) <- c("SchoolID2", "Fidelity4CPD")

fidelity4CPD.merge

#merge with main data set
Fidelity.prim4CPD <- merge(GfWdata3.prim, fidelity4CPD.merge,
by="SchoolID2", all.x=T)
summary(Fidelity.prim4CPD$Fidelity4CPD)
Fidelity.prim4CPD$Fidelity4CPD[is.na(Fidelity.prim4CPD$Fidelity4CPD)] <- 0
summary(Fidelity.prim4CPD$Fidelity4CPD)

#merge with secondary data set
Fidelity.sec4CPD <- merge(GfWdata3.sec, fidelity4CPD.merge, by="SchoolID2",
all.x=T)
summary(Fidelity.sec4CPD$Fidelity4CPD)
Fidelity.sec4CPD$Fidelity4CPD[is.na(Fidelity.sec4CPD$Fidelity4CPD)] <- 0
summary(Fidelity.sec4CPD$Fidelity4CPD)

#####
#FIDELITY ANALYSIS, FULL - NON-IMPUTED

model.coeffs.fid4 <- NULL
effect.size.fid4 <- NULL
FID4.n.primary <- NULL

#define consecutive numbers for bootstrap selection within schools
Fidelity.prim4CPD$number <- 1:nrow(Fidelity.prim4CPD)

#Actual bootstrap
for (boot in 1:1000) {

#generate list of student codes to select from
set.seed<-boot
id.list <- tapply(Fidelity.prim4CPD$number,
INDEX=Fidelity.prim4CPD$SchoolID2, sample, replace=T)
id.list <- unlist(id.list)
boot.data <- Fidelity.prim4CPD[id.list, ]

#Formula in SAP
#KS2past ~ KS1 + REG + GfW + (1|school)
prim.model <- lmer(CalcTotal_Overall12 ~ KS1_WRITPOINTS + Region +
I(Fidelity4CPD/100) + (1 | SchoolID2), data=boot.data)
model.coeffs.fid4 <- rbind(model.coeffs.fid4, fixef(prim.model))

FID4.n.primary <- rbind(FID4.n.primary,
length(prim.model@frame$KS1_WRITPOINTS))

varcomp <- data.frame(VarCorr(prim.model))
#collecting coefficient separately and the two variances
effect.size.fid4 <- rbind(effect.size.fid4, c(
summary(prim.model)$coefficients[4,1], varcomp[1,4], varcomp[2,4]

```

```

))

rm(prim.model, boot.data, varcomp)

if (boot%%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap

mean(FID4.n.primary)
sd(FID4.n.primary)

#coefficient data
apply(model.coeffs.fid4, MARGIN=2, mean)
apply(model.coeffs.fid4, MARGIN=2, sd)
apply(model.coeffs.fid4, MARGIN=2, quantile, probs=c(.025, .975))
hist(model.coeffs.fid4[,4])

effect.size.read(effect.size.fid4)

#####
#FIDELITY ANALYSIS, FSM-ONLY - NON-IMPUTED

model.coeffs.fid4.fsm <- NULL
effect.size.fid4.fsm <- NULL
FID4.n.primary.fsm <- NULL

#Actual bootstrap
for (boot in 1:1000) {

boot.dat <- subset(Fidelity.prim4CPD, EVERFSM_ALL_SPR17==1)
boot.dat$number <- 1:nrow(boot.dat)

#generate list of student codes to select from
set.seed<-boot
id.list <- tapply(boot.dat$number, INDEX=boot.dat$SchoolID2, sample,
replace=T)
id.list <- unlist(id.list)
boot.data <- boot.dat[id.list, ]
rm(boot.dat)

#Formula in SAP
#KS2past ~ KS1 + REG + GfW + (1|school)
prim.model <- lmer(CalcTotal_Overall12 ~ KS1_WRITPOINTS + Region +
I(Fidelity4CPD/100) + (1 | SchoolID2), data=boot.data)
model.coeffs.fid4.fsm <- rbind(model.coeffs.fid4.fsm, fixef(prim.model))

FID4.n.primary.fsm <- rbind(FID4.n.primary.fsm,
length(prim.model@frame$KS1_WRITPOINTS))

varcomp <- data.frame(VarCorr(prim.model))
#collecting coefficient separately and the two variances
effect.size.fid4.fsm <- rbind(effect.size.fid4.fsm, c(
summary(prim.model)$coefficients[4,1], varcomp[1,4], varcomp[2,4]
))

rm(prim.model, boot.data, varcomp)

if (boot%%10==0) {

```

```

plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap

mean(FID4.n.primary.fsm)
sd(FID4.n.primary.fsm)

#coefficient data
apply(model.coeffs.fid4.fsm, MARGIN=2, mean)
apply(model.coeffs.fid4.fsm, MARGIN=2, sd)
apply(model.coeffs.fid4.fsm, MARGIN=2, quantile, probs=c(.025, .975))
hist(model.coeffs.fid4.fsm[,4])

effect.size.read(effect.size.fid4.fsm)

#####
#bootstrap of IMPUTED PRIMARY OUTCOME FULL SAMPLE

primary.coeffs.i.fid4 <- NULL
effect.size.i.fid4 <- NULL
FID4.n.primary.i <- NULL

t1 <- Sys.time()
#Actual bootstrap
for (boot in 1:1000) {

if (is.data.frame(imputed.data$imputations[boot][[1]])==F) {
  capt.conv <- c(capt.conv, boot)
} #end of if-skip

if (is.data.frame(imputed.data$imputations[boot][[1]])==T) {

boot.dat <- imputed.data$imputations[boot][[1]]
boot.dat <- merge(boot.dat, fidelity4CPD.merge, by="SchoolID2", all.x=T)
boot.dat$Fidelity4CPD[is.na(boot.dat$Fidelity4CPD)] <- 0
boot.dat$number <- 1:nrow(boot.dat)

#generate list of student codes to select from
set.seed<-boot
id.list <- tapply(boot.dat$number, INDEX=boot.dat$SchoolID2, sample,
replace=T)
id.list <- unlist(id.list)
boot.data <- boot.dat[id.list, ]

#Formula in SAP
#KS2past ~ KS1 + REG + GfW + (1|school)
prim.model <- lmer(CalcTotal_Overall12 ~ KS1_WRITPOINTS + Region +
I(Fidelity4CPD/100) + (1| SchoolID2), data=boot.data)
primary.coeffs.i.fid4 <- rbind(primary.coeffs.i.fid4, fixef(prim.model))

FID4.n.primary.i <- rbind(FID4.n.primary.i,
length(prim.model@frame$KS1_WRITPOINTS))

varcomp <- data.frame(VarCorr(prim.model))
#collecting coefficient separately and the two variances
effect.size.i.fid4 <- rbind(effect.size.i.fid4, c(
summary(prim.model)$coefficients[4,1], varcomp[1,4], varcomp[2,4]
))

rm(prim.model, boot.dat, boot.data, varcomp)

```



```

} #end of if-compute

if (boot%%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap primary outcome
Sys.time() - t1

#
mean(FID4.n.primary.i)
sd(FID4.n.primary.i)

#coefficient data
head(primary.coeffs.i.fid4)
apply(primary.coeffs.i.fid4, MARGIN=2, mean, na.rm=T)
apply(primary.coeffs.i.fid4, MARGIN=2, sd, na.rm=T)
apply(primary.coeffs.i.fid4, MARGIN=2, quantile, probs=c(.025, .975) ,
na.rm=T)

effect.size.read(effect.size.i.fid4)

#####
#Only FSM, imputed

primary.coeffs.i.fid4.fsm <- NULL
effect.size.i.fid4.fsm <- NULL
FID4.n.primary.i.fsm <- NULL

t1 <- Sys.time()
#Actual bootstrap
for (boot in 1:1000) {

if (is.data.frame(imputed.data$imputations[boot][[1]])==F) {
  capt.conv <- c(capt.conv, boot)
} #end of if-skip

if (is.data.frame(imputed.data$imputations[boot][[1]])==T) {

boot.dat <- imputed.data$imputations[boot][[1]]
boot.dat <- merge(boot.dat, fidelity4CPD.merge, by="SchoolID2", all.x=T)
boot.dat$Fidelity4CPD[is.na(boot.dat$Fidelity4CPD)] <- 0
boot.dat <- subset(boot.dat, EVERFSM_ALL_SPR17==1)
boot.dat$number <- 1:nrow(boot.dat)

#generate list of student codes to select from
set.seed<-boot
id.list <- tapply(boot.dat$number, INDEX=boot.dat$SchoolID2, sample,
replace=T)
id.list <- unlist(id.list)
boot.data <- boot.dat[id.list, ]

#Formula in SAP
#KS2past ~ KS1 + REG + GfW + (1|school)
prim.model <- lmer(CalcTotal_Overall2 ~ KS1_WRITPOINTS + Region +
I(Fidelity4CPD/100) + (1| SchoolID2), data=boot.data)
primary.coeffs.i.fid4.fsm <- rbind(primary.coeffs.i.fid4.fsm,
fixef(prim.model))

```

```

FID4.n.primary.i.fsm <- rbind(FID4.n.primary.i.fsm,
length(prim.model@frame$KS1_WRITPOINTS))

varcomp <- data.frame(VarCorr(prim.model))
#collecting coefficient separately and the two variances
effect.size.i.fid4.fsm <- rbind(effect.size.i.fid4.fsm, c(
summary(prim.model)$coefficients[4,1], varcomp[1,4], varcomp[2,4]
))

rm(prim.model, boot.dat, boot.data, varcomp)

} #end of if-compute

if (boot%%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap primary outcome
Sys.time() - t1

#
mean(FID4.n.primary.i.fsm)
sd(FID4.n.primary.i.fsm)

#coefficient data
head(primary.coeffs.i.fid4.fsm)
apply(primary.coeffs.i.fid4.fsm, MARGIN=2, mean, na.rm=T)
apply(primary.coeffs.i.fid4.fsm, MARGIN=2, sd, na.rm=T)
apply(primary.coeffs.i.fid4.fsm, MARGIN=2, quantile, probs=c(.025, .975) ,
na.rm=T)

effect.size.read(effect.size.i.fid4.fsm)

#####
#SECONDARY OUTCOMES

#check availability of outcome
names(Fidelity.sec4CPD)
table(Fidelity.sec4CPD$KS2_WRITTAOUTCOME_Code)

#if re-opened workspace:
#library(multilevel)
#library(lme4)

t1 <- Sys.time()
#####
#bootstrap of SECONDARY OUTCOME 1 // WRIT
#define variables capturing estimates:
secondary.coeffs.writ.fid4 <- NULL
effect.size.writ.fid4 <- NULL
FID4.n.secondary.writ <- NULL

#define consecutive numbers for bootstrap selection within schools
Fidelity.sec4CPD$number <- 1:nrow(Fidelity.sec4CPD)

#Actual bootstrap
for (boot in 1:1000) {

set.seed<-boot
id.list <- tapply(Fidelity.sec4CPD$number,
INDEX=Fidelity.sec4CPD$SchoolID2, sample, replace=T)

```

```

id.list <- unlist(id.list)
boot.data <- Fidelity.sec4CPD[id.list, ]

#Formula in SAP
#KS2past ~ KS1 + REG + GfW + (1|school)
writ.model <- lmer(KS2_WRITTAOUTCOME_Code ~ KS1_WRITPOINTS + Region +
I(Fidelity4CPD/100) + (1| SchoolID2), data=boot.data)
secondary.coeffs.writ.fid4 <- rbind(secondary.coeffs.writ.fid4,
fixef(writ.model))

FID4.n.secondary.writ <- rbind(FID4.n.secondary.writ,
length(writ.model@frame$KS1_WRITPOINTS))

varcomp <- data.frame(VarCorr(writ.model))
#collecting coefficient separately and the two variances
effect.size.writ.fid4 <- rbind(effect.size.writ.fid4, c(
summary(writ.model)$coefficients[4,1], varcomp[1,4], varcomp[2,4]
))

rm(writ.model, boot.data, varcomp)

if (boot%%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap primary outcome

Sys.time() - t1

#
mean(FID4.n.secondary.writ)
sd(FID4.n.secondary.writ)

#coefficient data
head(secondary.coeffs.writ.fid4)
apply(secondary.coeffs.writ.fid4, MARGIN=2, mean)
apply(secondary.coeffs.writ.fid4, MARGIN=2, sd)
apply(secondary.coeffs.writ.fid4, MARGIN=2, quantile, probs=c(.025, .975))
hist(secondary.coeffs.writ.fid4[,4])

effect.size.read(effect.size.writ.fid4)

#####
#bootstrap of SECONDARY OUTCOME 2 // GPS

#check variable
table(Fidelity.sec4CPD$KS2_GPSMRK)

secondary.coeffs.gps.fid4 <- NULL
effect.size.gps.fid4 <- NULL
FID4.n.secondary.gps <- NULL

#define consecutive numbers for bootstrap selection within schools
Fidelity.sec4CPD$number <- 1:nrow(Fidelity.sec4CPD)

#Actual bootstrap
for (boot in 1:1000) {

set.seed<-boot

```

```

id.list <- tapply(Fidelity.sec4CPD$number,
INDEX=Fidelity.sec4CPD$SchoolID2, sample, replace=T)
id.list <- unlist(id.list)
boot.data <- Fidelity.sec4CPD[id.list, ]

#Formula in SAP
#KS2past ~ KS1 + REG + GfW + (1|school)
gps.model <- lmer(KS2_GPSMRK ~ KS1_WRITPOINTS + Region +
I(Fidelity4CPD/100) + (1| SchoolID2), data=boot.data)
secondary.coeffs.gps.fid4 <- rbind(secondary.coeffs.gps.fid4,
fixef(gps.model))

FID4.n.secondary.gps <- rbind(FID4.n.secondary.gps,
length(gps.model@frame$KS1_WRITPOINTS))

varcomp <- data.frame(VarCorr(gps.model))
#collecting coefficient separately and the two variances
effect.size.gps.fid4 <- rbind(effect.size.gps.fid4, c(
summary(gps.model)$coefficients[4,1], varcomp[1,4], varcomp[2,4]
))

rm(gps.model, boot.data, varcomp)

if (boot%%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap primary outcome

Sys.time() - t1

#
mean(FID4.n.secondary.gps)
sd(FID4.n.secondary.gps)

#coefficient data
head(secondary.coeffs.gps.fid4)
apply(secondary.coeffs.gps.fid4, MARGIN=2, mean)
apply(secondary.coeffs.gps.fid4, MARGIN=2, sd)
apply(secondary.coeffs.gps.fid4, MARGIN=2, quantile, probs=c(.025, .975))
hist(secondary.coeffs.gps.fid4[,4])

effect.size.read(effect.size.gps.fid4)

#####
#bootstrap of SECONDARY OUTCOME 3 // READ

#check variable
table(Fidelity.sec4CPD$KS2_READMRK)

secondary.coeffs.read.fid4 <- NULL
effect.size.read.fid4 <- NULL
FID4.n.secondary.read <- NULL

#define consecutive numbers for bootstrap selection within schools
Fidelity.sec4CPD$number <- 1:nrow(Fidelity.sec4CPD)

#Actual bootstrap
for (boot in 1:1000) {

```

```

set.seed<-boot
id.list <- tapply(Fidelity.sec4CPD$number,
INDEX=Fidelity.sec4CPD$SchoolID2, sample, replace=T)
id.list <- unlist(id.list)
boot.data <- Fidelity.sec4CPD[id.list, ]

#Formula in SAP
#KS2past ~ KS1 + REG + GfW + (1|school)
read.model <- lmer(KS2_READMRK ~ KS1_WRITPOINTS + Region +
I(Fidelity4CPD/100) + (1| SchoolID2), data=boot.data)
secondary.coeffs.read.fid4 <- rbind(secondary.coeffs.read.fid4,
fixef(read.model))

FID4.n.secondary.read <- rbind(FID4.n.secondary.read,
length(read.model@frame$KS1_WRITPOINTS))

varcomp <- data.frame(VarCorr(read.model))
#collecting coefficient separately and the two variances
effect.size.read.fid4 <- rbind(effect.size.read.fid4, c(
summary(read.model)$coefficients[4,1], varcomp[1,4], varcomp[2,4]
))

rm(read.model, boot.data, varcomp)

if (boot%%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap primary outcome

Sys.time() - t1

#
mean(FID4.n.secondary.read)
sd(FID4.n.secondary.read)

#coefficient data
head(secondary.coeffs.read.fid4)
apply(secondary.coeffs.read.fid4, MARGIN=2, mean)
apply(secondary.coeffs.read.fid4, MARGIN=2, sd)
apply(secondary.coeffs.read.fid4, MARGIN=2, quantile, probs=c(.025, .975))
hist(secondary.coeffs.read.fid4[,4])

effect.size.read(effect.size.read.fid4)

#####
###secondary imputed

#SECONDARY OUTCOMES -- IMPUTED FIDELITY

t1 <- Sys.time()
#####
#bootstrap of SECONDARY OUTCOME 1 // WRIT
#define variables capturing estimates:
secondary.coeffs.writ.fid4.i <- NULL
effect.size.writ.fid4.i <- NULL
FID4.n.secondary.writ.i <- NULL

#Actual bootstrap
for (boot in 1:1000) {

```

```

boot.dat <- imputed.data.secondary$imputations[boot][[1]]
boot.dat <- merge(boot.dat, fidelity4CPD.merge, by="SchoolID2", all.x=T)
boot.dat$Fidelity4CPD[is.na(boot.dat$Fidelity4CPD)] <- 0
boot.dat$number <- 1:nrow(boot.dat)

#generate list of student codes to select from
set.seed<-boot
id.list <- tapply(boot.dat$number, INDEX=boot.dat$SchoolID2, sample,
replace=T)
id.list <- unlist(id.list)
boot.data <- boot.dat[id.list, ]
rm(boot.dat)

#Formula in SAP
#KS2past ~ KS1 + REG + GfW + (1|school)
writ.model.i <- lmer(KS2_WRITTAOUTCOME_Code ~ KS1_WRITPOINTS + Region +
I(Fidelity4CPD/100) + (1| SchoolID2), data=boot.data)
secondary.coeffs.writ.fid4.i <- rbind(secondary.coeffs.writ.fid4.i,
fixef(writ.model.i))

FID4.n.secondary.writ.i <- rbind(FID4.n.secondary.writ.i,
length(writ.model.i@frame$KS1_WRITPOINTS))

varcomp <- data.frame(VarCorr(writ.model.i))
#collecting coefficient separately and the two variances
effect.size.writ.fid4.i <- rbind(effect.size.writ.fid4.i, c(
summary(writ.model.i)$coefficients[4,1], varcomp[1,4], varcomp[2,4]
))

rm(writ.model.i, boot.data, varcomp)

if (boot%%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap primary outcome

Sys.time() - t1

#
mean(FID4.n.secondary.writ.i)
sd(FID4.n.secondary.writ.i)

#coefficient data
head(secondary.coeffs.writ.fid4.i)
apply(secondary.coeffs.writ.fid4.i, MARGIN=2, mean)
apply(secondary.coeffs.writ.fid4.i, MARGIN=2, sd)
apply(secondary.coeffs.writ.fid4.i, MARGIN=2, quantile, probs=c(.025,
.975))
hist(secondary.coeffs.writ.fid4.i[,4])

effect.size.read(effect.size.writ.fid4.i)

#####
#bootstrap of SECONDARY OUTCOME 2 // GPS

t1 <- Sys.time()
secondary.coeffs.gps.fid4.i <- NULL
effect.size.gps.fid4.i <- NULL
FID4.n.secondary.gps.i <- NULL

```

```

#Actual bootstrap
for (boot in 1:1000) {

boot.dat <- imputed.data.secondary$imputations[boot][[1]]
boot.dat <- merge(boot.dat, fidelity4CPD.merge, by="SchoolID2", all.x=T)
boot.dat$Fidelity4CPD[is.na(boot.dat$Fidelity4CPD)] <- 0
boot.dat$number <- 1:nrow(boot.dat)

#generate list of student codes to select from
set.seed<-boot
id.list <- tapply(boot.dat$number, INDEX=boot.dat$SchoolID2, sample,
replace=T)
id.list <- unlist(id.list)
boot.data <- boot.dat[id.list, ]
rm(boot.dat)

#Formula in SAP
#KS2past ~ KS1 + REG + GfW + (1|school)
gps.model.i <- lmer(KS2_GPSMRK ~ KS1_WRITPOINTS + Region +
I(Fidelity4CPD/100) + (1| SchoolID2), data=boot.data)
secondary.coeffs.gps.fid4.i <- rbind(secondary.coeffs.gps.fid4.i,
fixef(gps.model.i))

FID4.n.secondary.gps.i <- rbind(FID4.n.secondary.gps.i,
length(gps.model.i@frame$KS1_WRITPOINTS))

varcomp <- data.frame(VarCorr(gps.model.i))
#collecting coefficient separately and the two variances
effect.size.gps.fid4.i <- rbind(effect.size.gps.fid4.i, c(
summary(gps.model.i)$coefficients[4,1], varcomp[1,4], varcomp[2,4]
))

rm(gps.model.i, boot.data, varcomp)

if (boot%%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap primary outcome

Sys.time() - t1

#Data for table 1

mean(FID4.n.secondary.gps.i)
sd(FID4.n.secondary.gps.i)

#coefficient data
head(secondary.coeffs.gps.fid4.i)
apply(secondary.coeffs.gps.fid4.i, MARGIN=2, mean)
apply(secondary.coeffs.gps.fid4.i, MARGIN=2, sd)
apply(secondary.coeffs.gps.fid4.i, MARGIN=2, quantile, probs=c(.025, .975))
hist(secondary.coeffs.gps.fid4.i[,4])

effect.size.read(effect.size.gps.fid4.i)

#####
#bootstrap of SECONDARY OUTCOME 3 // READ

t1 <- Sys.time()

```

```

secondary.coeffs.read.fid4.i <- NULL
effect.size.read.fid4.i <- NULL
FID4.n.secondary.read.i <- NULL

#Actual bootstrap
for (boot in 1:1000) {

boot.dat <- imputed.data.secondary$imputations[boot][[1]]
boot.dat <- merge(boot.dat, fidelity4CPD.merge, by="SchoolID2", all.x=T)
boot.dat$Fidelity4CPD[is.na(boot.dat$Fidelity4CPD)] <- 0
boot.dat$number <- 1:nrow(boot.dat)

#generate list of student codes to select from
set.seed<-boot
id.list <- tapply(boot.dat$number, INDEX=boot.dat$SchoolID2, sample,
replace=T)
id.list <- unlist(id.list)
boot.data <- boot.dat[id.list, ]
rm(boot.dat)

#Formula in SAP
#KS2past ~ KS1 + REG + GfW + (1|school)
read.model.i <- lmer(KS2_READMRK ~ KS1_WRITPOINTS + Region +
I(Fidelity4CPD/100) + (1| SchoolID2), data=boot.data)
secondary.coeffs.read.fid4.i <- rbind(secondary.coeffs.read.fid4.i,
fixef(read.model.i))

FID4.n.secondary.read.i <- rbind(FID4.n.secondary.read.i,
length(read.model.i@frame$KS1_WRITPOINTS))

varcomp <- data.frame(VarCorr(read.model.i))
#collecting coefficient separately and the two variances
effect.size.read.fid4.i <- rbind(effect.size.read.fid4.i, c(
summary(read.model.i)$coefficients[4,1], varcomp[1,4], varcomp[2,4]
))

rm(read.model.i, boot.data, varcomp)

if (boot%%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap primary outcome

Sys.time() - t1

#Data for table 1

mean(FID4.n.secondary.read.i)
sd(FID4.n.secondary.read.i)

#coefficient data
head(secondary.coeffs.read.fid4.i)
apply(secondary.coeffs.read.fid4.i, MARGIN=2, mean)
apply(secondary.coeffs.read.fid4.i, MARGIN=2, sd)
apply(secondary.coeffs.read.fid4.i, MARGIN=2, quantile, probs=c(.025,
.975))
hist(secondary.coeffs.read.fid4.i[,4])

effect.size.read(effect.size.read.fid4.i)

```



```

#saved all the above:
save.image("L:\\Jan\\001_Analysis\\Include_in_report_MAY2018\\2018_07_07_a1
12_fidelity4CPD.RData")

#####
### Fidelity analysis in 19 intervention schools
### Prepared as support for the process analysis section
### when working on the revised report (08.07.2018)

library(Rcmdr)

#Read in the fidelity ratings from the teams
NewFidelity <- readXL("L:/Jan/000_ORIGINAL/Fidelity scores.xlsx",
  rownames=FALSE, header=TRUE, na="", sheet="Sheet1",
  stringsAsFactors=TRUE)

fix(NewFidelity)
names(NewFidelity)
names(NewFidelity)[1] <- "SchoolID2"
names(NewFidelity)

#test for interrater effect
icc.19 <- lmer(Total ~ 1 + (1|Observer), data=NewFidelity)
summary(icc.19)
#The ICC is at .428, which means that there may be a string clustering
effect
#analysis with control for observer _and_ fixed effects as well

#Louise has confirmed via email that ratings were correctly scored
#i.e. that my data contained the reverse-scored items (8.7.2018)
#to control for team effect, averages for fixed effects analysis:
mean.exe <- mean(NewFidelity[NewFidelity$Observer=="E", ]$Total)
mean.yrk <- mean(NewFidelity[NewFidelity$Observer=="Y", ]$Total)
mean.exe
mean.yrk

#merge with main data set
Nineteen.prim <- merge(GfWdata3.prim, NewFidelity, by="SchoolID2")
length(Nineteen.prim[,1])
summary(Nineteen.prim$Total)
(unique(GfWdata3.prim$SchoolID2))
GfWdata3.prim$SchoolID2

#FIDELITY ANALYSIS, with N=19 schools FULL - NON-IMPURED
model.coeffs.19 <- NULL
effect.size.19 <- NULL
s19.n.primary <- NULL

#define consecutive numbers for bootstrap selection within schools
Nineteen.prim$number <- 1:nrow(Nineteen.prim)

#Actual bootstrap
for (boot in 1:1000) {

#generate list of student codes to select from
set.seed<-boot
id.list <- tapply(Nineteen.prim$number, INDEX=Nineteen.prim$SchoolID2,
  sample, replace=T)
id.list <- unlist(id.list)
boot.data <- Nineteen.prim[id.list, ]

```

```

#Formula in SAP
#KS2past ~ KS1 + REG + GfW + (1|school)
prim.model <- lmer(CalcTotal_Overall12 ~ KS1_WRITPOINTS + Region + Total +
Observer + Observer*Total + (1 | SchoolID2), data=boot.data)
model.coeffs.19 <- rbind(model.coeffs.19, fixef(prim.model))

s19.n.primary <- rbind(s19.n.primary,
length(prim.model@frame$KS1_WRITPOINTS))

varcomp <- data.frame(VarCorr(prim.model))
#collecting coefficient separately and the two variances
effect.size.19 <- rbind(effect.size.19, c(
summary(prim.model)$coefficients[4,1], varcomp[1,4], varcomp[2,4]
))

rm(prim.model, boot.data, varcomp)

if (boot%%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap

mean(s19.n.primary)
sd(s19.n.primary)

#Data for table 1
#coefficient data
apply(model.coeffs.19, MARGIN=2, mean)
apply(model.coeffs.19, MARGIN=2, sd)
apply(model.coeffs.19, MARGIN=2, quantile, probs=c(.025, .975))
hist(model.coeffs.19[,4])

effect.size.read(effect.size.19)

#with fixed effects/ assessment unit centred
#FIDELITY ANALYSIS, with N=19 schools FULL - NON-IMPUTED
#centering
#since centering is done on level2
#the actual averages cannot be exactly 0

split1 <- subset(Nineteen.prim, Observer=="Y")
split2 <- subset(Nineteen.prim, Observer=="E")
nrow(split1)
nrow(split2)

split1$Total.cent <- split1$Total - mean.yrk
mean(split1$Total)
mean(split1$Total.cent)

split2$Total.cent <- split2$Total - mean.exe
mean(split2$Total)
mean(split2$Total.cent)

Nineteen.prim2 <- rbind(split1, split2)
mean(Nineteen.prim2$Total)
mean(Nineteen.prim2$Total.cent)

model.coeffs.19c <- NULL
effect.size.19c <- NULL

```

```

s19c.n.primary <- NULL

#define consecutive numbers for bootstrap selection within schools
Nineteen.prim2$number <- 1:nrow(Nineteen.prim2)

#Actual bootstrap
for (boot in 1:1000) {

#generate list of student codes to select from
set.seed<-boot
id.list <- tapply(Nineteen.prim2$number, INDEX=Nineteen.prim2$SchoolID2,
sample, replace=T)
id.list <- unlist(id.list)
boot.data <- Nineteen.prim2[id.list, ]

#Formula in SAP
#KS2past ~ KS1 + REG + GfW + (1|school)
prim.model <- lmer(CalcTotal_Overall12 ~ KS1_WRITPOINTS + Region +
Total.cent + Observer + (1 | SchoolID2), data=boot.data)
model.coeffs.19c <- rbind(model.coeffs.19c, fixef(prim.model))

s19c.n.primary <- rbind(s19c.n.primary,
length(prim.model@frame$KS1_WRITPOINTS))

varcomp <- data.frame(VarCorr(prim.model))
#collecting coefficient separately and the two variances
effect.size.19c <- rbind(effect.size.19c, c(
summary(prim.model)$coefficients[4,1], varcomp[1,4], varcomp[2,4]
))

rm(prim.model, boot.data, varcomp)

if (boot%%10==0) {
plot(x=1, y=boot, xlim=c(0,2), ylim=c(0, 1000))
} #end of plot if-clause

} # end of bootstrap

mean(s19c.n.primary)
sd(s19c.n.primary)

#Data for table 1
#coefficient data
apply(model.coeffs.19c, MARGIN=2, mean)
apply(model.coeffs.19c, MARGIN=2, sd)
apply(model.coeffs.19c, MARGIN=2, quantile, probs=c(.025, .975))
hist(model.coeffs.19c[,4])

effect.size.read(effect.size.19c)

```

You may re-use this document/publication (not including logos) free of charge in any format or medium, under the terms of the Open Government Licence v3.0.

OGL This information is licensed under the Open Government Licence v3.0. To view this licence, visit <http://www.nationalarchives.gov.uk/doc/open-government-licence/>

Where we have identified any third party copyright information you will need to obtain permission from the copyright holders concerned. The views expressed in this report are the authors' and do not necessarily reflect those of the Department for Education.

This document is available for download at www.educationendowmentfoundation.org.uk



Education
Endowment
Foundation

The Education Endowment Foundation
5th Floor, Millbank Tower
21–24 Millbank
London
SW1P 4QP
www.educationendowmentfoundation.org.uk