Grammar and Writing Research Project: Statistical Analysis Plan

Jake Anders, Julie Dockrell, Sue Sing, Carole Torgerson, Dominic Wyse

Trial registration: https://sreereg.icpsr.umich.edu/framework/pdf/index.php?id=2540

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). *Englicious* combines formal grammar teaching, that is specified in England's national curriculum, with emphasis on how this grammar links with some of the processes of writing. 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 England's 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. As of November 2020, over 10,000 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.)

Design

- Two-armed cluster school-level cluster randomised
- Stratification: Proportion of FSM-eligible students; proportion of EAL students; randomisation batch
- Primary outcome: Writing attainment measured using GL Assessment Progress in Writing
- Secondary Outcome: Sentence combining

We aimed to recruit 60 schools to 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 2021

Randomisation and follow up

Randomisation was carried out at the school/teacher-level (only one teacher was recruited within each school, hence these two are indistinguishable). It was not deemed 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 was always planned to be carried out in two batches for reasons of delivery and recruitment practicality, with the aim being that each batch to be the same size, with 30 Year 2 teachers (infant pupils aged 6-7) randomly allocated to two groups (EI and control) in equal proportions. However, initially due to challenges with recruitment and then due to large changes to delivery caused by COVID-19 disruption ultimately randomisation was carried out as follows:

- Batch 1 (November 2019): 24 schools were allocated in equal proportions (12 to treatment; 12 to comparison)
- Batch 2 (January 2020): 40 schools were allocated in equal proportions (20 to treatment; 20 to comparison)

This provided a sample of 64 schools, slightly above our 60-school recruitment target. We moved ahead with intervention delivery in Spring/Summer 2020 on the basis of these allocations. However, COVID-19 disruption meant that this was abandoned, and delivery was deferred to Summer 2021 instead. This meant that the planned sample of pupils would all be too old to take part in the planned delivery and, in a substantial number of schools, the participating Year 2 teacher would not necessarily be teaching Year 2 in Summer 2021. This large change and the COVID-19 disruption more generally meant that 22 schools across Batch 1 and 2 schools (12 treatment; 10 control) decided not to take further part in the project.

Based on these changes, it was necessary to re-collect pupil data and baseline assessment. In addition, we discussed whether to re-randomise the remaining schools. This was rejected on the grounds that teachers allocated to the treatment group had received training and so it would not be realistic to treat them as true comparators if allocated to the comparison group in a randomisation. As such, remaining Batch 1 and 2 schools were retained in their allocated treatment group. Further recruitment was carried out in in Autumn 2020 and subsequent pupil data collection and baseline testing in Spring 2021 resulting in the following groups and allocations:

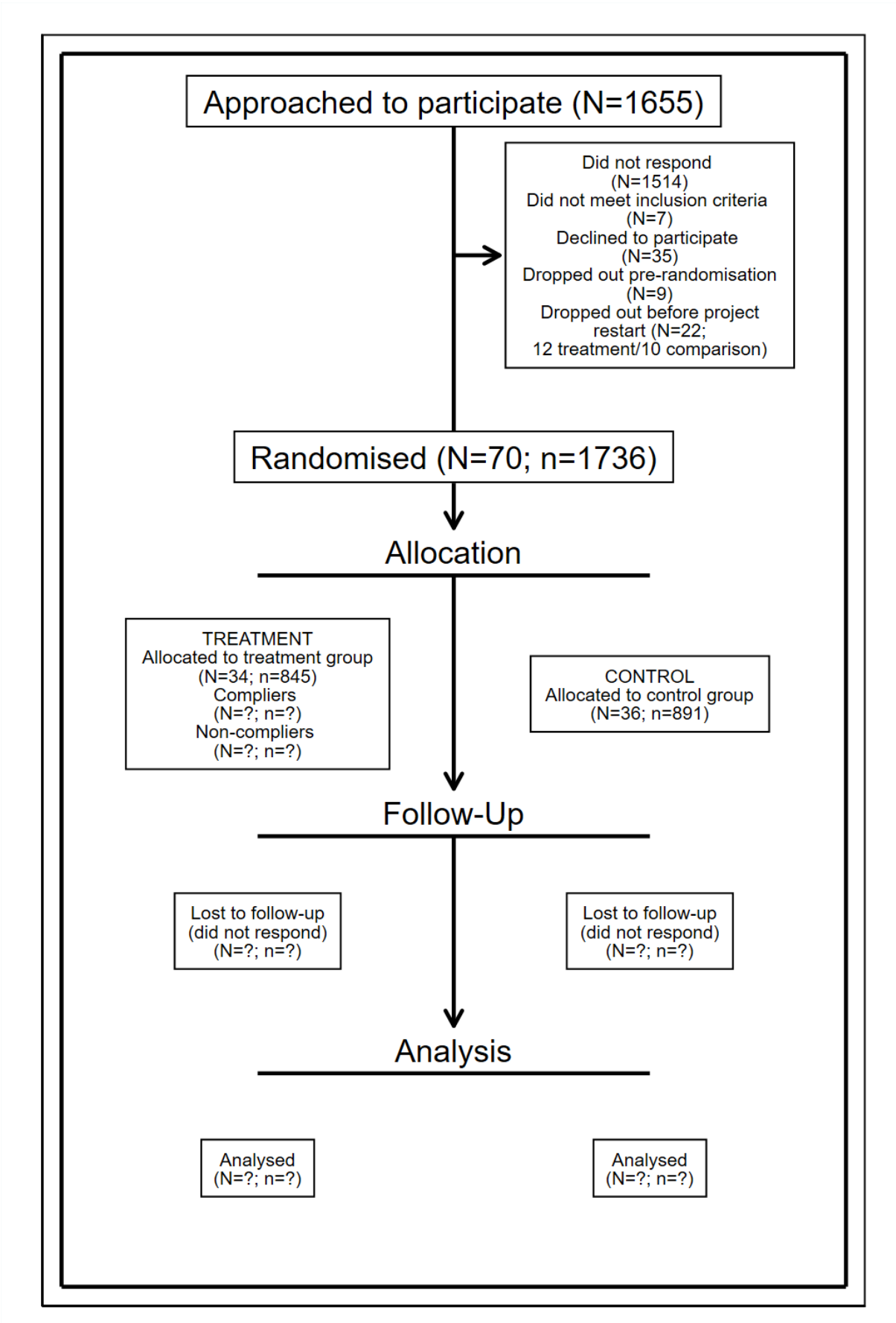
- Batch 1 (updated): 18 schools in equal proportions (9 treatment; 9 comparison)
- Batch 2 (updated): 24 schools with slightly fewer treated schools remaining (11 treatment; 13 comparison)
- Batch 3 (March 2021): 28 schools were allocated in equal proportion (14 to treatment; 14 to comparison)
- OVERALL: 70 schools (34 treated; 36 control)

Randomisation within each batch was, as anticipated, carried out 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 was carried out using a script executed in Stata to allow for replication. The scripts for this purpose will be included in the project report.

A full CONSORT diagram will be reported in the project report. A CONSORT diagram up to the point of randomisation is reported as Figure 1.

Figure 1. Provisional CONSORT diagram



Sample size and statistical power update

Preliminary power calculations indicated that if we achieve the target sample of 60 schools this would be sufficient to detect a minimum effect size of 0.240.[1] These were carried out under the following assumptions:

- stratified school-level cluster randomisation (this is identical to class-level randomisation since there will only be one teacher participating per school) of 60 schools in 8 strata (4 within each of 2 randomisation batches) to two equal-sized arms (i.e., 30 per arm),
- 15 pupils 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 model covariates at both individual- and cluster-level (McConnell & Vera-Hernández, 2015) using 4 regressors,
- usual assumptions of two-tailed significance tests at 0.05-level and power of 0.8.

After all three randomisation batches (removing 22 schools – 12 treatment; 10 control – who did not participate in the updated pupil data and baseline collection ahead of revised delivery in summer 2021) the data that we now hold provides the following new information relevant to our assumptions (while maintaining the other assumptions as they were at design stage):

- stratified school-level cluster randomisation of 70 schools in 12 strata (4 within each of 3 randomisation batches) with a 0.486 allocation to treatment (i.e., 34 to treatment and 36 to control);
- these schools have 24.8 (arithmetic); 24.2 (geometric); 23.3 (harmonic) pupils per school on average (harmonic mean used as more conversative/robust to variance in number of pupils per school);
- 11 baseline covariates included in model to explain outcome measure variance.

Based on this new information, we provide two revised MDES estimates as at the point of Batch 3 randomisation (other assumptions maintained given no basis to update). The first is carried out on the assumption of no attrition from this point (and provides comparability with estimates often reported at this point) while the second makes more realistic/pessimistic assumptions about school- and pupil-level attrition in collection of outcome measures (to provide an estimate of an MDES we think we are more likely to achieve either due to this or due to optimism bias in other estimates):

- No attrition (70 schools, 23 pupils per school, 0.486 allocation ratio): 0.211
- Realistic/pessimistic (51 schools, 19 pupils per school, 0.486 allocation ratio): 0.256

Outcome Measures

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

[1] Note that due to a typo in the protocol this was reported as 0.25.

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 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).

In addition, we 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 and accuracy of the sentences that children generate. One mark is awarded if the sentence is different from the child's other sentences on the test paper (zero marks in total if the sentence is not different from previous sentences); one mark is awarded if the sentence is written using standard English grammar; and one mark is awarded for semantic meaning (the sentence makes sense on its own). 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 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: 1-2 weeks from the end of the intervention. This is a deviation from protocol in which we planned for more delayed testing but has been necessary due to delivery alterations caused by COVID-19.

Also due to alterations to delivery and continuing COVID-19 restrictions, both outcome and pre-test measures will be administered by classroom teachers, which means that administration will not be blinded to treatment allocation. However, we argue that the nature of the tasks and the scripting of the introduction etc. minimises the potential effects of this on test performance among pupils in the class.

Tests will be marked by a team recruited specifically for this purpose (largely drawn from doctoral students at UCL) as follows:

- Markers will receive training on the marking of all tasks to ensure consistency in their approach. The approach that they will take has been set out in a Marking Guide, which will be shared with all markers and published as an appendix to our final report. The training will be followed by practice marking of 4-5 scripts which are not part of the sample. These will be analysed for agreement and error patterns, including calculation of inter-rater reliability statistic compared to agreed mark by senior raters. If inter-rater reliability is below 0.6 then further training would be carried out with further 4-5 scripts analysed in the same way.
- A 20% sample of first 100 scripts of each task marked by each marker will also be marked by a second marker. If inter-rater reliability of these falls below 0.7 then senior raters will mark this sample and calculate inter-rater reliability of each marker compared to the senior raters. Further training will be provided for the marker if their inter-rater reliability falls below 0.7 in this check against the senior raters (the training will follow the approach above and then returning to the same 20% sampling process). If inter-reliability is above 0.7 in these checks then we will move to 1%

sampling of further scripts for double marking in order to monitor the remainder of the marking process.

- Markers will be kept blinded from whether any given test they are allocated is treatment or control, since this would have substantial potential to introduce bias.
- Markers will be allocated a mix of tests from treatment and control groups to reduce risk that tester effects could drive results at the margin.
- Markers will mark in batches of different tasks to prevent the possibility of their perception of one task shaping their marking of another task, particularly across pre- and post-tests.

We will plot a pooled histogram of each of the outcome measures in order to understand the distribution, potential ceiling/floor effects and, hence, the potential need for robustness checks resulting from this. Tobit regression modelling (otherwise following the analysis plans set out below) would be used as a robustness check in the event of ceiling and/or floor effects.

To provide guidance for future projects using these or similar outcome measures, we will estimate:

- The intra-cluster correlation of our primary and secondary outcome measures using an empty variance components model: $Y_{ij} = \alpha + \phi_j + \varepsilon_{ij}$ where individual i is nested in school j, Y_{ij} is the measure of interest, ϕ_j is a school-level random effect, and ε_{ij} is an individual-level error term. The school-level random effect is assumed to be normally distributed and uncorrelated with the individual-level errors. The intra-cluster correlation itself will be estimated from this model using the following equation: $\rho = \frac{var(\phi_j)}{var(\phi_j)+var(\varepsilon_{ij})}$

- The bivariate proportion of post-test measure variance explained by the analogous pre-test measure (i.e., between pre-test sentence construction measure and post-test sentence construction measure; between PIE short writing task and PIE long writing task) will be estimated. This will be done using a hierarchical linear model of the post-test with school-level random intercepts and the relevant pre-test measure as the only predictor. Within-group (pupil-level), between-group (school-level) and overall variance explained will be extracted from this model.
- The proportion of primary and secondary outcome measures explained by the baseline covariates included in the primary analysis model will be estimated. This will be done using a hierarchical linear model of the outcome measure with school-level random intercepts and the baseline covariates reported below as predictors. Within-group (pupil-level), between-group (school-level) and overall variance explained will be extracted from this model.

Analysis plan

Imbalance analyses

We will check for balance of sample as randomised and primary analysis sample for the following characteristics:

- Pre-test measures (sentence combining score and PIE short writing task raw scores)
- Pre-test measures in previous year for batch 1 and 2 (given main pre-test unblinded due to COVID-19-related delays)
- FSM
- EAL
- Gender

- School's average KS1 points score from 2018/19

We will do this by reporting means and standard deviations for the treatment and control group and calculating absolute standardised differences (Imbens & Rubin, 2015) (i.e., the absolute value of the mean difference divided by the sample standard deviation) between the treatment and control groups. These provide a simple, scale-free measure of differences that is easy to interpret.

In the event of imbalance on key characteristics exceeding 0.1 standard deviations – and where this characteristic is not already included in our planned modelling – we will re-estimate the primary analysis model in order to check the robustness of our findings to attempting to control statistically for this imbalance.

We will plot the pre-test measures by treatment status using overlapping histograms.

Primary outcome 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.

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

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

where Y_{ij} is the long form writing task score for individual i nested in school j, $Treat$ is our school-level treatment indicator, $\boldsymbol{X}$ is a vector of school- and pupil-level covariates to improve precision (further details below), η_j is a vector of randomisation stratification variables, and ε is a pupil-level error term.

Statistical inference will be carried out using randomisation inference (Athey & Imbens, 2017),[2] taking into account the clustering of pupils into schools and the stratified randomisation. This represents a change from plans initially reported in the protocol (to use classical school-level clustered standard errors, which will be reported as a robustness test in an appendix) but provides a clearer basis for meaningful inference about how likely our results were due to chance.

The vector of school- and pupil-level covariates included in the model will be as follows:

- gender,
- whether English is an Additional Language,
- eligibility for free school meals,
- baseline writing task scores (PIE short score and baseline sentence combining score),
- means of all of the above at the class/school level,
- school's average KS1 points score from 2018/19.

Secondary outcome analysis

We will conduct one secondary outcome analysis:

[2] Athey, S., & Imbens, G. W. (2017). The Econometrics of Randomized Experiments. In Handbook of Economic Field Experiments (Vol. 1, pp. 73–140). Elsevier. https://doi.org/10.1016/bs.hefe.2016.10.003

- Sentence combining: Same as the primary outcome analysis modelling except replace Y_{ij} with the end of project sentence combining score.

Sensitivity analyses

We will re-estimate the primary analysis model following the advice of Lin (2013)[3] and Negi & Wooldridge (2020)[4] in appropriate specification of impact evaluation models. This includes the treatment indicator as a main variable, then includes all covariates de-meaned, and includes a full set of interactions between the demeaned covariates and the treatment indicator.

$$Y_{ij} = \alpha + \beta_1 Treat_j + \boldsymbol{\gamma}' \widetilde{\boldsymbol{X}}_{ij} + \delta \widetilde{\boldsymbol{X}}_{ij} * \boldsymbol{Treat}_j + \eta_j + \varepsilon_{ij}$$

where $\widetilde{\boldsymbol{X}}_{ij}$ are the de-meaned model covariates included in the primary outcome ITT model, and $\delta \widetilde{\boldsymbol{X}}_{ij} * \boldsymbol{Treat}_j$ are the interactions between these de-meaned covariates and the treatment variables. Our primary outcome of interest remains β_1, which may recover a more precisely estimated intention to treat effect in the presence of treatment heterogeneity. In other respects, the specification and method of analysis remains the same as the primary outcome ITT model.

Robustness analyses

Due to having to pause and re-start the trial due to COVID-19, the intervention is being implemented with a different cohort of children than initially anticipated. This means that pre-test data had to be recollected which, in the case of those who had already been randomised, the baseline data were collected unblinded. Furthermore, quite a few schools dropped out due to the COVID-19 disruption. We check the robustness of our findings as follows:

- Removing the main pupil-level and school-level average baseline scores from the primary analysis and replacing them with school-level average baseline scores from ahead of initial randomisation (for Batch 1 and 2 only) (may be replaced with school-level average baseline scores of a random sub-sample of these if there are marking capacity constraints)
- Carrying out Batch 3 sub-group analysis only (further detail reported under sub-group analyses below)

We will consider these additional analyses to cast doubt on our primary analysis where the impact estimates from these additional models is statistically significantly different from the primary analysis model (primary analysis estimate is not within the 95% confidence interval of alternative model and alternative analysis estimate is not within the 95% confidence interval of the primary model). Should such a difference occur then we will carry out further investigation to understand the reason for the difference.

Graphical analyses

We will plot the primary and secondary outcome measures by treatment status using overlapping histograms. This will be done:

[3] Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. The Annals of Applied Statistics, 7(1), 295–318. https://doi.org/10.1214/12-AOAS583

[4] Negi, A., & Wooldridge, J. M. (2020). Revisiting regression adjustment in experiments with heterogeneous treatment effects. Econometric Reviews, 1–31. https://doi.org/10.1080/07474938.2020.1824732

- Unconditional: simple overlapping histograms of raw primary and secondary outcome measures
- Adjusted: overlapping histograms of primary and secondary outcome measures which have been adjusted using the primary analysis model covariates in order to align this graphical analysis with the statistical analysis model. Specifically, this will be constructed by estimating the following linear regression model and predicting the residuals: $Y_{ij} = \alpha + \boldsymbol{\gamma}'\boldsymbol{X}_{ij} + \eta_j + \varepsilon_{ij}$ (i.e., the primary analysis model without the treatment indicator)

Sub-group analyses

We will carry out the following sub-group analyses:

- Batch 3 only (given large delays to batches 1 and 2, and re-collection of pre-test measures post-analysis)
- Whether English is an Additional Language (measured using school-reported pre-randomisation data collection)
- Eligibility for Free School Meals (measured using school-reported pre-randomisation data collection)

For the Batch 2 sub-group analysis, we will estimate the same analysis model as for the primary analysis, except restricted only to pupils in schools that were randomised as part of batch 3.

For all other sub-group analysis, we will estimate the differential effect for this sub-group using an interaction of the treatment and an indicator variable for the sub-group of interest using the following linear regression model:

$$Y_{ij} = \alpha + \beta_1 Treat_j + \beta_2 SubGroup_{ij} + \beta_3 Treat_j * SubGroup_{ij} + \boldsymbol{\gamma}'\boldsymbol{X}_{ij} + \eta_j + \varepsilon_{ij}$$

where $SubGroup_{ij}$ is an indicator variable for the sub-group of interest. In other respects, the specification and method of analysis remains the same as the primary outcome ITT model. The primary coefficient of interest is β_3, i.e. on the interaction between the treatment and the sub-group of interest, which recovers the difference between the overall treatment effect and the sub-group treatment effect. We will also report $\beta_1 + \beta_3$ which recovers the sub-group treatment effect.

Missing data

Missingness analysis will be used as sensitivity analysis. We will base confirmation of the effectiveness of the treatment on complete case analysis only but assess the sensitivity of the estimate to missingness using the estimates from these additional analyses. If the complete case analysis model implies effectiveness but the missingness sensitivity analysis estimate does not then we must assume that the missing data is missing not at random to such an extent as to invalidate our conclusion of effectiveness, which we would state in the reporting of the evaluation.

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

In the case of missing data for model covariates, in line with Groenwold et al. (2012)[5] we propose to use a missing indicator strategy as the most appropriate way to handle missing

[5] Groenwold, R. H. H., White, I. R., Donders, A. R. T., Carpenter, J. R., Altman, D. G., & Moons, K. G. M. (2012). Missing covariate data in clinical research: When and when not to use the missing-

predictor data in the context of a randomised controlled trial while respecting the principle of intention to treat. We will check the robustness of our findings to carrying out the following approach (simultaneously for all covariates with missingness). For all model covariates, we will replace missing values with their mean value (for continuous variables) or their modal value (for categorical values) and include an additional indicator variable which is set to 1 when this covariate is missing and 0 otherwise.

In the case of missing primary outcome data which exceeds 10% of sample as randomised, we will estimate a probit regression model of whether the outcome variable is missing. We will then check the robustness of our findings to adding covariates to the primary analysis model which are statistically significant predictors in this model of primary outcome missingness.

If both missing covariate data and missing outcome data occur, these strategies will be combined into a single model i.e. with additional covariates that predict missing outcomes and a missing indicator strategy for missing covariates.

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, we will obtain 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/implementation index, where the number of successfully delivered classes in the sequence will be standardised to have mean zero and standard deviation one to aid interpretation;
2. As a binary measure, where classes in which teachers 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.

For the continuous measure:

We will use instrumental variables analysis to estimate the effect of increasing implementation by one standard deviation. We will estimate the instrumental variables model using two stage least squares (2SLS) regression by estimating a (first stage) model of the implementation index, as follows:

$$Index_j = \alpha + \beta_1 Treat_j + \boldsymbol{\gamma}' \boldsymbol{X}_{ij} + \eta_j + \varepsilon_{ij}$$

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

$$Y_{ij} = \alpha + \beta_1 \widehat{Index}_j + \boldsymbol{\gamma}' \boldsymbol{X}_{ij} + \omega_{ij}$$

where $\widehat{Index}_j$ are the predicted values of treatment receipt derived from the first stage model, and ω is an error term. Our primary outcome of interest will be β_1, which should recover the effect of a one standard deviation increase in implementation (i.e., a one standard deviation increase in the number of lessons in the sequence successfully delivered). We will conduct

indicator method for analysis. Canadian Medical Association Journal, 184(11), 1265. https://doi.org/10.1503/cmaj.110977

this analysis using the ivregress functionality of Stata to make necessary adjustments to standard errors (which will also be clustered at school level) due to the instrumental variables approach (note that we retain classical inference for this analysis).

Example syntax for this model is reported in the analysis syntax appendix.

For the binary measures:

We will use Complier Average Causal Effect (CACE)[6] analysis to estimate intervention effects on treated children. We will estimate the CACE using two stage least squares (2SLS) regression by estimating a (first stage) model of compliance, as follows:

$$Comply_j = \alpha + \beta_1 Treat_j + \boldsymbol{\gamma}' \boldsymbol{X}_{ij} + \eta_j + \varepsilon_{ij}$$

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

$$Y_{ij} = \alpha + \beta_1 \widehat{Comply}_j + \boldsymbol{\gamma}' \boldsymbol{X}_{ij} + \eta_j + \omega_{ij}$$

where $\widehat{Comply}_j$ are the predicted values of treatment receipt derived from the first stage model, and ω is an error term. Our primary outcome of interest will be β_1, which should recover the effect of the intervention among compliers. We will conduct this analysis using the ivregress functionality of Stata to make necessary adjustments to standard errors (which will also be clustered at school level) due to the instrumental variables approach (note that we retain classical inference for this purpose).

Example syntax for this model is reported in the analysis syntax appendix.

Effect size calculation

Cohen's d effect size will be calculated as follows:

$$d = \frac{\overline{x_1} - \overline{x_2}}{\widehat{s*}}$$

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

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

where n_1 is the sample size in the control group, n_2 is the sample size in the treatment group, s_1 is the standard deviation of the control group, and s_2 is the standard deviation of the treatment group (all estimates of standard deviation used are unconditional).

Ninety-five per cent confidence intervals (95% CIs) of the effect size will be estimated by inputting the upper and lower confidence limits of $\widehat{\beta_1}$ from the regression model into the effect size formula.

[6] Gerber AS, Green DP. (2012). Field Experiments: Design, analysis and interpretation. WW Norton and Company, New York.

Appendix: Analysis Syntax

In this appendix, we provide indicative analysis syntax to implement key statistical models specified in the Statistical Analysis Plan using Stata. Eventual syntax may have small changes (e.g., variable name changes) that do not affect the syntax's implementation of the models specified above. Variables are as specified in the Statistical Analysis Plan.

ICC Estimation

```
xtset schoolid

xtreg piw_long

local piw_icc : di %7.2fc e(sigma_u) / (e(sigma_u) + e(sigma_e))
```

is a hierarchical linear regression model with school-level random intercepts (defined by the `schoolid` is a school identifier) estimated on individual-level full randomised sample data, where `piw_long` is the Progress in Writing Long Task raw score.

This will also be repeated for the secondary outcome measure (`sc_outcome`).

Pre-test/post-test variance explained

```
xtset schoolid

xtreg piw_long piw_short

local piw_bivar_r2_pupil : di %7.2fc e(r2_w)

local piw_bivar _r2_school : di %7.2fc e(r2_b)

local piw_bivar _r2_overall : di %7.2fc e(r2_o)
```

is a hierarchical linear regression model with school-level random intercepts (defined by the `schoolid` is a school identifier) estimated on individual-level full randomised sample data, where `piw_long` is the Progress in Writing Long Task raw score and `piw_short` is the Progress in Writing Short Task raw short (pre-test).

This will also be repeated for the Sentence Combining task secondary outcome measure (`sc_outcome`) with the pre-test measure replaced with the Sentence Combining baseline task (`sc_baseline`).

Baseline/outcome variance explained

```
xtset schoolid

xtreg piw_long `xs'

local piw_r2_pupil : di %7.2fc e(r2_w)

local piw_r2_school : di %7.2fc e(r2_b)

local piw_r2_overall : di %7.2fc e(r2_o)
```

is a hierarchical linear regression model with school-level random intercepts (defined by the `schoolid` is a school identifier) estimated on individual-level full randomised sample data, where `piw_long` is

the Progress in Writing Long Task raw score and `xs` is a local macro of baseline covariates included in the primary analysis model to improve the precision of the treatment estimate (including pre-test measures).

This will also be repeated for the Sentence Combining task secondary outcome measure (`sc_outcome`).

Primary intention to treat (ITT) analysis:

```
ritest treat _b[treat], r(2000) strata(stratum) cluster(schoolid) seed(TBA): ///

      regress piw_long i.treat `xs' i.stratum
```

is a linear regression model estimated on individual-level full randomised sample data, where `piw_long` is the Progress in Writing Long Task raw score (corresponding to Y in the regression equation), `treat` is a binary treatment variable (corresponding to $Treat$ in the regression equation), `xs` is a local macro of baseline covariates included to improve the precision of the treatment estimate (corresponding to $\boldsymbol{X}_{ij}$ in the regression equation, and including pre-test measures), `stratum` is a categorical randomisation stratification variable (corresponding to η_j in the regression equation), and `schoolid` is a school identifier (corresponding to j in the regression equation).

Sensitivity analysis:

```
local xsint ""

foreach x in `xs' {

cap drop `x'_dm

sum `x'

gen `x'_dm = `x' - r(mean)

local xsint "`xsint' `x'_dm c.`x'_dm#1.treat"

}

ritest treat _b[treat], r(2000) strata(stratum) cluster(schoolid) seed(TBA): ///

      regress piw_long i.treat `xs' `xsint' i.stratum
```

is a linear regression model estimated on individual-level full randomised sample data, where `piw_long` is the Progress in Writing Long Task raw score (corresponding to Y in the regression equation), `treat` is a binary treatment variable (corresponding to $Treat$ in the regression equation), `xs` is a local macro of baseline covariates included to improve the precision of the treatment estimate (corresponding to $\boldsymbol{X}_{ij}$ in the regression equation, and including pre-test measures), `stratum` is a categorical randomisation stratification variable (corresponding to η_j in the regression equation), and `schoolid` is a school identifier (corresponding to j in the regression equation).

Graphical analysis:

```
regress piw_long `xs' i.stratum
```

```
cap drop piw_long_analysis_resid

predict piw_long_analysis_resid, resid

twoway (histogram piw_long if treat==1, freq color(black%50)) ///

       (histogram piw_long if treat==0, freq fcolor(none) lcolor(black)) ///

       , legend(order(1 "Treatment" 2 "Control")) xtitle("Unadjusted outcome
score") ytitle("Density")

twoway (histogram piw_long_analysis_resid if treat==1, freq color(black%50)) ///

       (histogram piw_long_analysis_resid if treat==0, freq fcolor(none)
lcolor(black)) ///

       , legend(order(1 "Treatment" 2 "Control")) xtitle("Adjusted outcome score")
ytitle("Density")
```

plots the raw values of the Progress in Writing Long Task raw score (`piw_long`) in two overlapping histograms by treatment status, and the adjusted values of `piw_long` obtained from a linear regression model estimated on individual-level full randomised sample data, where `piw_long` is the Progress in Writing Long Task raw score is regressed on `xs` (a local macro of baseline covariates included to improve the precision of the treatment estimate, corresponding to $\boldsymbol{X}_{ij}$ in the regression equation, and including pre-test measures), and `stratum` (a categorical randomisation stratification variable, corresponding to η_j in the regression equation), from which the residual `piw_long_analysis_resid` is predicted, and plotted in a histogram by the two values of `treat`, which is the binary treatment variable.

Sub-group analysis:

```
ritest treat (_b[1.treat#1.subgroup])(_b[1.treat]+_b[1.treat#1.subgroup]), r(2000)
strata(stratum) cluster(schoolid) seed(TBA): ///

       regress piw_long i.treat i.subgroup 1.treat#1.subgroup `xs' i.stratum
```

is a linear regression model estimated on individual-level full randomised sample data where `piw_long` is the Progress in Writing Long Task raw score (corresponding to Y in the regression equation), `treat` is a binary treatment variable (corresponding to $Treat$ in the regression equation), subgroup is an indicator variable for membership of the sub-group of interest (corresponding to $SubGroup$ in the regression equation), `xs` is a local macro of baseline covariates included to improve the precision of the treatment estimate (corresponding to $\boldsymbol{X}_{ij}$ in the regression equation, and including pre-test measures), `stratum` is a categorical randomisation stratification variable (corresponding to η_j in the regression equation), and `schoolid` is a school identifier (corresponding to j in the regression equation).

Implementation Index analysis:

```
ivregress 2sls piw_long `xs' i.stratum (index = treat), ///

       vce(cluster schoolid)
```

is an instrumental variables (two stage least squares) regression model estimated on individual-level full randomised sample data where `piw_long` is the Progress in Writing Long Task raw score (corresponding to Y in the regression equation), `treat` is a binary treatment variable (corresponding to $Treat$ in the regression equation), `index` is a continuous indicator of school compliance defined in

the main body of the SAP, `xs` is a local macro of baseline covariates included to improve the precision of the treatment estimate (corresponding to $\boldsymbol{X}_{ij}$ in the regression equation, and including pre-test measures), `block` is a categorical randomisation stratification variable (corresponding to η_j in the regression equation), and `schoolid` is a school identifier (corresponding to j in the regression equation).

CACE analysis:

```
ivregress 2sls piw_long `xs' i.stratum (comply = treat), ///
       vce(cluster schoolid)
```

is an instrumental variables (two stage least squares) regression model estimated on individual-level full randomised sample data where `piw_long` is the Progress in Writing Long Task raw score (corresponding to Y in the regression equation), `treat` is a binary treatment variable (corresponding to $Treat$ in the regression equation), `comply` is an indicator of school compliance defined in the main body of the SAP, `xs` is a local macro of baseline covariates included to improve the precision of the treatment estimate (including pre-test measures), `stratum` is a categorical randomisation stratification variable (corresponding to η_j in the regression equation), and `schoolid` is a school identifier (corresponding to j in the regression equation).