Tips by Text
Statistical Analysis Plan

Evaluator (institution): IES and NIESR
Principal investigator(s): Anneka Dawson and Lucy Stokes

Education Endowment Foundation

PROJECT TITLE	Tips by Text, a two-armed randomised trial
DEVELOPER (INSTITUTION)	Behavioural Insights Team
EVALUATOR (INSTITUTION)	IES and NIESR
PRINCIPAL INVESTIGATOR(S)	Anneka Dawson and Lucy Stokes
SAP AUTHOR(S)	Anneka Dawson, Lucy Stokes, Ceri Williams, David Wilkinson
TRIAL DESIGN	Two-arm individual level randomised controlled trial
TRIAL TYPE	Efficacy
PUPIL AGE RANGE AND KEY STAGE	EYFS Reception – Year 1
NUMBER OF SCHOOLS	109
NUMBER OF PUPILS	2,646
PRIMARY OUTCOME MEASURE AND SOURCE	York Assessment of Reading for Comprehension
SECONDARY OUTCOME MEASURE AND SOURCE	Strengths and Difficulties Questionnaire

SAP version history

VERSION	DATE	REASON FOR REVISION
1.1	22/4/2021	Updated to note inclusion of Bayesian analysis as an additional analysis, prompted by fewer than expected completed post-tests as a result of Covid-19.
1.0 [*original*]	5/10/2020	*N/A*

Table of Contents

Introduction

Tips by Text is a text message curriculum developed by Professor Susanna Loeb and colleagues (Director, Annenberg Institute and Professor in Education and International and Public Affairs, Brown University) which aims to improve the developmental outcomes of young children. The curriculum has been adapted to the UK context by the Behavioural Insights Team (BIT).

In this evaluation, we aim to test the impact of Tips by Text compared to business as usual using a two-arm randomised controlled trial (RCT). The text messages are designed to provide parents of children in reception year (4-5 years old) with information and strategies to help their children's development outside of the school environment. Three types of messages are sent every week including during the school holidays which parents can tailor to work with their children. The three types of messages are: "FACT" texts, designed to inform and motivate parents by highlighting the importance of a particular skill or set of skills; "TIP" texts, designed to minimise the cognitive, emotional, and time burdens of engaged parenting by include short, simple, and highly-specific activities for parents to do with their children that build on existing family routines; and "GROWTH" texts, which provide encouragement, reinforcement, and extend the TIP texts.

Ultimately 109 schools from the North-East of England were recruited to participate in the trial. Parents of reception children in these schools were randomly allocated to one of two groups; (1) the control group (no text messages) or (2) the treatment group who receive three text messages per week. It was intended that parents allocated to the treatment group would receive text messages from early November 2019 to late July 2020 (9 months in total). In practice, the period of text message delivery has been extended to end October 2020 as a result of the Covid-19 crisis so the children will have moved into Year 1 and therefore be aged 5-6 by the post-test (please see updates to the trial protocol for further details).

The primary analysis will examine the impact of Tips by Text on children's literacy outcomes using the York Assessment for Reading Comprehension (YARC) and then secondary analysis will explore socio-emotional outcomes using the Strengths and Difficulties Questionnaire.

The primary research question is therefore:

RQ1. Does the Tips by Text intervention improve Reception children's literacy outcomes?

The secondary research questions are:

RQ2. Does the Tips by Text intervention improve Reception children's social development skills?
RQ3. Does the Tips by Text intervention improve literacy outcomes differentially for children eligible for Free School Meals (FSM)?
RQ4. Does the Tips by Text intervention improve literacy outcomes differentially for children with low and high baseline attainment?

The trial protocol identified two additional secondary research questions:

- Does the Tips by Text intervention improve Reception children's numeracy outcomes?

- Does the Tips by Text intervention improve Reception children's language and communication outcomes?

However, these two research questions will no longer be explored within the trial, as they relied on data from the Early Years Foundation Stage Profile, which, in line with other national assessments, has no longer been conducted for this cohort of children given the Covid-19 crisis.

Design overview

Please ensure all details are in line with the latest version of the protocol.

Trial design, including number of arms		Two-arm randomised controlled trial
Unit of randomisation		Parent
Stratification variables (if applicable)		None
Primary outcome	variable	Literacy attainment
	measure (instrument, scale, source)	YARC total score, YARC Early Reading, GL Assessment
Secondary outcome(s)	variable(s)	Socio-emotional outcomes
	measure(s) (instrument, scale, source)	Total Difficulties Score, Strengths and Difficulties Questionnaire
Baseline for primary outcome	**variable**	Literacy attainment
	measure (instrument, scale, source)	YARC score (YARC Early Reading sound deletion and sound isolation subscales), GL Assessment
Baseline for secondary outcome	**variable**	-
	measure (instrument, scale, source)	-

In collaboration with the delivery team, it was agreed that the trial would use individual level randomisation (at the parent level, in order to avoid a scenario whereby twins are assigned to different groups) to reduce the number of schools needed for the trial and to maximise the power to detect an effect. The trial has two arms so that half the eligible children's parents within participating schools were allocated to the treatment arm (receiving the programme) and half to the control arm (who do not receive the programme). All reception classes were randomised, so that for each reception class half were randomly allocated to receive the intervention and half do not. Although we cannot rule out the possibility of contamination of the control group from the sharing of Tips by Text content, it is expected by the delivery team

that the number of parents who speak to each other about the text messages, or directly share the messages with one another, will be limited.

Randomisation was conducted using simple randomisation, within classes (or effectively within schools where there was only one reception class). As the number of schools recruited exceeded the delivery target, as per the trial protocol, in schools with multiple reception classes, only one class was selected (at random) to complete the pre-tests. This means that while a total of 2,646 pupils were in classes eligible for testing, a total of 3,662 pupils form part of the overall trial. Pupils in the classes that were not selected for testing were still randomised; the original intention being that outcomes from the Early Years Foundation Stage Profile (EYFSP) would be analysed as secondary outcomes for the full sample of 3,662 pupils. As documented in the protocol revisions (version 1.1), analysis of EYFSP outcomes is no longer possible, as in line with other national assessments, these assessments did not take place in summer 2020 due to the Covid-19 pandemic. While beyond the scope of the current planned analysis, this larger sample could potentially be followed up in future longitudinal analysis.

Due to a small number of schools having an earlier October half-term (2019), randomisation took place in two batches, enabling parents in these schools to start receiving the texts immediately following half-term. Five schools were randomised in the first batch (20th October 2019) and the remaining 104 schools in the second batch (30th October 2019).

Sample size calculations overview

		Protocol		Randomisation	
		OVERALL	FSM	OVERALL	FSM
Minimum Detectable Effect Size (MDES)		0.10	0.14	0.10	0.15
Pre-test/ post-test correlations	level 1 (pupil)	0.4	0.4	0.4	0.4
	level 2 (class)	-	-	-	-
	level 3 (school)	-	-	-	-
Intracluster correlations (ICCs)	level 2 (class)	-	-	-	-
	level 3 (school)	0.10	0.10	0.10	0.10
Alpha		0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8
One-sided or two-sided?		2	2	2	2
Average cluster size		26	10	24	9
Number of schools*	intervention	105	105	109	109
	control	105	105	109	109
	total	105	105	109	109
Number of pupils	intervention	1365	525	1325	510
	control	1365	525	1321	508
	total	2730	1050	2646	1018

*Note that as randomisation took place within schools (classes), all schools form part of both the intervention and control group.

The sample size was determined with the aim of establishing a MDES of 0.1, given the previous research by York and Loeb (2018), which found an impact of the programme on literacy of 0.11 standard deviations. MDES calculations were conducted using Optimal Design software.

All estimates are based on standard EEF assumptions of 80% power and 5% significance level. We assume the pre-test explains 40 per cent of the variation in the post-test scores[1], and that 10 per cent of variance is explained by the blocking variable[2] (or effectively here, an ICC of 0.10). To achieve an MDES of 0.1 this suggested a required sample size of around 105 schools based on an average of 26 pupils per school (assuming an average of 30 pupils per school, and then allowing for around 5 per cent of pupils to withdraw before randomisation and around 10 per cent of the remainder lost to follow-up). As set out in the trial protocol, if it proved feasible to recruit around 105 schools (and ultimately 109 were recruited), then only 1 class per school receives the assessment that forms the primary outcome for the trial.

Department for Education Statistics for January 2019 show that 15.8 per cent of pupils in primary schools were eligible for and claiming free school meals (Department for Education, 2019). However, schools with above average proportions of pupils eligible for free school meals (FSM) were targeted in recruitment. On the assumption that on average 10 pupils within the reception year of a school will be eligible for FSM, equivalent to around 38 per cent of pupils (at the time of writing the protocol, this was the average percentage of pupils eligible for FSM in recruited schools), and keeping all other assumptions the same, this results in an MDES of 0.14. The trial is not however designed with the aim of detecting an effect for the FSM subgroup as the primary population of interest.

At the point of randomisation, 109 schools were recruited and 2,392 pupils had completed pre-tests. In total, 2,646 pupils were in the classes selected for the pre-test, and were randomised; all of these pupils will be approached for post-test.[3] This means the maximum available sample at post-test is 2,646 (note this excludes 2 pupils who were withdrawn from the study after randomisation, based on the available information at the point of preparing this SAP). Based on these numbers, and keeping all other assumptions the same as at protocol stage, this gives an MDES of 0.10. There will inevitably be some degree of attrition from the trial by the point of analysis. If we assume that ultimately 100 schools remain in the trial, with an average cluster size of 20, the MDES stands at 0.11. For the FSM subgroup, based on post-randomisation numbers, the MDES stands at 0.15. Given the likely difficulties in conducting assessments as a result of Covid-19, it is probable that attrition will be higher; with attrition of around 20 per cent of schools, this would result in an MDES of 0.12. Thus even with fairly substantial attrition, the trial would still be able to detect a relatively small effect. To help place this in context, in line with EEF guidance, an effect size of between 0.10 and 0.18 is considered to be equivalent to two months' additional progress.

[1] To our knowledge, there is currently no publicly available information on the likely correlation between pre- and post-test scores for our primary outcome measure, the York Assessment of Reading for Comprehension. The efficacy trial of the Nuffield Early Language Intervention (implemented with a similar age group) found that around 55% of the variation in post-test scores (a composite language score) was explained by the pre-test and pupil characteristics (Sibieta, 2016). This would be higher due to the additional inclusion of pupil characteristics; we therefore use a lower estimate of 40 per cent for our assumption.

[2] As noted earlier, randomisation was conducted within schools, or blocks. The sample size calculations require an estimate of the percentage of the variance explained by the "blocking variable", in this case, the percentage of the variance explained by the school (Spybrook et al., 2011). Although based on secondary schools, previous research has indicated that schools explain around 10 per cent of the variation in pupil attainment at the end of Key Stage 4 (Wilkinson et al., 2018)

[3] Except in the event of further withdrawals. Pupils who have subsequently left their school will also not be tested.

Analysis

The estimated impact will be based on the difference between those assigned to the treatment and control groups, regardless of contamination of the control group or drop out. This is in order to estimate the "intention-to-treat" (ITT) effect. In addition, the regression models used for the primary analysis will include controls for prior attainment (that is, using the YARC total score as captured at pre-test), as well as a dummy variable to indicate those schools which were randomised in the first batch.

Estimated impact in terms of pupil's outcomes will be converted into a Hedges' g effect size. This will use the estimated total pooled standard deviation of the treatment and control groups, rather than the within-school pooled standard deviation as the former is a more conservative approach.

For each measure, we will estimate outcomes using a linear regression model including a dummy variable indicating trial arm allocation, and school fixed effects. We prefer this approach over multi-level analysis which relies on an assumption that random effects are independent of regressors (Ebbes, 2004).

Primary outcome analysis

The primary outcome to be assessed in this trial is literacy, as measured by the York Assessment of Reading for Comprehension (YARC). The version of the YARC suitable for 4-7 year olds is called Early Reading and covers four dimensions: sound isolation, sound deletion, letter sound knowledge, and early word recognition. To reduce testing time and burden on the school, only the former two measures were used as a pre-test as they are considered the most sensitive and more appropriate for the younger age of the children by GL Assessment and then all four dimensions are to be used at post-test. The primary outcome will be a "total" YARC score, where we will standardise each of the components to have a mean of zero and standard deviation of one. These will then be added together to create a composite measure and re-standardised (calculated where all four components are non-missing). We propose this approach, rather than simply summing together the four scales, given the subscales are not of equal length and thus a simple sum would give greater weight to the longer subscales. The same approach will be adopted with the two subscales used for the pre-test.

As there is just one primary outcome no adjustments for multiple comparisons will be applied.

The equation to be estimated is:

$$Y_{ijt} = \alpha + \beta_1 Treat_i + \beta_2 Y_{ijt-1} + \beta_3 \gamma_j + s_j + \varepsilon_{ij}$$

where i are pupils and j are schools, Y_{ijt} is the YARC post-test score, $Y_{ijt-1}$is the YARC pre-test score, $Treat_i$ is our treatment indicator (a dummy variable where 1 represents being allocated to receive the intervention and 0 represents allocation to the control group, i.e. not receiving the intervention), γ_j is a dummy variable indicating randomisation batch, s_j represent school fixed effects and ε being an error term. Where pre-test scores are missing, these will be imputed using the principles set out in the missing data section. As stated above, 2,392 of the 2,646 pupils in the classes eligible for testing completed a pre-test (this can be missing, for example, where pupils were absent on the day of testing, where special educational needs meant the test wasn't suitable, and where it was not possible to return to the school to conduct

assessments in the time available[4]). All 2,646 pupils are being approached for post-testing in order to maximise the potential available sample (rather than restricting this to pupils who completed pre-tests only). Thus for some pupils with post-tests, pre-test scores will be missing.

As this is effectively a multi-site trial, school identifiers will be included as fixed effects. This is in line with the EEF Statistical Analysis Guidance for an efficacy trial.

The analysis will be conducted in Stata (version 15).

Secondary outcome analysis

The approach to the secondary outcome analysis will follow the same approach used for the primary outcome. We do not have a measure of the secondary outcome at baseline, but will include the YARC pre-test score as it is still expected that this should account for some of the variance. In all other respects the analysis will remain the same. Thus our equation to be estimated is:

$$YS_{ijt} = \alpha + \beta_1 Treat_i + \beta_2 Y_{ijt-1} + \beta_3 \gamma_j + s_j + \varepsilon_{ij}$$

where YS_{ijt} is the SDQ score, instead of the YARC post-test score (and all other elements are as defined above).

More precisely, our outcome measure will be the total difficulties score from the SDQ. This is a standard SDQ measure which is constructed from summing together responses on four of the five SDQ subscales, giving a score which ranges from 0 to 40. Following existing conventions in using the SDQ (Vaz et al., 2016), if at least three of the five items within a subscale are completed, any remaining missing scores will be replaced by the mean for that subscale. If fewer than three items are completed, no score will be computed for that subscale, and if any of the four subscales are missing, the total difficulties score will not be computed. The extent of missingness in these data, including by treatment and control arm, will be presented in the report. We will also run separate models for three components; 'internalising problems' (which combines the emotional symptoms and peer relationship problems scales, 10 items), 'externalising problems' (which combines the conduct problems and hyperactivity scales, 10 items) and the prosocial scale (5 items). This division of the scales has been shown to be more appropriate for low-risk or general population samples (Goodman et al., 2010).

Subgroup analyses

We will conduct separate analysis of the subgroup of pupils eligible for Free School Meals. Pupils eligible for FSM will be identified using the variable EVERFSM_6_P available from the NPD. This analysis will be conducted for both the primary and secondary outcome; all other subgroup and additional analyses will be conducted only for the primary outcome.

We will run analyses interacting treatment allocation with FSM status, as well as running separate models for the FSM and non-FSM subgroups. The same specification will be used as for the primary analysis. The effect size for FSM pupils will be presented on the basis of the subgroup model, as per the EEF statistical analysis guidance.

Given existing research has suggested differences in impact according to prior attainment (e.g. York and Loeb, 2018), as specified in the trial protocol, we will also conduct a separate subgroup analysis for “high” and “low” attaining pupils, based on the pre-test assessments. These groups will be defined by dividing the sample in half, based on the median score on the

[4] Further details of the reasons for missingness will be presented in the evaluation report.

pre-test, which follows the approach adopted by York and Loeb (2018). Again we will report results both using interaction effects and as separate subgroups. The same model specification will be used as for the primary analysis.

Additional analyses

At the time of preparing this SAP, it is unclear whether some testing may need to happen remotely, rather than in person, as a result of Covid-19 restrictions. If remote testing is used, we will run an additional model for our primary analysis that also includes a dummy variable that captures whether the assessment was conducted remotely or not (i.e. in order to check for mode effects). We will also explore the sensitivity of the results to running a model which excludes any pupils for whom the post-tests were conducted remotely. Another potential implication of Covid-19 is that it may prove necessary to extend the post-testing period. The main post-testing period is November-December 2020; if this extends into January 2021, we will run an additional model that includes a dummy variable for any assessments conducted in January, to assess whether there are any implications for the results of the extended testing period.

We will also run an additional model for the primary analysis that additionally includes school by treatment interaction effects as well as school effects. The inclusion of interaction effects allows for the possibility that the effect of the intervention varies across schools (Singh et al., 2019), and so it is useful to check the robustness of our main results allowing for this.

We will also run a model that excludes pre-test scores and simply regresses post-test scores on treatment indicator and randomisation batch, to assess the sensitivity of results to the inclusion of the pre-test. This is of particular relevance given the fact that there are indications of floor effects in the pre-test data, even though the subscales used at pre-test are intended to be suitable for this age group based on guidance from the test provider. We will also run the model restricting to the sample for which pre-test scores are available (i.e. without imputing pre-test scores where these are missing).

We will conduct exploratory analysis that examines separately whether there is an impact of the intervention on each of the four YARC subscales administered at post-test: letter sound knowledge, early word recognition, sound isolation, and sound deletion. This will follow the same model specification for the primary outcome, and adjust for the total YARC pre-test score in each case.

Analyses in response to missing data are discussed further below.

Update v1.1: Due to the third national lockdown and associated disruption for schools as a result of Covid-19, the decision was taken to stop post-testing in January 2021. This has resulted in a smaller number of completed post-assessments (around a third of the anticipated number). The smaller than anticipated sample size available for analysis inevitably has implications for the ability of the originally proposed analysis to detect a statistically significant effect (more generally, there is of course wider debate around the use of p-values to judge statistical significance, see for example, the discussion in Nuzzo, 2014). Bayesian methods can be used as an alternative means of evaluating effectiveness and may, though not in all circumstances, offer benefits when working with smaller sample sizes (McNeish, 2016).

While we will carry out the original analysis using frequentist methods as documented in the original version of this SAP, we will also undertake an additional exploratory analysis using a Bayesian approach. This will provide exploratory evidence into the probability that the

intervention has a positive effect. This additional analysis will be carried out for both primary and secondary outcomes, and will be estimated using the eefAnalytics package available within Stata (Vallis et al., 2021). For this analysis we will adopt non-informative priors; this is to ensure that the results are in large part driven by the data, rather than on a subjective judgement regarding priors (Uwimpuhwe et al., 2020). For consistency, we will adopt the same assumptions regarding non-informative priors as used in the analysis by Uwimpuhwe et al. (2020) which uses vague Gaussian priors ($N(0,10^6)$) for each of the regression parameters.The report will present the effect sizes and credible intervals estimated through the Bayesian approach; along with the posterior probability that the intervention has an effect size of at least 0.1. This follows the recommendation made by Uwimpuhwe et al. (2020) regarding a suitable threshold for assessing effectiveness of educational interventions, but is also consistent with the fact that the trial was designed with the aim of establishing a MDES of 0.1, given the previous research for the US by York and Loeb (2018), which found an impact of the programme on literacy of 0.11 standard deviations. For completeness, we will also report the probability that the intervention has an effect size of at least zero (that is, that it has a positive effect). The results from the frequentist analysis will remain the main results on which the report will focus, with the results from the Bayesian approach forming an additional exploratory analysis, to provide further insight into the sensitivity of results. Any substantive differences in the results obtained from the two approaches will be discussed in the evaluation report.

Longitudinal follow-up analyses[5]

A longitudinal follow-up analysis is beyond the scope of the current planned analysis. A longitudinal analysis could potentially be conducted in future to investigate whether any impact on longer-term outcomes is evident. Relevant outcomes for longitudinal analysis would be outcomes as measured in national assessments at KS1 (end Year 2) and KS2 (end of primary school). Any analysis of KS1 outcomes would need to be based on teacher assessments as schools are not required to submit information on test scores to DfE and thus only information on teacher assessments is available in the NPD. In addition, it could also be of interest to explore any impact on the phonics screening check usually undertaken at end Year 1. Any such analyses would be conducted according to the principles outlined in EEF Longitudinal Analyses Guidance.

Imbalance at baseline

We will report on the characteristics of the sample both at randomisation and at the point of analysis. This will include school characteristics (including Ofsted rating, school type (academy status), urban/rural location, number of pupils, percentage FSM pupils, percentage EAL pupils, percentage SEN pupils, and average school level performance at end Key Stage 2) and pupil characteristics (including age, gender, FSM eligibility, pre-test score).

We will present pupil characteristics on the basis of:

- all participating pupils as at the point of randomisation (regardless of whether in class selected for testing in schools with multiple classes).
- participating pupils at the point of randomisation in classes selected for testing
- pupils in the final analysis sample (that is, those with post-test YARC scores)

[5] Please see the longitudinal analysis guidance.

Reporting will follow the standard EEF template, with means and standard deviations reported for continuous variables and counts and percentages in each category given for categorical variables. In line with EEF guidance, differences in pupil-level pre-tests (i.e. the YARC total score) will be reported as effect sizes. We will report the correlation between pre and post-test scores and compare these against the assumptions made in the sample size calculations. Histograms of pre and post-test scores will also be presented.

We will assess balance by calculating absolute standardised differences (Imbens & Rubin, 2015) between the treatment and control groups and these will be presented in the report. Differences of greater than 10% will be considered as indicative of imbalance. If imbalance is observed we will run an additional sensitivity analysis incorporating any variables on which imbalance is present as additional covariates into the primary outcome model.

Missing data

We will report the number of complete cases (those without missing data). We will consider separately missingness in outcome data at post-test and missingness at pre-test, and will report the distribution of missing observations by treatment arm.

In the event of greater than 5% missing data at either cluster[6] or individual level we will conduct further investigation into the mechanisms of missingness.

Any multiple imputation undertaken in response to missing data will focus on the primary analysis model, in line with EEF guidance. The extent to which multiple imputation is appropriate depends on the missingness mechanism.

Missing post-test data

If more than 5% of outcome data (total YARC post-test score) is missing (that is, as a percentage of the randomised sample) we will investigate the extent to which baseline characteristics (at school and pupil level)[7] are correlated with missingness, using logistic regression, where the dependent variable is a binary indicator for missingness. If this shows significant associations with any of the characteristics, we would conduct an additional analysis including those covariates in the primary analysis model to assess the robustness of the main results.

Missing pre-test data

Again, where more than 5% of data is missing (that is, as a percentage of the randomised sample) we will undertake further investigation into the nature of missingness. This will follow the same approach outlined above of a logistic regression to explore the extent to which observed baseline characteristics are correlated with missingness.

Multiple imputation

Where the analysis above shows significant associations with the included characteristics, and where missingness does not depend on unobserved characteristics, the missing data are considered to be missing at random (MAR), and in this case, proceeding with multiple

[6] That is, if data are missing for 5% of clusters.
[7] The characteristics to be included in the model will be pupil age in months, gender, FSM status and pre-test score where available, and a school identifier.

imputation (MI) is appropriate (in practical terms we will proceed with MI if any significant associations with observed characteristics are established). This MI would use those variables shown to be significant in the logistic regression alongside an indicator of treatment arm and school, using ten imputations[8]. Where appropriate, MI will be explored for imputing both pre-test and post-test data. We will then compare the estimated treatment effect based on the imputed data with that from the primary analysis.

If, however, missing data depend on unobserved variables, the missing observations are missing not at random (MNAR), and this cannot be resolved through MI. In this case we would conduct further sensitivity analysis exploring the robustness of the results to differing specifications, including complete case analysis only; exclusion of the pre-test from the model, and using null imputation to impute missing pre-test scores.

Compliance

We will construct a measure of compliance based on information on whether parents opted out[9] of receiving the texts and information on delivery of text messages. Thus compliance is effectively determined at parent level (which is applied at pupil level in the analysis – in most cases this is effectively the same, except in the case of siblings). The delivery team will provide the evaluation team with information on which parents opted out of receiving text messages and when this occurred, as well as information on whether messages were successfully delivered. This will allow the evaluation team to construct a measure of compliance, which will take a value between 0 and 1, set proportionately according to the number of messages received as a proportion of total possible messages[10] (thus a value of 1 represents a parent who received all possible messages, and a value of 0.5 would represent a parent who received half of all messages).

We will use this information on compliance in conducting a Complier Average Causal Effect (CACE) analysis to estimate intervention effects on treated pupils. This will be conducted for the primary outcome only. An instrumental variable (IV) approach will be used. We will estimate the CACE using two stage least squares (2SLS) regression by estimating a (first stage) model of compliance, using the measure of compliance described above and group allocation as the IV. The predicted values from the first stage are then used in the estimation of a model of our outcome measure. As per EEF analysis guidance, results for the first stage will be reported alongside with the correlation between the instrument and the endogenous variable and results of a F test. This analysis will be conducted using the ivregress command in Stata, which allows the necessary adjustments to be made to standard errors in response to the clustered nature of the data.

Non-compliance will also be examined through the implementation and process evaluation.

[8] We propose 10 imputations given previous research indicating that at least a reasonable number of imputations (defined as 5 or more) should be performed (Hayati Rezvan et al., 2015), and as traditionally between 2 and 10 imputations were considered sufficient (Rubin, 1987).

[9] Note that for a small number of parents/pupils, while they did not actively opt-out, messages were stopped (for example on request of the school due to specific circumstances). For the purposes of compliance analysis this will be treated in the same manner as opt-out.

[10] The total number of possible messages is the same for all parents; even though parents randomised within the first batch began receiving messages one week earlier, the total number of possible messages received over the course of the intervention is the same.

Intra-cluster correlations (ICCs)

We will estimate the ICCs for the pre-test and post-test at school-level using empty hierarchical linear models including school-level random effects as follows:

$$Y_{ij} = \beta_0 + \eta_j + \varepsilon_{ij}$$

where Y_{ij} is the pre- or post-test of individual i in school j, β_0 is a constant term, η_j is a school-level random effect and ε_{ij} is an individual-level idiosyncratic error term. The ICC estimate is recovered as follows:

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

Effect size calculation

Effect sizes will be calculated using Hedges' g, following the standard approach for EEF trials as set out in the EEF analysis guidance. This will therefore be calculated as:

$$ES = \frac{(Y_T - Y_{C)adjusted}}{s*}$$

Where $(Y_T - Y_{C)adjusted}$ is the adjusted difference in means between the treatment and control groups as recovered from the regression model, and $s*$ is the pooled unconditional variance of the treatment and control groups. All relevant parameters will be provided in the report so that readers are able to compute alternative definitions of effect sizes.

Ninety-five per cent confidence intervals will be reported around the resulting effect, calculated by inputting the lower and upper confidence limits for the coefficient on the treatment variable from the regression model into the effect size formula.

References

Department for Education (2019) *Schools, pupils and their characteristics: January 2019*, Department for Education. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/812539/Schools_Pupils_and_their_Characteristics_2019_Main_Text.pdf

Ebbes, P., Bockenholt, U. and Wedel, M. (2004) 'Regressor and random-effects dependencies in multilevel models', *Statistica Neerlandica*, 58 (2), pp.161–178.

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 38, 581–586. http://dx.doi.org/10.1111/j.1469-7610.1997.tb01545.x

Goodman A, Lamping DL, Ploubidis GB (2010) When to use broader internalising and externalising subscales instead of the hypothesised five subscales on the Strengths and Difficulties Questionnaire (SDQ): data from British parents, teachers and children. *Journal of Abnormal Child Psychology*, 38, 1179-1191.

Hayati Rezvan, P., Lee, K.J. & Simpson, J.A. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. BMC Med Res Methodol 15, 30 (2015). https://doi.org/10.1186/s12874-015-0022-1

Imbens, G. W., and Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge University Press. https://doi.org/10.1017/CBO9781139025751

McNeish, D. (2016) On Using Bayesian Methods to Address Small Sample Problems, *Structural Equation Modeling: A Multidisciplinary Journal*, 23:5, 750-773, DOI: 10.1080/10705511.2016.1186549

Nuzzo, R. (2014) "Statistical Errors: P Values, the 'Gold Standard' of Statistical Validity, are not as Reliable as Many Scientists Assume." *Nature* 506 (7487): 150–152. doi:10.1038/506150a

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. New York: Wiley

Sibieta, L. (2016) *Nuffield Early Language Intervention. Evaluation Report and Executive Summary*. Education Endowment Foundation. Available at: https://www.nuffieldfoundation.org/wp-content/uploads/2016/02/EEF-Evaluation-report-Nuffield-Early-Language-Intervention.pdf

Spybrook, J., Bloom, H., Congdon, R., Hill, C., Martinez, A., Raudenbush, S. (2011) Optimal Design Plus Empirical Evidence: Documentation for the "Optimal Design" Software.

Uwimpuhwe, G., Singh, A., Higgins, S. and Kasim, A. (2020) Application of Bayesian posterior probabilistic inference in educational trials, *International Journal of Research & Method in Education*, DOI: 10.1080/1743727X.2020.1856067.

Vallis, D. Singh, A., Uwimpuhwe, G., Higgins, S. Xiao, Z., De Troyer, E. and Kasim, A. (2021). "EEFANALYTICS: Stata module for Evaluating Educational Interventions using Randomised Controlled Trial Designs," Statistical Software Components S458904, Boston College Department of Economics, revised 04 Apr 2021.

Vaz, S., Cordier, R., Boyes, M., Parsons, R., Joosten, A., Ciccarelli, M., Falkmer, M., & Falkmer, T. (2016). Is Using the Strengths and Difficulties Questionnaire in a Community Sample the Optimal Way to Assess Mental Health Functioning?. *PloS one*, 11(1), e0144039. https://doi.org/10.1371/journal.pone.0144039

Wilkinson, D., Bryson, A. and Stokes, L. (2018) Assessing the variance in pupil attainment: how important is the school attended? *National Institute Economic Review*, 243(1): R4-R16.

York, B. N. and Loeb. S. (2018). *One Step at a Time: The Effects of an Early Literacy Text Messaging Program for Parents of Preschoolers*, NBER Working Paper 20659. Available at: http://www.nber.org/papers/w20659