# Statistical Analysis Plan for Grammar for Writing

## University of York

Education Endowment Foundation

| | |
|---|---|
| **INTERVENTION** | **Grammar for Writing** |
| **DEVELOPER** | University of Exeter |
| **EVALUATOR** | University of York |
| **TRIAL REGISTRATION NUMBER** | ISRCTN 83236864 |
| **TRIAL STATISTICIAN** | Jan R. Böhnke, Dundee Centre for Health and Related Research, School of Nursing and Health Sciences, University of Dundee, Dundee, j.r.boehnke@dundee.ac.uk |
| **TRIAL CHIEF INVESTIGATOR** | Louise Tracey, Department of Education, University of York |
| **SAP AUTHOR** | Jan R. Böhnke, Dundee Centre for Health and Related Research, School of Nursing and Health Sciences, University of Dundee, Dundee, j.r.boehnke@dundee.ac.uk |
| **SAP VERSION** | 2.0 |
| **SAP VERSION DATE** | 11th September 2017 |
| **EEF DATE OF APPROVAL** | 7th November 2017 |
| **DEVELOPER DATE OF APPROVAL** | 21st September 2017 |

## Table of contents

## Introduction

The 'Grammar for Writing' programme draws on the concept of improving children's grammar in parallel with their writing by using a contextual approach. It is a way of teaching writing that assumes that rather than teaching grammatical rules in the abstract, teachers should help students to understand how linguistic structures convey meaning. Consequently, the programme aims to improve writing by developing students' understanding of grammatical choices. Underpinned by key pedagogical principles, Grammar for Writing is embedded in the context of teaching about writing genres. The core elements of the programme encompass the use of grammar terms, linking grammar effects in writing, and using talk to develop discussion about choices and effects. The programme is designed to be delivered by teachers as standalone units of work or as a series of units within a whole class setting. Each unit is around 4 weeks' worth of work.

Whilst the concept behind Grammar for Writing is promising, it currently lacks conclusive evidence in the primary school phase. There have been several, developer-led trials in secondary schools which have demonstrated positive results (Jones, Myhill, & Bailey, 2013; Jones et al., 2013). An efficacy trial funded by the Education Endowment Foundation (Torgerson et al., 2014) looked at whole-class and small group delivery in a 4-week version of the programme adapted for Year 6 after Key Stage 2 SaTs assessments. However, this found only limited effects measured by children's performance on the GL Progress in English assessment. For the whole-class intervention, there was a small and statistically non-significant effect (*ES* = 0.1). The impact for those additionally taught Grammar for Writing in small groups was slightly higher than for those taught in small groups without Grammar for Writing (*ES* = 0.24).

For the purposes of the current trial, two units of work (narrative writing and persuasive writing) have been delivered within a whole-class context during Year 6. There were 4 days of CPD provided to teachers through the school year, co-delivered by the University of Exeter and Babcock LDP, an education support and improvement service which provides training within the school sector. This CPD included the provision of teaching materials for the two units of work.

The primary research question is:

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

A secondary research question asks whether or not Grammar for Writing impacts on other literacy outcomes for Year 6 pupils. Finally, given that the Grammar for Writing programme aims to increase teachers' grammar knowledge and subsequently increase students' literacy outcomes, a mediation hypothesis relating to the impact of Grammar for Writing on teacher knowledge and said grammar knowledge on student outcomes will be tested.

## Study design

### Population including eligibility criteria

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 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 students) were kept to a minimum by deliberately targeting larger schools for recruitment.[1] Half of the schools were recruited from the North

[1] At time of writing only one school had 18 students; all other schools had more than 20.

East[2] and the other half from across the rest of England. There was a high proportion of disadvantaged schools.[3]

### Sample size

The aim of the recruitment was to have 150 schools participate in the study. This total was determined by power calculations (see below). The unit for randomisation was schools participating in this study, such that there would be 75 treated schools and 75 control schools.

Recruitment was conducted by the developer team (University of Exeter), with support from the evaluation team (University of York). For pragmatic reasons specific regions were targeted in addition to the North-East: the North West, South West and London.[4] A primarily dual approach was then adopted. Firstly, all schools in the local authorities in those target areas were systematically identified and approached. Secondly, 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. Finally a number of untargeted, opportunistic approaches were made using social media (ie. Twitter and Facebook). One thousand five hundred and seventy-one schools (1,571) were approached by the developer team to participate in the study. One hundred and ninety-five schools (195) expressed an interest in taking part. Of these, 155 schools were recruited to the study and randomised. This resulted in a total of 77 schools assigned to treatment, and 78 to control. Overall, $N$ = 312 teachers were recruited, an average of 2 teachers per school. There were 144 teachers in control schools (1.9 average per school) and 168 teachers in intervention schools (average 2.2 per school).

### Description of trial design

This study is a two-arm effectiveness RCT with randomisation occurring at the school-level to reduce the possibilities of contamination that could occur when using a within-school design.

*Control schools*

Control schools were informed of their allocation and requested to continue their teaching as usual. Control schools will receive £500 on completion of all requested measures at the end of the intervention period (July 2017). This payment can then be used towards funding Grammar for Writing training if desired. Head teachers and teachers participating in the study in control schools all received two newsletters during the study keeping them informed about the evaluation and the training they could choose to take at the end of the trial (December 2016, May 2017). Control schools were also contacted to organise the end of year writing assessments. A small sub-sample ($N$ = 5) were contacted to organise a classroom observation and Year 6 teacher and literacy co-ordinator interviews as part of the process evaluation of this study and all these schools agreed (see protocol).

*Treatment schools*

All Year 6 teachers in schools that were selected into the treatment group were expected to undertake three days of CPD, with a fourth day offered to schools on future planning for Year 6 writing using Grammar for Writing programme principles. Teachers received the four days of CPD in October 2016, November 2016, March 2017 and May 2017. The training was delivered by the University of Exeter and Babcock LDP. The CPD included the provision of teaching materials for the two units of work. One hundred and twenty-four of the 158 teachers

---

[2] 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)

[3] As defined in the National Student Database, 105 out of the 155 recruited schools had FSM on average at 29% or over as reported by the schools themselves or taken from the Schools Directory 2016/17.

[4] As defined by the former designated Government Office Regions. (http://webarchive.nationalarchives.gov.uk/20080728115009/http://www.dcsf.gov.uk/rsgateway/region1.shtml)

in the intervention schools (77%) attended all of the first three CPD days. Head teachers and Year 6 teachers in the intervention schools received two newsletters during the study keeping them informed about the evaluation (December 2016, May 2017). Intervention schools were also contacted to organise the end of year writing assessments. A small sub-sample ($N$ = 10) were contacted to organise a classroom observation and Year 6 teacher and literacy co-ordinator interviews as part of the process evaluation of this study and all these schools agreed.

For the purposes of this trial, two units of work were delivered within a whole-class context: narrative writing and persuasive writing. The two units of work were designed to be delivered to classes after the associated CPD days 2 and 3. The first unit, on narrative writing, consisted of a series of hour-long lessons to be delivered daily over a four-week period in the Spring Term. The second unit, on persuasive writing, consisted of two-weeks' worth of lessons to be delivered in the Summer Term 2017.

Intervention schools received the Grammar for Writing programme, materials and training at the reduced rate of £500. The amount paid to both control and intervention schools reflects the fact that the burden placed on schools is not high and it avoids potential ethical problems if the intervention is shown not to be effective. In addition teachers will receive an extra payment of £20 in vouchers in exchange for completing the pre- and post-intervention on-line surveys (see below).

As described in more detail below, two different measurement schedules were used, both consisting of one baseline and one follow-up assessment:

- For students, individual KS1 results will be collated from the NPD.
- The independent writing assessment (primary outcome, see below) and KS2 results (secondary outcome, see below) collected after the delivery of the second Grammar for Writing unit will be used as the primary and secondary outcomes.
- For all teachers (i.e. both groups) a grammar quiz was assessed in June 2016 through to October 2016 to gather baseline data on teachers' grammar skills and a follow-up assessment was performed in June/July 2017. Table 1 provides an overview of both measurement schedules and milestones for the study.

*Table 1*. Milestones and assessments scheduled for the Grammar for Writing trial

| Date | Measure |
|---|---|
| May/ June 2012 | KS1 assessments (from NPD) |
| June-October 2016* | Teacher baseline survey (online collection by evaluation team) |
| July-October 2016* | Randomisation |
| October 2016 | CPD 1 |
| November 2016 | CPD 2 |
| Spring Term 2017 | Delivery of Unit 1 (narrative writing) |
| March 2017 | CPD 3 |
| Summer Term 2017 | Delivery of Unit 2 (persuasive writing) |
| May 2017 | KS2 assessments (available from NPD ~October 2017) |
| May 2017 | CPD 4 (after KS2 assessments) |
| June 2017 | Writing assessments |
| June 2017 | Teacher post-test survey |

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

## Protocol changes

No changes to the protocol were made after acceptance of an amended protocol on the 17th January 2017.

## Randomisation

Schools were only eligible for randomisation after their head teacher signed the Memorandum of Understanding; written consent by teachers had been obtained up front; and when pre-test data requested in the Memorandum of Understanding was provided (including student UPNs, teacher contact details, and completion of teacher pre-intervention survey). A request to complete the teacher survey was sent to teachers after consent was obtained although the requirement to complete a survey was specified in that consent.

Since the recruitment phase went on for longer than expected the randomisation was performed by Louise Elliott (York) in the months July 2016 to October 2016. Schools were stratified by region (North-East/not-North-East) and then randomised using minimisation. Minimisation uses algorithms to minimise imbalance at baseline in expectation and permits ongoing allocation, so schools can be randomised and informed of their allocation soon after recruitment. Randomisation was conducted and recorded using MinimPy software (Saghaei & Saghaei, 2011; v3.0; default settings).

Schools were stratified by region (North East / not-North East) and were then randomised in six batches:

- 15 July 2016, 36 schools randomised
- 17 August 2016, 34 schools
- 12 September 2016, 31 schools
- 22 September 2016, 32 schools
- 30 September 2016, 19 schools
- 5 October 2016, 3 schools.

In total there are 77 schools allocated to treatment, and 78 schools allocated to control or treatment as usual. No school has withdrawn consent to include their data in the study.

## Calculation of sample size

The statistical power of the proposed analyses was estimated using the formula provided as a standard by the EEF[5].

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

The following assumptions made were:

- Students per school per class: 25 (i.e. $n$ = 50 per treatment per school)
- Between-school pre-post correlation (squared): $R_1^2 = 0.53$
- Intraclass correlation: $\rho = 0.15$
- Criterion for statistical significance: $p < .05$ and statistical power: 0.80 (consequently $M_{J\text{-}K} = 2.85$)

This would result in a *MDES* = .18. Further, we would expect stratification variables to explain some of the variance (Explained variance between schools $R_2^2 = .10$), which would lead to a *MDES* = .17.

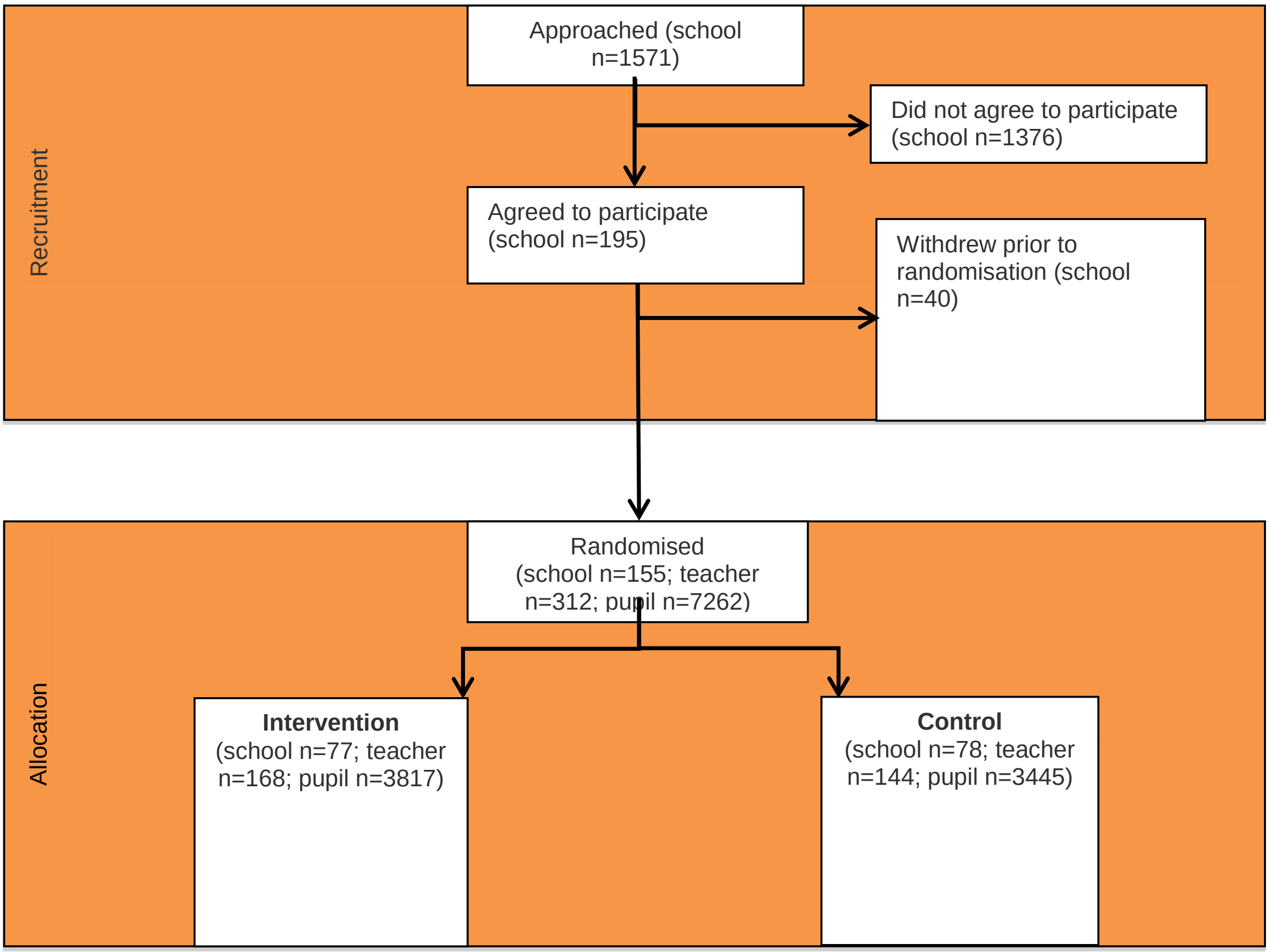
Assuming 16 FSM students per school, this sample of 150 schools would enable an effect size of *MDES* = 0.18 (with stratification *MDES* = .17) to be detected in the FSM sub-sample (defined by NPD EVERFSM).

[5] Education Endowment Foundation, 31.10.2013, Pretesting in EEF Evaluations.

## Follow-up

At the stage of writing no definite follow-up and flow information is available beyond the one presented above. Figure 1 presents the CONSORT flowchart based on available numbers. Numbers presented on students may at this stage include opt-outs since definite information was not available. No information on follow-up is available at the moment, but 20 treatment schools have not tested any students at follow up for the primary outcome. These schools did not withdraw from the NPD data collection.

**Figure 1: CONSORT flowchart for the Grammar for Writing trial (as of 26. July 2017)**




## Outcome measures

### *Primary outcome*

The primary outcome will be the combined results of two tasks selected from past Key Stage 2 ($KS2_{past}$) writing assessments which were in use pre-2013. These past Key Stage 2 assessments have been chosen because the current writing assessment for KS2 consists of a portfolio of teacher-assessed work which, whilst externally moderated, is judged to be 'working toward', 'working at' or 'working above' the expected standard for the end of Key Stage 2. The advantage of using past KS2 writing tasks is that they can be administered in controlled conditions within schools and have a set marking scheme which is sufficiently graded 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 longer written task and one shorter written task, covering both persuasive and narrative writing. The developer team will remain blind to their exact content. The assessments will be administered in schools independently by the National Foundation for Educational Research (NfER) to ensure controlled conditions and to reduce any burden on schools. They will be marked by a team of experienced assessors, unaware of allocation, at the University of York. All assessors will receive training from the University of York and the marking will be moderated. As narrative and persuasive writing are implicit within the KS2 curriculum, this primary measure will not be inherent to treatment – teachers in the control condition will also be teaching their students to write narratively and persuasively.

The result of the task is scored between 0 (which means none of the criteria for the lowest scoring band has been met for the assessment focuses: (1) ‘sentence structure and punctuation’; (2) ‘text structure and organisation’; and (3) ‘composition and effect’) and 40 (which means all the criteria for each of the three assessment criteria have been met to a high standard). The fourth assessment focus, ‘handwriting’ for which 0-3 marks could be obtained will not be included in the outcome scoring as this is not a focus of the programme.

### *Secondary outcomes*

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 include KS2 scores on each element of literacy (writing; reading; grammar, punctuation and spelling[6]) separately and together for the same students. Using primarily nationally collected data minimises cost and the burden on schools and students. 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) will be focused on ensuring that students succeed on them. With the addition of the past KS2 writing tasks, the proposed outcome measures will provide a measure of all-round performance on literacy, and, specifically, the legacy effect on writing.

The Key Stage raw scores will be 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. 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’.

### *Other measures*

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

An intermediary measure will be the teacher ‘grammar quiz’ developed for inclusion in the pre- and post-test teacher survey. The pre-test ‘quiz’ was developed by the developer team for use as part of the ‘Grammar for Writing’ training. 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 by the evaluation team at York. The intermediate quiz will result in a score

[6] The currently confirmed NPD read-out variables for this will be KS2_READMRK for reading; KS2_GPSMRK as a score for grammar, punctuation and spelling; and WRITTAOUTCOME as the teacher assessed writing score.

between 0 (no task correct) and 30 (all tasks correct); the final score ranging from 0 to 29, respectively.

The pre- and post-test teacher surveys were delivered online using Qualtrics (Qualtrics, Provo, UT). In addition to the grammar quizzes for statistical analysis, the surveys gathered data regarding teachers' professional and academic backgrounds, linguistic subject knowledge, confidence in teaching literacy, specifically grammar and writing, any schemes of work used in their literacy teaching and contextual classroom factors, which will be used to describe the sample of teachers in greater detail (a full descriptive table will be provided in the appendix to the report).

Emails and reminder phone calls are used to encourage completion and teachers will receive an extra payment of £20 in vouchers in exchange for completing the pre- and post-intervention on-line surveys.

## Analysis

All analyses will be conducted by Jan R. Böhnke (JRB, Dundee) who will be blind to group identity. The evaluation team in York will liaise with EEF to ensure that no data from the teacher survey will be shared with JRB which could disclose group allocation indirectly (e.g., days of CPD attended, evaluation data concerning CPD etc.).

All data will be presented descriptively with means, standard deviations, and medians for quantitative outcomes and category frequencies for categorical data (see Planned Table 3). All continuous variables will be grand mean centred. The analysis will be conducted for all students providing at least demographic information at baseline. All statistical analyses will be reported for complete cases as well as corrected for missing data and drop-outs (which we expect to be low in relative frequency for all NPD data we use, but can happen on our primary outcome measure, which is not drawn from the NPD). Bootstrapped confidence intervals are used to judge the statistical significance of the intervention effect.

### *Primary intention-to-treat (ITT) analysis*

The impact evaluation will use Hierarchical Linear Models (HLM), a mixed effects model in which students are nested within schools. This makes it possible to separate within-school variation in the outcome from between-school variation. The analysis will be intent-to-treat, which means that schools will be treated according to the condition they were allocated (control or Grammar for Writing), not which they actually received.

This study was planned for a single primary outcome, the writing assessment developed by the team from previously used Key Stage 2 assessments ($KS2_{past}$) and to answer the question 'how effective Grammar for Writing is in improving the writing skills in Year 6 students?' In accordance with the power analysis, pre-test data from the Key Stage 1 (KS1) writing results will be used as a student-level covariate (*KS1*) without random variation across schools. An individual student *i*'s $KS2_{past}$ result in a specific school will be modelled as depending on school *j*'s average $KS2_{past}$ attainment (random school-level intercept; $\mu_{0j}$) and a random error term ($\varepsilon_{ij}$). Each school's average $KS2_{past}$ performance ($\mu_{0j}$) will be predicted by an overall intercept (average performance; $\gamma_{00}$), ; 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):

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

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

$$\mu_{1j} = \gamma_{10} \tag{3}$$

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

```
KS2past ~ KS1 + REG + GfW + (1|school)
```

The intervention will be evaluated as having shown an effect in this trial when the average bootstrapped point estimate for the coefficient of the intervention effect $(\gamma_{02})$ is positive (i.e. on average intervention schools achieve higher scores on $KS2_{\text{past}}$) and the 95%-bootstrap confidence interval of this coefficient does not include 0.

The analysis will be cluster-bootstrapped as applied in previous projects (Hanley, Böhnke, Slavin, Elliott, & Croudace, 2016; Huang, in press): From each school a random sample of the same size as its actual sample is drawn (with replacement) and across these school-wise bootstrap samples, the mixed model is then estimated.[7] This process is repeated $b = 1000$ times and for a 95%-confidence interval the statistical estimates (here the $\gamma_{03}$ values) are saved and their top and bottom 2.5%-quantiles are identified. The average of the bootstrapped values will be treated as the point estimate and will be reported in all coefficient tables. No *p*-values will be reported for any analysis.

### Imbalance at baseline

EverFSM and KS1 results will be presented with means, standard deviations, and medians for KS1 and category frequencies for EverFSM. Imbalance will be judged to be present if the KS1 standardised mean difference ≥ .10 (KS1); when the standardised differences of proportions $w \geq .05$ (Faul et al., 2007) for the cross-tabulation of intervention group and FSM, respectively. If imbalance is detected for EverFSM it will be included as student-level covariate in all outcome analyses (KS1 is included in all analyses as per power analysis).

### Missing data

As already clarified, we do not expect large amounts of missing data in this study, since nearly all data used is drawn from the NPD. Nevertheless, the primary outcome measure is prone to some drop-out since it is independently collected. 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 students 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. For the ITT analysis of the primary outcome we will use multiple imputation by chained equations (MICE; Azur, Stuart, Frangakis, & Leaf, 2011) to impute missing values.

As with other imputation techniques, MICE 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) will be used for

[7] 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].

this analytic step and in this specific case, the following variables will be entered into the algorithm:

- Gender, EverFSM, and the KS1 result ("baseline data"; independent of whether they have or don't have missing data);
- The primary and secondary outcome variables ("follow-up"; which are 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 code whether baseline data is missing (yes/ no) or only follow-up data (yes/ no; see below; no missing data since coded from available missingness patterns; see below).

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

MICE does not define a specific model for the missingness mechanism, which is why it is not in all cases seen as preferable where details about missingness processes are available (especially in longitudinal studies). But 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 will condition predictions of the MICE procedure on whether any data for a respondent is missing at baseline (i.e. some problem with NPD data retrieval or documentation) or whether any data is missing at follow-up (not shown up to primary outcome assessment or NPD retrieval problem for secondary outcomes).

### *Non-compliance with intervention*

In this study only a single on-treatment analysis will be performed: Schools and teachers will be allocated to the group according to their factual post-hoc participation in the study. While in the ITT analysis schools and teachers are allocated according to the randomisation result to either treatment or control group, this analysis will allocate them according to actual participation status:

- Schools will be allocated according to the group they were actually part of (if any incorrectly allocated);
- Teachers will be scored according to the number of CPD days they attended.

The treatment assignments developed from both coding procedures are then used instead of the ITT treatment allocation and the analyses for primary and secondary outcomes will be re-run for each of these variables.

The analysis of the potential mediating effect of grammar knowledge (see below) is more appropriate to test one of the key instrumental hypotheses of the intervention. And the subgroup analysis for high vs. low implementation fidelity more appropriate to test the potential moderating effect of implementing Grammar for Writing to differing degrees of quality.

### *Secondary outcome analyses*

The analyses of the secondary outcomes (see footnote 6) investigate whether or not Grammar for Writing impacts on other literacy outcomes. The analytic approach will use exactly the same procedure and model as for the primary outcome, with the only difference that instead of $KS2_{past}$ the secondary outcome variables will be used as dependent variables. The intervention will be evaluated as having shown a potential effect on a secondary outcome when the 95%-bootstrap confidence interval of the coefficient ($\gamma_{02}$; see formula 2 above) does 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 research.

### *Additional analyses*

The only additional analysis concerns the link between teachers' grammar knowledge and programme impact, testing the third hypothesis of the study. Grammar for Writing should increase teachers' grammar knowledge and subsequently increase students' literacy outcomes. Consequently a mediation hypothesis relating to the impact of Grammar for Writing on teacher knowledge and said grammar knowledge on student outcomes will be tested.

The measure of teachers' ($N$ = 312) grammar knowledge is their performance in the second grammar quiz at the end of the intervention. The scores of this quiz will be reported (Mean, Median, SD), including Cronbach-$\alpha$ and the pre-post correlation in the control group as reliability estimates. The scores will be compared using a bootstrapped $t$-test across the two intervention groups. If the bootstrapped 95%-confidence interval of the bootstrapped $t$-values does not include 0 and the average $t$-value is positive (indicating higher attainment in the group of teachers who received the intervention), Grammar for Writing will be evaluated as having shown a potential effect on teachers' grammar knowledge.[8]

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 will be 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).

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

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

$$\mu_{1j} = \gamma_{10} \tag{6}$$

$$\mu_{2j} = \gamma_{20} + u_{20} \tag{7}$$

A potential mediation effect would be detected if the bootstrapped 95%-confidence interval of the product of the coefficients $\mu_{2j}$ and $\gamma_{20}$ 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.

[8] Further analyses on the data will be conducted to evaluate the validity of the grammar quiz. This will entail 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. And factor and Rasch Model analyses will be conducted to gauge the plausibility of both quizzes representing the same trait. These analyses are post-hoc evaluations of how well the measure performed and will form appendices to the full report.

### *Subgroup analyses*

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

1. students eligible for FSM;
2. boys and girls;
3. high and low achievers on the pre-test (KS1; median-split based on all observed scores); and
4. high and low implementation fidelity within treatment schools.[9]

The multilevel model described for the primary outcome will be 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 will 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 $KS2_{past}$ result in a specific school will be modelled as depending on school *j*'s average $KS2_{past}$ attainment (random school-level intercept; $\mu_{0j}$), previous attainment (*KS1*), and a random error term ($\varepsilon_{ij}$). For the test for subgroup effects, a coefficient for one of the student-level variables described above is added (*Subgroup*) as a random slope. Each school's average $KS2_{past}$ performance ($\mu_{0j}$) will be predicted by an overall intercept (average performance; $\gamma_{00}$); 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:

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

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

$$\mu_{1j} = \gamma_{10} \tag{10}$$

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

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

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

[9] The teacher survey assessed several variables that are available to proxy the degree to which teachers implemented the programme as designed (beyond analyses regarding teachers' programme dosage mentioned above). For example approximately three quarters of teachers stated that they made changes to the programme. The survey also asks for what changes teachers made. These changes will be classified as programme-conform vs. non-conform by the evaluation team with advice from the developer team. This assessment will be turned into an individual score for each teacher (0 = no or only conform changes; +1 per non-conform change) and these scores will be used to classify teachers per median split.

The intervention will be evaluated as having shown a potential interaction with the specified subgroup variable when the 95%-bootstrap confidence interval of ($\gamma_{21}$; formula 11) does not include 0.

Only when this effect is found to be statistically significant will more detailed reporting on subgroup statistics be done (means, SDs). The exception is FSM for which details will be reported anyway.

No subgroup analyses will be performed that have not been defined in the protocol. The analysis for high vs. low fidelity will be performed by the team in York, since it would un-blind the main analyst to the intervention allocation.

### *Effect size calculation*

Effect sizes will be calculated based on the total variance in the models. For two-level models (see definition of error terms above):

$$\boldsymbol{ES} = \frac{\boldsymbol{Effect}}{\sqrt{\boldsymbol{u_{00}} + \boldsymbol{\varepsilon_{ij}}}}$$

Confidence intervals will be bootstrapped. Here, *Effect* is either a difference in group means for models without covariates; or the coefficient from the estimated model (e.g., $\boldsymbol{\gamma_{02}}$ in the analysis of the primary outcome; formula 2). Both effect sizes will be reported for any model-based inference.

## Report tables

The following tables are planned to be included in the main report. A technical appendix will cover all detailed model results, including estimated coefficients for the models that were run and intermediate steps.

**Planned Table 1: Summary of impact on primary outcome**

| Group | Effect size (95% confidence interval) | Estimated months' progress | EEF security rating | EEF cost rating |
|---|---|---|---|---|
| **GfW vs. control** | | | | |
| **GfW FSM vs. control** | | | | |

**Planned Table 2: Minimum detectable effect size at different stages**

| Stage | N [schools/ students] (n=intervention ; n=control) | Correlation between pre-test (+other covariates) & post-test | ICC | Blocking/ stratification or pair matching | Power | Alpha | Minimum detectable effect size (MDES) |
|---|---|---|---|---|---|---|---|
| **Protocol** | | | | | | | |
| **Randomisation** | | | | | | | |
| **Analysis (i.e. available pre- and post-test)** | | | | | | | |
| **FSM-only** | | | | | | | |

**Planned Table 3: Baseline comparison**

| Variable | Intervention group | | Control group | |
|---|---|---|---|---|
| School-level (categorical) | n/N (missing) | Percentage | n/N (missing) | Percentage |
| e.g. Academy | | | | |
| e.g. Ofsted rating<br>Outstanding<br>Good | | | | |
| | | | | … |
| School-level (continuous) | n (missing) | Mean/ median | n (missing) | Mean/ median |
| Number of Y6 students | | | | |
| Average KS1 result | | | | |
| Fidelity of implementation Measure | | | -- | -- |
| Percentage of Students with at least one missing value | | | | |
| Teacher level (continuous) | | | | |
| Grammar quiz, pre-test, raw data | | | | |
| Grammar quiz, post-test, raw data | | | | |
| Grammar quiz, pre-test, ranks | | | | |
| Grammar quiz, pre-test, ranks | | | | |
| Grammar quiz, differences in ranks | | | | |
| Student-level (categorical) | n/N (missing) | Percentage | n/N (missing) | Percentage |
| Eligible for FSM | | | | |
| Gender | | | | |
| WRITTAOUTCOME<br>working towards<br>working at<br>working above’ | | | | |
| Student-level (continuous) | n (missing) | [Mean or median] | n (missing) | [Mean or median] |
| KS2past | | | | |
| KS1 Result | | | | |
| KS2_READMRK | | | | |
| KS2_GPSMRK | | | | |

**Planned Table 4: Primary analysis; unadjusted and adjusted effect size estimates**

| | Raw means | | | | Effect size | | |
|---|---|---|---|---|---|---|---|
| | Intervention group | | Control group | | | | |
| Outcome | n (missing) | Mean (95% CI) | n (missing) | Mean (95% CI) | n in model (intervention; control) | Unadjusted Hedges g (95% CI) | Adjusted effect based on analytic model (95% CI) |
| ***KS2**$_{past}$* | | | | | | | |
| ***KS2**$_{past}$*, **imputed data** | | | | | | | |

**Planned Table 5: Secondary analysis; unadjusted and adjusted effect size estimates**

| | Raw means | | | | Effect size | | |
|---|---|---|---|---|---|---|---|
| | Intervention group | | Control group | | | | |
| Outcome | n (missing) | Mean (95% CI) | n (missing) | Mean (95% CI) | n in model (intervention; control) | Unadjusted Hedges g (95% CI) | Adjusted effect based on analytic model (95% CI) |
| **WRITTAOUTCOME** | | | | | | | |
| **WRITTAOUTCOME, imputed data** | | | | | | | |
| **KS2_READMRK** | | | | | | | |
| **KS2_READMRK, imputed data** | | | | | | | |
| **KS2_GPSMRK** | | | | | | | |
| **KS2_GPSMRK, imputed data** | | | | | | | |

**Planned Table 6: Subgroup analysis; unadjusted and adjusted effect size estimates**

| | Raw means | | | | Effect size | | |
|---|---|---|---|---|---|---|---|
| | Intervention group | | Control group | | | | |
| Outcome | n (missing) | Mean (95% CI) | n (missing) | Mean (95% CI) | n in model (intervention; control) | Unadjusted Hedges g (95% CI) | Adjusted effect based on analytic model (95% CI) |
| ***FSM-only*** | | | | | | | |
| ***KS2***$_{past}$ | | | | | | | |
| ***KS2***$_{past}$**, imputed data** | | | | | | | |
| | | | | | | | |
| ***Female students*** | | | | | | | |
| ***KS2***$_{past}$ | | | | | | | |
| ***KS2***$_{past}$**, imputed data** | | | | | | | |
| | | | | | | | |
| ***Male students*** | | | | | | | |
| ***KS2***$_{past}$ | | | | | | | |
| ***KS2***$_{past}$**, imputed data** | | | | | | | |
| | | | | | | | |
| ***KS1, upper 50%*** | | | | | | | |
| ***KS2***$_{past}$ | | | | | | | |
| ***KS2***$_{past}$**, imputed data** | | | | | | | |
| | | | | | | | |
| ***KS1, lower 50%*** | | | | | | | |
| ***KS2***$_{past}$ | | | | | | | |
| ***KS2***$_{past}$**, imputed data** | | | | | | | |
| | | | | | | | |
| ***Low intervention fidelity schools*** | | | | | | | |
| ***KS2***$_{past}$ | | | | | | | |
| ***KS2***$_{past}$**, imputed data** | | | | | | | |
| | | | | | | | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| ***High intervention fidelity schools*** | | | | | | | |
| ***KS2***$_{past}$ | | | | | | | |
| ***KS2***$_{past}$**, imputed data** | | | | | | | |

## 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. https://doi.org/10.1002/mpr.329

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

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.

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, 175-191.

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/

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.

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. (in press). Using Cluster Bootstrapping to Analyze Nested Data With a Few Clusters. *Educational and Psychological Measurement*, *0*(0), 0013164416678980.

Jones, S., Myhill, D., & Bailey, T. (2013). Grammar for writing? An investigation of the effects of contextualised grammar teaching on students' writing. *Reading and Writing*, *26*(8), 1241–1263. https://doi.org/10.1007/s11145-012-9416-1

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.

Li, T., Hutfless, S., Scharfstein, D. O., Daniels, M. J., Hogan, J. W., Little, R. J. A., … Dickersin, K. (2014). Standards should be applied in the prevention and handling of missing data for patient-centered outcomes research: a systematic review and expert consensus. *Journal of Clinical Epidemiology*, *67*(1), 15–32. https://doi.org/10.1016/j.jclinepi.2013.08.013

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.

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

Raudenbush, S. W. et al. (2011). *Optimal Design Software for Multi-level and Longitudinal Research (Version 3.01)*. Retrieved from www.wtgrantfoundation.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. (2007). Model selection and model averaging after multiple imputation. Computational Statistics & Data Analysis, 71, 758-770.

Skrondal, A., & Rabe-Hesketh, S. (2005). *Multilevel longitudinal modeling using Stata*. College Station, TX: Stata Press.

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/evaluation/projects/lets-think-secondary-science/