Grammar and Writing Evaluation Protocol

Jake Anders, Dominic Wyse, Bas Aarts, Julie Dockrell, Carole Torgerson, Sue Sing, Yana Manyukhina, Ian Cushing

Intervention

Englicious is an approach to grammar teaching underpinned by linguistics research that is supported by an extensive set of website resources (www.englicious.org). The approach aims to make learning about grammar fun and appealing, and stimulates pupils to learn about grammar in a hands-on way. For teachers the website provides a wide variety of innovative teaching materials, including lesson plans, interactive exercises, projects, videos, a glossary, etc., as well as background materials to improve their understanding of grammar. It helps teachers deliver the National Curriculum requirements for English grammar, and to prepare their pupils for the Grammar, Punctuation and Spelling tests which are optional at KS1 and statutory at KS2. What makes *Englicious* unique is that it is informed by modern linguistics (Aarts 2011, Aarts, Mehl and Wallis, 2016; Aarts and Smith-Dennis, 2018), and makes full use of digital technologies such as tablets, apps and interactive whiteboards. Currently over 6,500 teachers have signed up to use *Englicious*. The resources on *Englicious* are tailored for particular year groups and address specific grammatical topics. For example, in order to teach pupils that adverbs (part of the KS1 National Curriculum specification for Year 2) can be moved around in sentences *Englicious* offers a lesson plan with an associated interactive activity that teaches the idea of adverb mobility in a playful way. (See http://bit.ly/2sP9DaQ and the link to the activity.)

The *Englicious* approach is being fully manualised in order to be the intervention to be tested in the Grammar and Writing Research Project (GWRP). In order to manualise the intervention the team will use and adapt lesson plans and exercise materials that are already currently available on the website, cross-referring to the specifications for Year 1 and Year 2 pupils in England's National Curriculum programmes of study. Where necessary new materials and lesson plans will be developed.

Significance

Learning to read and write well is essential for all areas of the primary school curriculum, but also to access secondary-level learning. Not acquiring reading and writing to age-appropriate levels can have profound impact on life-chances, including lifetime earnings (The Government Office for Science, 2008). A fundamental aspect of becoming literate is mastery of the grammar of language, and particularly the ability to form grammatically conventional sentences in writing. Understanding and effectively using sentences is not only essential for pupils' progression in writing, but is also the most important focus for the study of grammar as part of linguistics. Grammar is of renewed interest in England because the National Curriculum implemented from 2014 requires much more attention to the teaching of grammar in primary schools than previous national curricula in England.

Worldwide there are a small number of experimental trials that have investigated grammar teaching to improve writing in the primary phase (for overviews see Graham et al., 2012; Wyse & Torgerson, 2017). In England, only one robust experimental trial that focused on grammar for writing at primary school level has been conducted (Torgerson et al. 2014), albeit with junior pupils. Worldwide, no trials have been carried out with infant pupils (ages 6-7) who, in England, are nevertheless required to be taught grammar, punctuation and spelling as part of the National Curriculum. This grammar teaching is specified in the 'writing – vocabulary, grammar and punctuation' statutory requirements of the National Curriculum which require year two pupils (age 6-7) to write compositions that use standard English grammar and punctuation accurately, and understand key grammatical concepts as specified in 'English Appendix 2'. We propose the world's first research on how six-year-old

to seven-year-old children's writing might be improved through teaching about grammar using innovative and engaging web-based resources. New knowledge is needed about the teaching of grammar for writing, to inform teaching of the National Curriculum now and for future developments in national curricula.

Existing research on children's written text production reveals a range of interventions that have been devised to support children's writing. These interventions span word, sentence and text level skills. Recent meta analyses (Koster, Tribushinina, De Jong & Van den Bergh, 2015; Graham, McKeown, Kiuhara, & Harris, 2012) found positive effects for 'sentence combining' and negative effects for explicitly teaching grammar (see also Graham & Perin, 2007), leading researchers to conclude that traditional grammar instruction that is focused on developing extensive meta-linguistic knowledge about grammatical structure and rules is not a means to improve writing (Troia, 2014). However, positive effects for the teaching of grammar have been found in both single case meta analyses (Rogers & Graham, 2008) and studies with older pupils in England (Myhill, Jones, Lines & Watson, 2012). The failure to examine the impact of writing interventions during the initial stages when children learn to write is a significant limitation in understanding the efficacy of instruction in grammar. The ability to combine sentences that are syntactically correct develops at around the age of 7 when both knowledge of grammar and sentence combining are important predictors of writing proficiency (Berninger et al., 2011). Initial stages in the development of written text production offer the opportunity to examine the ways in which children's written grammatical competence can be developed.

Aims

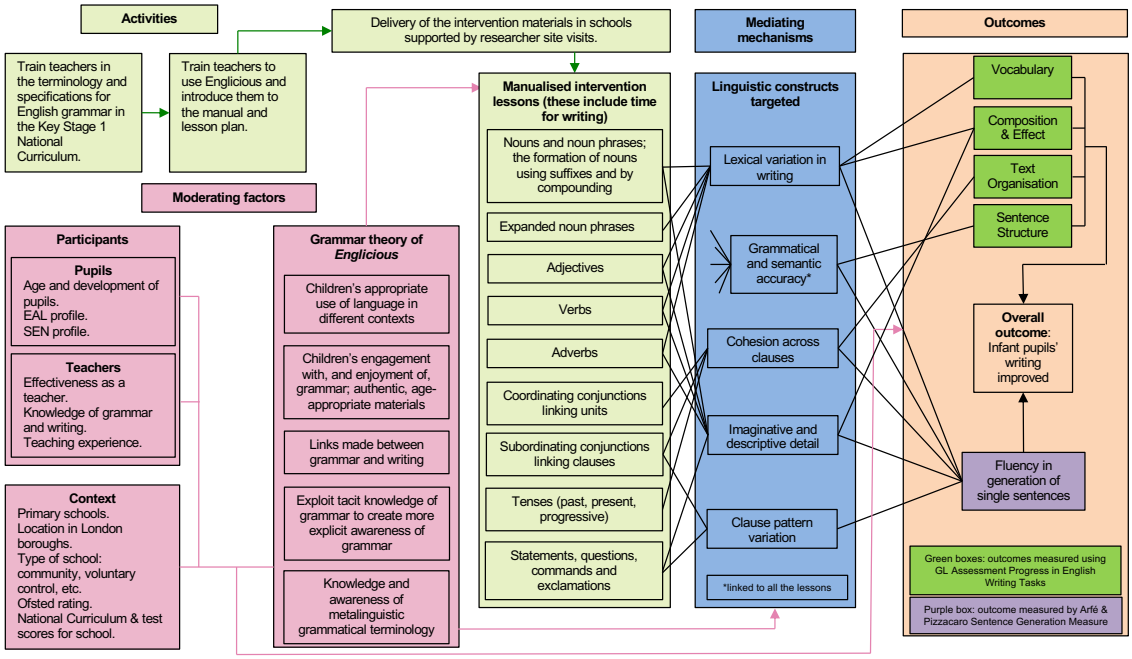
This GWRP will evaluate the effectiveness of the manualised *Englicious* approach to teaching grammar with the aim of improving pupils' writing. *Englicious* is an innovative web-based approach to grammar teaching. It has not yet been subject to an experimental trial of its effectiveness. We propose a randomised controlled efficacy trial as the most robust and appropriate method to test the effectiveness of this teaching intervention.

The aims of the project are, a) to contribute to the knowledge base in relation to the ways in which the teaching of grammar affects pupils' writing; b) to support teachers in relation to the teaching of writing required by the National Curriculum in England, and c) to bring new evidence to the attention of policy makers who are responsible for the teaching of English, and specifically grammar, in the National Curriculum.

The implications of the work for policy will include recommendations in relation to the teaching of writing in England's National Curriculum. The outcomes will also be of interest to other countries, where English is the dominant language in schools, because the teaching of grammar is included as part of most national curricula for primary schools. The findings are likely to be taken up by teachers and their organisations with an interest in the teaching of grammar for writing. Education policymakers with responsibility for the curriculum are also likely to be interested.

The overarching mixed methods research design includes both quantitative and qualitative work with complementary aims (further details below). An important aim of the overall research design is to explore the logic model (Figure 1) that represent our hypotheses about the ways in which the intervention might result in certain outcomes.

Figure 1: Logic Model



We will revise the logic model on the basis of the data, their analyses, and the outcomes of the project.

Research questions (RQs):

1. To what extent is the web-based *Englicious* intervention (EI) effective in improving pupils' writing?
2. To what extent does the teaching in the intervention classes show fidelity to Englicious?
3. In what ways does the teaching in the intervention classes differ from the control classes?
4. How do Year 2 teachers deliver the requirements of grammar in England's National Curriculum?
5. Overall, what explanations does the qualitative analysis provide in relation to the logic model and the outcomes of the assessments of children's writing?
6. In what ways do the outcomes of the research have implications for the teaching of writing in the National Curriculum for primary schools in England?
7. What are the main implications for teacher practice as a result of implementing EI, and, more generally, for evidence-informed teaching of writing?
8. What are teachers' beliefs, knowledge and experience of teaching grammar for writing?

The overall aim of the research design is to generate and analyse "data to examine how an intervention is put into practice, how it operates to achieve its intended outcomes, and the factors that influence these processes" (Humphrey et al., 2016, p. 2). Substantively the research seeks to explore and develop an understanding of the teaching of grammar for writing, for children age six to seven, that is advocated by the Englicious intervention, compared to the teaching in the 'business as usual' conditions in the control classes. The study will contribute to knowledge of the role of grammar teaching in teaching pupils to write. More specifically, the qualitative research will enable greater understanding of the logic and efficacy of the intervention. Work prior to the start of the research has developed a logic model which relates to the study as a whole (see Figure 1) which will continue to be refined throughout the project by this work.

The main research outcome of the research will be a contribution to knowledge in the field of the teaching and learning of writing for infant children. The research will be the first of its kind carried out with children aged six to seven. The outcomes will inform the practice of teachers in primary schools who have to implement the National Curriculum requirements for grammar and writing. The outcomes of the research will also inform National Curriculum policy, in particular the ways in which grammar teaching in the National Curriculum can be, and is being, implemented to support infant children's writing.

Design

The research project as a whole is a mixed-methods research design that combines a randomised controlled trial (with quantitative measurement of pupil outcomes) and the explanatory power of in-depth qualitative inquiry (including process evaluation). Following the baseline measures, the main qualitative data collection and analysis is completed, followed in sequence by the final testing of pupils' writing. The findings of the qualitative research will enable additional explanations of the outcomes of the quantitative work, whether the quantitative outcome is null or statistically significant. The main value of the qualitative research is to provide evidence of the ways in which *Englicious* is implemented, including its unique features compared with the business as usual control condition. This capacity to explain the efficacy of *Englicious* is given added depth by virtue of the reflections on some of the general practice of teaching grammar as part of the National Curriculum that the data collection will reveal. Ultimately the explanatory aspect of the design draws on both qualitative and quantitative outcomes to understand the effectiveness of the logic model that underpins *Englicious* (to this end the design reflects relatively recent developments in mixed methods methodology which emphasise "purposive integration" of quantitative and qualitative elements, Sammons and Davis, 2017).

Quantitative design

This RCT element of this mixed methods study is designed as a two-armed school-level cluster randomised controlled trial. We aim to recruit 60 schools, who will be randomly allocated to the following conditions in equal proportions:

- **Intervention**: Receipt of *Englicious* training in December 2019 or January 2020 (depending on randomisation batch)
- **Waitlist**: Business as usual until after outcome data collection in Summer 2019

Qualitative design

20% of the teachers in the intervention (n=6) and 20% in the control will be selected for the basis of 12 case studies in each group. The samples of teachers will be stratified by years of teaching experience, their school Ofsted grade, and the percentage of free school meals in the school. These criteria are deemed the most relevant for the selection because they relate to teaching quality, and one key school and pupil variable. In all schools selected we

anticipate significant numbers of pupils with EAL and broad ranges of pupil attainment consistent with our focus on schools in London. A selection of teachers will then be made by random selection in order to achieve maximum variation in teacher characteristics across both intervention and control groups.

Each case study teacher will be visited for a total of two full working days – once between the beginning and middle of the term of the intervention, and once towards the end. Each full day visit will consist of the following elements:

1. an overview interview to elicit awareness of the intervention (approximately 20 minutes), or, in the case of control schools, of their approach to teaching grammar. The first interview will also enable the collection of baseline information about the teacher and the school relevant to the research;

2. a semi-structured observation of at least one full lesson where the intervention or other literacy teaching is the focus;

3. a concluding interview using examples from the observed teaching as a stimulus to elicit greater depth of understanding about the implementation (approximately 30 minutes).

Target Population and Sampling

Grammar is taught by all primary schools in England that are required to follow the 2014 National Curriculum. To increase external validity, rather than advertising and asking teachers to self-select, our initial approach will be to randomly draw a sample of schools from the population of primary schools in London using the Department for Education (DfE) schools database as sampling frame. These will be contacted by our research team, requesting that a single teacher in each school takes part. As it is not practical or ethical to require participation, refusals are possible (in which case we will randomly select a replacement), but we will take steps to maximise participation of the first schools we approach. Ultimately, however, it is possible that this does not allow us to recruited the required sample, in which case we will revert to a wider recruitment strategy.

Primary schools in London are some of the most diverse in England, taking account of socio-economic status and the languages spoken. This diversity provides a suitably wide spread of pupil language experience and therefore a test of the efficacy of the approach for mother tongue speakers of English and other languages. We appreciate that sampling only from London reduces the generalisability of our findings, however we believe our approach to sampling is based on appropriate cost-benefit considerations (in the context of an efficacy trial) as we build the evidence base around *Englicious*.

The school sampling process was carried out as follows using a script executed in Stata to allow for replication. This script is available in an appendix to this document:

- We start with a list of all schools in England obtained from the DfE website. Additional data from other DfE sources is also obtained as follows:
    - merging in data from the school level National Pupil Database (NPD) to obtain school-level proportions of children for whom English is an Additional Language (EAL) and who have ever been classified as Free School Meals (FSM) children, in all schools
    - merging in data from the Office for Standards in Education (Ofsted) to obtain schools' most recent Ofsted inspection overall effectiveness grade
- We removed the following schools from the sampling frame:
    - schools located outside the 32 London boroughs and the City of London;
    - schools marked as closed, independent (i.e. fee-paying), or special schools;

    - those not identified by the DfE as primary schools and those with a lower age range starting above 6 (to exclude "junior" schools, which lack the year group that the intervention targets);
    - schools that are marked as 'inadequate' in their most recent Ofsted inspection (who are, hence, in *special measures*).
- A random number between zero and one is generated a for each school in this sampling frame. This is done using a 'stable seed', which ensures that if the process is repeated with the same seed we get the same outcomes, which is important to allow for replication. The value of the stable seed is obtained from a specified source and reported in the sampling appendix to this document.
- Half of the sample is placed into a high EAL group and half into a low EAL group
- Half of the sample is placed into a high Ever FSM group and half into a low Ever FSM group
- Four stratification groups defined by the intersection between these EAL and FSM groups (i.e. high/high, high/low, low/high, low/low) are formed. Note that these are not all the same size because of the correlation between EAL and FSM
- The schools within each stratification group are sorted by the random number they have been assigned (i.e. arbitrarily) and, taking into account the size of the stratification group as a proportion of the overall sample, an identifier of the appropriate number within this stratification group that are the ones we should contact first for recruitment are generated. A second cohort is generated using the same method, and so on.
- A spreadsheet is produced in which there are 60 recruitment groups (one for each space on the RCT) and the order in which schools should be contacted within these groups. This provides for predictability and minimises selection on unobservables in proceeding through the recruitment process.

Where schools decline to participate, we will continue to the next school in the list identified for this recruitment group identified by the sampling process (although, as noted above, we will attempt to maximise participation of the first schools sampled to maximise external validity).

Drawing on guidance issued by the EEF (2013), having randomly selected our initial target schools we will attempt to secure recruitment through existing connections, emails and telephone calls to the schools. We will prepare an information pack for schools making clear the aims of the research, what taking part would involve and the benefits to them and their pupils (whether they are allocated directly to treatment or a waiting list). With positive follow-up, we will arrange a more detailed conversation with a senior leader and proposed class teacher involved to ensure proper commitment; we will also ask schools to sign a 'memorandum of understanding' (based on that suggested by EEF, 2013) committing to play a full part in the research regardless of the arm to which they are allocated to ensure they understand this and that it is important to the research. It is better for schools to decide against participation at this stage than once the trial has begun.

As part of the school recruitment process, schools would be expected to identify the teacher who would participate in Englicious training and, subsequently, deliver the appropriate lessons to the pupils within their class (if they are randomly allocated to the treatment group). Thus, the sample of pupils for this trial are those within the class of recruited teachers.

Schools will not be considered fully recruited for the purposes of the evaluation until all required pre-trial data have been collected from schools, consent procedures have been carried out (possibly with minimum threshold – see further discussion below), and pre-testing has been carried out. If possible, we will slightly over-recruit to guard against school-level attrition, although we will work extremely hard to minimise this given that attrition after

randomisation poses a threat to internal validity of the trial. Maximising retention will involve ensuring we continue to communicate with all schools throughout the trial period to ensure they do not lose interest or contact with us. We will provide a named contact to all schools should they have any queries at any point throughout the evaluation. We will minimise the burden on schools, especially those allocated to the control/waitlist group. We will need to arrange testing at the end of the trial period, but will minimise the work that this involves for schools by arranging test administration ourselves (this is also important to blinding of outcomes; the RAs conducting testing will also be kept blind to treatment arm of a school).

Randomisation

Randomisation will be at the school/teacher-level (only one teacher will be recruited within each school, hence these two are indistinguishable). It will not be practical to use pupil-level randomisation for an approach that requires training of teachers, since this would require schools to reorganise classes for our research, to which they would be very unlikely to agree.

Randomisation will be carried out in two batches for reasons of delivery and recruitment practicality. In each batch, 30 Year 2 teachers (infant pupils aged 6-7) will be randomly allocated to two groups (EI and control) in equal proportions. Group a) will receive training in using *Englicious* as part of their teaching of writing. Group b) will receive the offer of training in *Englicious* if it is demonstrated to be effective in the trial (consistent with the ethics guidance of only offering approaches likely to benefit pupils) or an alternative teacher development opportunity in knowledge about language should this not be appropriate. Training for this group will not occur until after the post-test has been completed among all participating schools (end of trial).

We will carry out each randomisation within stratification blocks to reduce the risk of imbalance on important characteristics between our resulting treatment and control groups. These stratification blocks will be formed by the intersection between equally-sized high and low EAL proportion, and high and low FSM proportion groups.

The randomisation process will be carried out using a script executed in Stata to allow for replication. The script will also be made publicly available for this purpose.

Sample size calculations

Preliminary power calculations indicate that if we achieve the target sample this will be sufficient to detect a minimum effect size of 0.25. These were carried out under the following assumptions. We will conduct school-level cluster randomisation (this is identical to class-level randomisation since there will only be one teacher participating per school) of 60 groups to two equal-sized arms (i.e. 30 per arm). We further assume 15 pupils are tested per school (giving a generous allowance for data processing objections and non-response), intra-cluster correlation of the outcome measure of 0.15, 0.49 of post-test variance (corresponding with test-retest correlation of 0.7) in outcome explained by pre-test at both individual- and cluster-level (McConnell & Vera-Hernández, 2015), and usual assumptions of two-tailed significance tests at 0.05-level and power of 0.8.

Outcome Measures

Quantitative data

Robust assessment of writing is challenging. However, we see it as central to the aims of this trial. The most appropriate approach for our research is a standardised measure of pupils' technical English skills (i.e. grammar, punctuation and spelling) and of writing. We have secured agreement from GL Assessment that we can use elements from their

Progress in English (PiE) test focussed on writing. Their more recent Progress Test in English (PTE) does not include a standardised writing element (partly because of the challenges posed by assessing writing), which is why we take this approach.

Our proposed primary outcome measure is raw score on the longer writing task drawn from the GL Assessment Progress in English (PiE) test, which is designed with links to the National Curriculum (however, these links are with the version of the national curriculum prior to the 2014 version; this older PiE is used rather than GL Assessment's more recent Progress Test in English–PTE–as this has removed assessment of writing). Nevertheless, we think its stated aims of assessing the following remain relevant: "Writing in narrative and non-narrative forms; using sentences that are grammatically correct; choosing words for variety and interest". The marking rubric is designed to assess attainment across four strands: composition and effect, text organisation, sentence structure, and vocabulary, with twice the weighting given to composition and effect. This primary outcome measure task has been designed and validated for use with our age group of interest by GL Assessment. As such, we believe the risks of ceiling and floor effects have been minimised. Our discussion of these in the application was primarily regarding the additional sentence construction measure (further discussion below), although we believe it is right to be aware of the risks of ceiling and floor effects and consider what action we can take to mitigate this, should it arise.

In addition, we plan to use a bespoke sentence construction test (based on that used by Arfé, Dockrell & De Bernardi, 2016) in part because an additional core measure of children's writing competence is the fluency with which single written sentences are generated. The ability to translate ideas into written sentences flexibly influences text production, and develops through primary school (Berninger, Nagy, & Beers, 2011). Fluency in this process (i.e. writing sentences) is directly associated with dimensions of writing proficiency including written text productivity, microstructural accuracy and overall text quality (Arfé & Pizzocaro, 2016; Arfé et al., 2016). Our proposed task assesses children's fluency in generating ideas in single written sentences. Children receive a sheet of lined paper with two-word pairs and are asked to generate as many different sentences as they can from the two words in five minutes (Arfé & Pizzocaro, 2016). A score of 1 will be given to each sentence that is both grammatically and semantically correct. In scoring sentence accuracy, errors in punctuation, capitalization or misspellings will not be not coded. Inter-rater reliability for a measure of this type has been found to be good (94%) and test-retest reliability at a two-month interval is .62 (Arfé et al., 2016). These tests are, or will be, designed to be appropriate to the age of the children involved, in order to avoid ceiling or floor effects, ensuring, for example, that there are items that all pupils should be able to make at least a good attempt at.

Some adaptation of these tasks will be necessary for use within the context of an RCT, including enforcing the length of writing time more strictly than would normally be the case (to ensure this does not differ systematically across treatment arms), and scripting the introduction of the tasks (both to avoid systematic differences across arm and to add context that would otherwise have been provided by earlier stages of the test).

Pupils will be tested twice: 1. Baseline: prior to the start of the intervention; 2. Immediate intervention effects: three months after the intervention (in line with standard practice). We propose to use the shorter writing task from the GL Assessment Progress in English as a baseline assessment rather than re-using the longer writing task for the following reasons:

- Since the evaluation design is a randomised controlled trial, there is no prima facie requirement to use a pre-test that is the same as the post-test (or even necessarily to use a pre-test at all): due to randomisation there should be no differences between the treatment and control group except those present due to chance. As such, the primary purpose of a pre-test is to improve the precision of the treatment effect estimate. We do note that it will also help to demonstrate the balance between the

groups in writing attainment at baseline (as noted in the proposal, we would do this using standardised differences as advocated by Imbens & Rubin, 2015).
- Given this primary purpose, when using a pre-test we should choose one that will maximise the outcome measure variance explained (i.e. the two are highly correlated) without negative knock-on effects on our primary measure of interest. While using the same test twice would likely be highly correlated, there is a risk that using precisely the same test twice would reduce the utility of the long writing task (primary outcome measure) in assessing pupils' writing ability, since their familiarity with the test may reduce the variation in scores on its second use. In addition, use of the longer assessment at both time points would increase costs (primarily of marking), so this does form part of a balancing of costs and benefits.
- As such, we plan to use a similar, but non-identical, measure. We believe there is a convincing case that they are aiming to measure the same constructs. Both tasks are drawn from GL Assesment's PiE and are assessed across the same dimensions of composition and effect, text organisation, sentence structure, and vocabulary.

It will not be possible for the researchers running the tests to be 'blind' to the group allocation, as they will also be collecting information about compliance in the case of treated schools (see below), however we think that the risk of this affecting test performance among pupils in the class should be minor. The UCL Institute of Education doctoral students who will mark the scripts will be allocated a mix of tests from treatment and control groups and will be kept blinded from which are which, since this would have a higher potential to introduce bias.

Qualitative data

The qualitative data collection will be conducted in primary school Year 2 classes during the time of the delivery of the intervention and the control 'business as usual' conditions. The qualitative data collection will include the following main elements undertaken towards the beginning and towards the end of the intervention period:

Interviews with teachers (who are teaching either the intervention or control condition):

- to understand their beliefs, knowledge and experience in relation to the teaching of grammar and the teaching of writing, including how they approach planning of the lessons;
- to evaluate the extent to which the teaching has fidelity with the intended intervention in the intervention classes;
- to probe the reasons for why decisions about teaching were made during observed lessons;
- to reflect on the nature of the differences between the intervention classes and the control classes;
- to understand more about how England's National Curriculum requirements for grammar impact on teaching in Year 2 classes.

Observations of literacy lessons (where *Englicious* is the approach, or literacy teaching in the control condition):

- to evaluate the extent to which the teaching shows fidelity to the intervention;
- to collect key examples of teaching moments to be discussed at the post-teaching interviews;
- to understand the differences and similarities between control classes and intervention classes;
- to compare observations of teaching with teachers' perceptions of their approaches to teaching described during interviews (triangulation);

- to understand more about how England's National Curriculum requirements for grammar impact on teaching in Year 2 classes.

Case studies: the analysis of the interview and observation data will be used to:

- produce holistic case studies of each teacher, focused on their teaching of writing, necessary to understand significant patterns and differences between intervention conditions and control conditions.
- reflect on the potential reasons for the experimental outcomes
- refine the logic model and its rationale

Two different semi-structured interview schedules will be used to reflect the beginning interview and the final interview (see Appendix 2). The interviews will be audio-recorded and fully transcribed. Observations of lessons will be recorded as field notes (hand-written and/or using digital devices), and will then be transferred to a proforma (see Appendix 3) as soon after the observation as possible.

The proforma will focus on the following areas, and therefore will structure the observation focuses:

- implementation environment – e.g. to what extent does the teacher's planning reflect the intervention, or other grammar teaching in a control condition? What is the nature of the classroom environment and ethos including physical characteristics? To what extent is there a supportive environment for the intervention in the school more generally?
- implementer characteristics – in what ways does the teacher-interaction reflect the objectives of the lesson and intervention (or other lesson planning) more broadly?
- participant characteristics – what is the nature of the pupils' response to the teaching?

Table 1 lays out the timeline for the qualitative data collection, and shows the use of research staff.

. The approach to staff expertise for the qualitative data collection combines the observations of research staff with knowledge of:

a) primary education and grammar for writing teaching expertise;
b) linguistics and the specific design of *Englicious*;
c) qualitative research data collection

The project's research fellows will provide continuity of observations across all teachers visited as part of the qualitative data collection. Dr Sing will collect qualitative data at all school visits. Dr Manyukhina will also visit four schools (two intervention; two control) with Dr Sing to assure quality of qualitative data (separate from the visits indicated in Table 1). This approach to data collection improves validity of the qualitative research through ongoing triangulation of observations by more than one observer at a minimum of 50% of school visits. In addition, the sequence of visits allows researchers with expertise in primary education, writing, and grammar (Profs Wyse and Aarts) to visit 50% of case study schools (intervention and control), including for the purposes of assuring the rigour of the work of the research associate.

Table 1: Timeline, and staff[1] involvement, for qualitative data collection

	Activity	**Staff carrying out work**
Autumn 2019	Agree dates for visits to schools	SS, DW, BA
December 2019	Agree final versions of interview questions and field note analysis proformas	All team members as appropriate
January to February 2020	First one-day visits to **intervention** schools 1 to 3: interviews and observations of teaching *Englicious*	DW & SS
	First one-day visits to **intervention** schools 4 to 6: interviews and observations of teaching *Englicious*	BA & SS
	First one-day visits to **control** schools 1 to 3: interviews and observations of teaching *Englicious*	DW & SS
	First one-day visits to **control** schools 4 to 6: interviews and observations of teaching *Englicious*	BA & SS
March to April	Second one-day visits to **intervention** schools 4 to 6: interviews and observations of teaching *Englicious*	BA & SS
	Second one-day visits to **intervention** schools 4 to 6: interviews and observations of teaching *Englicious*	DW & SS
	Second one-day visits to **control** schools 4 to 6: interviews and observations of teaching *Englicious*	BA & SS
	Second one-day visits to **control** schools 4 to 6: interviews and observations of teaching *Englicious*	DW & SS

Analysis plan

Quantitative analysis

The primary outcome will be specified as an intention to treat (ITT) analysis of the longer writing task from the GL Assessment Progress in English test (see further below). We will also explore the impact on compliers using an instrumental variables analysis (Angrist et al., 1996) based on an agreed definition of compliance (see below). We will use the shorter writing task from the PiE assessment as a pre-test and explore imbalance of this (using

[1] Professor Bas Aarts (BA); Professor Dominic Wyse (DW); Dr Yana Manyukhina (YM); Dr Sue Sing (SS)

absolute standardised differences, Imbens & Rubin, 2015) and other available school- and pupil-level characteristics between the treatment and control groups; pupil-level covariates will also be used to improve precision in the analysis model, while, given the clustered nature of the trial, we will calculate cluster-robust standard errors at the school level. The analysis model will be fully specified in the statistical analysis plan and changes will not be made once outcomes data have been collected. Planned pupil-level data would include, for example, gender, whether English is an Additional Language, and eligibility for free school meals, which could also be used in exploratory sub-group analyses, as part of the qualitative research, and to understand the external validity of our sample (see further below).

Our primary analysis model is expected to be a linear regression model of the following form:

$$Y_{ij} = \alpha + \beta_1 Treat_j + \beta_2 PreTest_{ij} + \boldsymbol{\gamma}' \boldsymbol{X}_{ij} + \varepsilon_{ij}$$

where Y_{ij} is the long form writing task score for individual i nested in school j, $PreTest_{ij}$ is the short form writing task score for the same pupil, $Treat$ is our school-level treatment indicator, $\boldsymbol{X}$ is a vector of school- and pupil-level covariates to improve precision (that will be finalised for the SAP), and ε is a pupil-level error term. Standard errors will be calculated taking into account clustering at school-level (j), which appropriately accounts for this clustering without making additional assumptions about the distribution of school effects (as is the case with multi-level modelling).

Qualitative analysis

The qualitative data analysis plan involves a first phase of standard Qualitative Data Analysis (QDA), including thematic coding facilitated by NVivo QDA software, followed by a second phase of QDA which will build short case study reports for each teacher using data identified from the first phase of QDA. Some high level codes will be pre-specified to ensure allignment with the overall aims of the process analysis. A final phase of analysis will seek overall synergies and contrasts, including with the findings of the quantitative data, in order to build a more sophisticated understanding of the strengths and weaknesses of the intervention in relation to the control conditions.

After each school/teacher visit, audio recordings of interviews will be sent for transcription, then imported into NVivo. Field notes will be typed up, if handwritten, and transferred to proformas (by the researcher who carried out the observations), sense-checked, the formatting cleaned, and imported to NVivo. All proformas will be read in full with analytic memos recorded. Using standard QDA approaches the data will be *coded* by Dr Sing and Dr Manyukhina to inform the development of case studies of the teachers. A 5% selection of coding will be selected at random for critical evaluation by Dr Prof Wyse, in addition to ongoing discussions about the progress of data analyses.

Analysis of interview data will be triangulated with analysis of lesson observation data to ensure robustness of the case studies. The first phase of case study development will be thematic, and will seek key incidences by drawing across a) the intervention data, and b) the control data. Case study narratives for each teacher will be written drawing on key moments in the data identified as part of the coding process. Once case studies have been drafted a further overall comparison will be made between intervention and control schools to seek explanations for the observed phenomena, and to search for and explain key differences. As an essential part of the mixed methods research design, later in the project once the outcomes of the quantitative analysis are known, the case studies will be reviewed again to seek further explanation for the outcomes, whether the quantitative outcomes are null or statistically significant. The qualitative data analysis and subsequent finalisation of case studies will be carried out by the research associates closely overseen by Professor Wyse. Part of the work to oversee quality will be regular meetings to review the emerging findings of the qualitative research.

Definition of compliance

As part of their training, participating teachers will be asked to log delivery of all planned lessons in a section of the paper manual provided. As part of the data collection at the end of the year, researchers carrying out testing in the schools will take a photograph of the log page in the paper manual. This will be used as a compliance measure in two ways:

1. As a continuous measure;
2. As a binary measure, where teachers who report having successfully delivered all classes in the sequence will be deemed to be compliant and those who do not report this are deemed non-compliant.

Ethics and registration

This project was approved through the UCL Institute of Education Research Ethics Committee approved process and allocated the project number REC1170. As we will be processing personal data the project was also allocated an ICO registration number as follows: Z6364106/2019/01/150 social research. The trial is registered on the Registry of Efficacy and Effectiveness Studies with Registry ID: #1920.1v1 (https://sreereg.icpsr.umich.edu/framework/pdf/index.php?id=2540).

Personnel

- Dominic Wyse: Principal Investigator
- Bas Aarts: Co-Investigator
- Jake Anders: Co-Investigator
- Julie Dockrell: Co-Investigator
- Carole Torgerson: Co-Investigator
- Yana Manyukhina
- Ian Cushing

Risks

There is a risk of lack of power in this trial. This is fairly common in efficacy trials and must be balanced against costs of expanding the approach to a larger sample without evidence of promise. As noted above, issues of floor effects in the measurement of writing may be a particular concern in the age group that we are targeting with potential implications in attenuating any effect of the intervention. Should there be evidence of ceiling effects in our measures, despite our attempts to use age-appropriate measures to avoid this issue, we would explore use of tobit regression models (in place of the linear regression models otherwise planned) to check the robustness of our findings to this issue (e.g. McBee, 2010). Clearly, however, this may not entirely remove the attenuation bias and could still have implications for statistical power.

An RCT should have a high degree of internal validity. However, there are factors that could undermine this. Perhaps the most likely is attrition, with schools uncooperative in retesting of their pupils (pre-test would be conducted prior to randomisation). Our use of a waitlist design (where the control group receives treatment after post-test, provided it shows promise, or if not then an offer of teacher development in knowledge about language) will help to minimise this possibility. Furthermore, we will develop strong working relationships with all participating teachers to this same end. It is also relevant to consider external validity, i.e. the generalisability of our results, which would be somewhat limited by concentrating on a single year group within schools in London (notwithstanding some similarities in the SES of pupils in other cities in England). However, external validity is a lower priority at efficacy trial

stage.

There are key risks in research projects of this type around recruitment and data collection. We have strengthened our team by including a researcher (Manyukhina) with strong skills in liaising with schools to contribute to management of both of these processes in order to reduce these concerns, rather than recruiting for this externally. Both processes will also be closely managed, and mitigation plans put in place should there be signs that they are going off track.

Timeline

Date	Activity
January 2019	• Project start
Jan-Mar 2019	• Data privacy impact assessment (DPIA). IOE ethics application and UCL data protection registration (including review of DPIA, legal basis for processing, and privacy information to be provided to parents). • Recruit researchers. • Manualise the *Englicious* intervention (EI; in accordance with the specifications of the National Curriculum for KS1), by working with a group of four year 2 teachers.
Mar 2019	• First Advisory Group Meeting
Apr-Jun 2019	• Recruitment of schools
Jun-Aug 2019	• Update programme documentation using teacher feedback.
Sep-Oct 2019	• Undertake ethics/data protection procedures, including provision of information sheets, consent forms (as appropriate) and data privacy notices to parents.
Nov-Dec 2019	• Pre-test all pupils using Progress Test in English (PTE) short test and sentence-construction measure. • Two batches of randomisation of teachers to two trial arms. • Train first batch intervention group teachers for EI.
January 2020	• Train second batch intervention group teachers for EI. • Second Advisory Group Meeting
Jan-Apr 2020	• Teaching of intervention and controls takes place. • Process evaluation data collected. • Ongoing analysis of process evaluation data.

	• Researchers analyse results of the pre-test PTE and sentence-construction measure.
Jun 2020	• Administer post-tests for all pupils (PTE long test and sentence-construction measure).
Jul-Sep 2020	• Marking of post-test PTE and sentence construction measure. • Begin research write-up and writing of academic papers. • Begin implementation of communications plan.
Sep 2020	• Third Advisory Group meeting
Sep-Oct 2020	• Analyse results using PIE post-test writing task and sentence construction measures. • Training of control group teachers in EI (or alternative, as appropriate).
Nov-Dec 2020	• Completion of research write up and writing of research papers. • Further communication plan actions.
Dec 2020	• Final Advisory Group meeting aligned with dissemination event • Project end.

References

Aarts, B. (2011) *Oxford Modern English Grammar*. OUP.

Aarts, B., S. Mehl & Wallis, S. (2016) Language learning at your fingertips. In: K. Corrigan & A. Mearns (eds.) *Creating & Digitizing Language Corpora*. Palgrave. 211-239.

Aarts, B. & E. Smith-Dennis (2018) Using corpora for English language teaching and learning. In: McIntyre, D. & Price, H. (eds.) *Applying linguistics: language and the impact agenda*. Routledge.

Angrist, J., Imbens, G. & Rubin, D. (1996) Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434). 444-455.

Arfé B., & Pizzocaro, E. (2016). Sentence generation in children with and without problems of written expression. In J. Perera, M. Aparici, E. Rosado, & N. Salas (Eds.), *Written and Spoken Development across the Lifespan*. Springer.

Arfé, B., Dockrell, J. E., & De Bernardi, B. (2016). The effect of language specific factors on early written composition: the role of spelling, oral language and text generation skills in a shallow orthography. *Reading and Writing, 29*(3), 501-527. doi:10.1007/s11145-015-9617-5

Berninger, V.W., Nagy, W. & Beers, S. (2011) Child writers' construction and reconstruction of single sentences and construction of multi-sentence texts: contributions of syntax and transcription to translation. *Reading and Writing*, 24(151). DOI: 10.1007/s11145-010-9262-y

EEF [Education Endowment Foundation] (2013) Recruitment and Retention Pack. EEF Guidance, Sept. 2013.

The Government Office for Science. (2008). *Foresight Mental Capital and Wellbeing Project (2008). Final Project Report.* London: The Government Office for Science.
Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology*, 99(3). 445-476. doi:10.1037/0022-0663.99.3.445
Graham, S., McKeown, D., Kiuhara, S. & Harris, K. (2012) A meta-analysis of writing instruction for students in the elementary grades. *Journal of Educational Psychology*, 104(4). 879-896. DOI: 10.1037/a0029185
Humphrey, N., Lendrum, A., Ashworth, E., Frearson, K., Buck, R. & Kerr, K. (2016). *Implementation & Process Evaluation (IPE) for Interventions in Education Settings: An Introductory Handbook*. EEF.
Imbens, G. & Rubin, D. (2015). *Causal Inference for Statistics, Social & Biomedical* Sciences. CUP.
Koster, M.,Tribushinina, E., De Jong, P.F. & Van den Bergh, B. (2015) .Teaching children to write: a meta-analysis of writing intervention research. *Journal of Writing Research*, 7(2). 249-274.
McBee, M. (2010) Modeling outcomes with floor or ceiling effects: an introduction to the Tobit Model. *Gifted Child Quarterly*, 54(4). 314-320. DOI: 10.1177/0016986210379095
McConnell, B. & Vera-Hernández, M. (2015) Going beyond simple sample size calculations: a practitioner's guide. IFS Working Paper W15/17, Institute for Fiscal Studies.
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. 139-166.
Rogers, L. & Graham, S. (2008) A meta-analysis of single subject design writing intervention research. *Journal of Educational Psychology*, 100(4), 879-906. DOI: 10.1037/0022-0663.100.4.879
Sammons, P. & Davis, S. (2017) Mixed Methods Approaches and their Application in Educational Research. Chapter 23 in *Dominic Wyse, Neil Selwyn, Emma Smith and Larry E. Suter (eds.) The BERA/Sage Handbook of Educational Research, Volume 1.* London: Sage.
Torgerson D., Torgerson, C., Mitchell, N., Buckley, H., Ainsworth, H., Heaps, C., & Jefferson, L. (2014) *Grammar for Writing: Evaluation Report and Executive Summary*. EEF report.
Troia, G. (2014). Evidence-based practices for writing instruction (Document No. IC-5). Retrieved from University of Florida, Collaboration for Effective Educator, Development, Accountability, and Reform Center website: http://ceedar.education.ufl.edu/tools/innovation-configuration/
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.