Statistical Analysis Plan
Maximising the Impact of Teaching Assistants

RAND Europe
Dr. Alex Sutherland

PROJECT TITLE	**Maximising the Impact of Teaching Assistants (MITA)**
DEVELOPER (INSTITUTION)	University College London-Institute of Education (UCL-IOE)
EVALUATOR (INSTITUTION)	RAND Europe & University of Cambridge
PRINCIPAL INVESTIGATOR(S)	Alex Sutherland
TRIAL (CHIEF) STATISTICIAN	Alex Sutherland
SAP AUTHOR(S)	Alex Sutherland (RAND), Janna van Belle (RAND), Sonia Ilie (Cambridge)
TRIAL REGISTRATION NUMBER	ISRCTN33648
EVALUATION PROTOCOL URL OR HYPERLINK	*https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Protocols/Round_10_-_Maximising_the_Impact_of_Teaching_Assistants_1.pdf*

SAP version history

VERSION	DATE	REASON FOR REVISION
1.0 [*original*]	11/10/2018	*Original version*

Table of Contents

Introduction

Description of the intervention

The aim of this efficacy trial is to assess whether 'Maximising the Impact of Teaching Assistants' (MITA) leads to improvements in pupil attainment in reading and maths. MITA is a whole school intervention that aims to better deploy teaching assistants (TAs) and improve pupil outcomes through: (i) higher quality TA-pupil interactions; (ii) improved classroom management and lesson planning; and (iii) allowing classroom teachers to work more with lower-achieving pupils. This is one of the first times a trial will test a whole school intervention aiming at improving how schools, teachers and TAs can improve the use of TAs in everyday classrooms.

The intervention consists of three levels of support: 1) training delivered to the Senior Leadership Team (SLT) in schools (two leaders from each school, including the head teacher) held in school 'clusters' throughout the year; 2) School visits from a National Leader of Education (NLE) (a practicing Senior leader) linked to the London Leadership Strategy ('LLS') who will provide support in identifying gaps in current practice and developing and implementing a change management plan; and 3) School training for all teachers and TAs on the 'scaffolding framework' focused on effective interactions with pupils.[1] The NLE consultant will provide continuous support to school staff between training sessions and this will also promote engagement with other elements of the intervention.

The training/support sessions that make up the intervention will be delivered across the course of the 2017/18 school year: there will be four half day SLT training sessions throughout the year, school visits from NLEs each term, and two half day training sessions for TAs and a twilight-length training for teachers delivered in spring term. In the following year, 2018/19, substantive changes developed during the training will be implemented by the schools.

The primary focus of the trial is the overall effect of this package on pupil attainment in reading and maths. The control condition will be 'business as usual'. This statistical analysis plan sets out how we will assess whether MITA leads to improvements on pupil reading and maths outcomes compared to 'business as usual'.

This project will test several hypotheses relating to the impact and delivery of MITA. Specifically that MITA:

1. has a positive effect on pupils' attainment, specifically:
    a) Better reading outcomes (vs controls) for Year 3 pupils.
    b) Better reading and mathematics outcomes (vs controls) for Year 6 pupils.
2. results in improved *deployment* of the school TA workforce.
3. results in change of school/classroom practices, specifically:
    a) Practices aimed at improved interactions between TAs and pupils.
    b) Practices aimed at fostering pupil independence.
4. has a positive effect on *pupils' engagement with learning*.

1 This framework is designed to support TAs to scaffold pupils' learning and foster independent learning where pupils are expected to self-scaffold while TAs observes their progress, intervening only when pupils show they are unable to proceed (Bosanquet et al, 2015).

Study design

The trial was planned as a stratified, two-arm, cluster-randomised controlled trial (cRCT), across 100 schools recruited from four geographical regions: 1) West Midlands; 2) Portsmouth, Havant, Fareham and Hampshire; 3) Barking & Dagenham, Redbridge, Havering, Thurrock and London; and 4) Suffolk. All Year 1 and Year 4 pupils are eligible for inclusion in the trial (see below).

Based on assumptions about class sizes (roughly 33 pupils per class in primary school) and the actual number of schools recruited (n=128), the study should involve around 13,000 pupils distributed over 128 schools and 400 classes, with 130 pupils per school.

The following eligibility criteria for participating schools applied:

- Primarily two or three form entry primary or junior school (one and four form entry schools will also be allowed, but we expect a small minority of schools will be in this category).
- No prior engagement in MITA and/or Maximising the Practice of Teaching Assistants MPTA training sessions.
- No substantial prior action(s) taken following recommendations from the EEF TA guidance or MITA or MPTA handbooks.
- Not in special measures OR facing imminent leadership changes.

Given the limited training resource, school recruitment will be on a 'first come first served basis' – meaning that schools will be accepted on the trial once they have completed the required paperwork and prerequisite tasks (provision of pupil UPNs, list of teachers and TAs, signing of a Memorandum of Understanding (MoU)). All schools who fulfil the inclusion/exclusion criteria and who volunteer for the trial will be eligible for the intervention.

Schools will be assigned to either treatment (MITA) or control (business as usual). All schools signing up will have a 50% chance to be assigned to the treatment group within the school geographical clusters and attainment profile (see randomisation section below). Given that this is a whole school intervention, all teachers and TAs (including those with different role titles, but ostensibly working in pupil/classroom support roles) in treatment schools, across all years, will be eligible for and will receive the intervention. Participating schools will be asked to sign a MoU that will outline the roles and responsibilities of all stakeholders involved and clearly set out the requirements for schools.

There is only one treatment condition in this trial: throughout the 2017/18 academic year the school will receive training and support to improve deployment of TAs. For control schools, it will be business as usual during the school year 2017/18. As an incentive, control schools will also receive additional continuing professional development (CPD) training during the summer of 2018, along with a one-off payment of £750, and a package of school supplies (books, games, etc.).

The **baseline data** for the year 3 pupils consists of an external reading test (ELMS Reading – see outcome measure section), which has been independently administered and invigilated by a third party provider (ACER) in June/July 2017, when these pupils were in year 1. For the year 6 pupils the baseline data consist of Key Stage 1 reading and mathematics results which reflect performance of these pupils at the end of year 2, two years before randomisation (NPD variables will be KS1_READWRITPOINTS and KS1_MATPOINTS).[2] Both baseline measures will be standardised for each year respectively before being pooled for analysis.

[2] Please note that Levels were still being used at this point, so the baseline and outcome measures for the older cohort will differ in that respect.

The **primary outcome** will be pupil attainment in Reading at the time of follow-up (in then year 3 pupils) as measured by an external reading test (again ELMS - Reading), and Maths and Reading at the time of follow-up (in the then Year 6 pupils) as measured by maths and reading results in the Key Stage 2 test. **Table 1** gives and overview of primary baseline and outcome measures.

Table 1: summary of primary baseline and outcome measures.

Year group	2016/17 (Baseline year)	2018/19 (Outcome year)
	Baseline (end of year 1 and 4)	Outcome end of year 3 and 6
Younger cohort (Start Yr1)	Reading - ACER, ELMs test	Reading – ACER, ELMs test
Older cohort (Start Yr4)	Maths/Reading (Key Stage 1)	Maths/Reading (Key Stage 2)

There are two **secondary outcomes** at different levels: i) pupil engagement (pupil level), and ii) change in teacher/TAs practice (teacher/TA level). ***Pupil engagement*** will be measured at the end of the follow-up year (Term 3 during the 2018/19 school year). A previously validated measure, the "Engagement vs. Disaffection with Learning: Student-report" (EvDLS) which has previously been used with the age groups in this trial will be used for this purpose (Skinner et al., 2008). ***Change in practice measures*** will be collected through the teacher and TA surveys that will be collected at baseline (before the start of the intervention), again at the end of the first academic year of implementation (i.e. May/June 2018), and again at the end of the follow-up year (May/June 2019). Information about TA practice will also be gathered through classroom observations and audio recordings made in a small number of lessons. Staff surveys, interviews and classroom observations will be used for triangulation to map change in practice across time. Box 1 below provides an overview of the timing of all the above activities, per type of school.

Box 1: Data collection

Year→	**Planning Year: 16/17**			**Intervention year: 17/18**			**Follow-up year 18/19**		
Term → / Activity↓	T1	T2	T3	T1	T2	T3	T1	T2	T3
Staff surveys (Heads/SLT, teachers and TAs)			$X_{I/C}$			$X_{I/C}$			$X_{I/C}$
Staff interviews – 3-5 interviews per school (5 intervention schools; 5 control schools)				$X_{I/C}$		X_{I}			$X_{I/C}$
Classroom observations (across 5 intervention schools; and 5 control schools)			$X_{I/C}$			X_{I}			$X_{I/C}$
Audio recordings of TA-pupil interactions			$X_{I/C}$			X_{I}			$X_{I/C}$

With good fidelity of implementation (see fidelity section below), the expectation is to see **change** on the following aspects of practice:

- **Deployment of TAs**: teachers spending *more* time (and TAs *less* time) with lower-attaining and SEND pupils.
- **Increased quality of TA-pupil interactions**: better TAs knowledge and use of effective interactions (e.g. pupils being given more 'wait time'; greater use of prompts

and clues; reduced use of correction; and more utilisation of self-scaffolding strategies).

- **Increased quality of preparations**: quality of TAs' pre-lesson preparation and TAs' feedback to teachers, improvements in opportunities for and quality of teacher-TA liaison (self-reported by TAs and teachers).

Randomisation

The unit of randomisation in this trial was schools – and these were allocated using a stratified design.

Allocation to treatment or control was conducted in Stata, per the procedure described below, on the 5th of July 2017 by Dr Sutherland, after the baseline external reading test for year one had been either completed or (in a few cases) scheduled, and after schools had provided complete UPN list for the year 4 pupils and returned MoUs. (For schools that had not yet completed baseline testing they were not informed of allocation until after testing had been completed.)

Strata were constructed from the school-cluster membership (geography-based strata, with the four regions described above), and from a mean-split school-level prior attainment. We created this by adding the KS2 English and Maths scores together and dividing by two. We then calculated the mean of this measure based on sample data, then classified schools as 'low KS2' or 'high KS2' according to whether they were above or below this mean. Those schools with exactly the same mean would have been allocated to 'low KS2', but there were no schools with that value. This means that, for example, a school might be from the 'West Midlands' and have a 'high' KS2 score.

Five schools did not have prior attainment data available so were randomly allocated to hi/lo (via a coin flip witnessed by another member of the team) prior to randomisation taking place. Schools were randomly allocated to one of the two arms of the trial within each stratum. Table 2 below shows actual allocations by region and prior attainment.

Table 2: MITA randomisation results

	Allocation		
Region + KS2 prior	Control	MITA school	Total
London+Hi KS2	17	16	33
London+Lo KS2	17	17	34
Portsmouth+Hi KS2	6	5	11
Portsmouth+Lo KS2	6	5	11
Suffolk+Hi KS2	2	2	4
Suffolk+Lo KS2	2	2	4
West Mids+Hi KS2	6	6	12
West Mids+Lo KS2	10	9	19
Total	66	62	128

Calculation of sample size

The MDES calculation is based on the primary outcome for this trial, which is a pooled measure of pupil attainment (see descriptions below). The target number of schools was a minimum of 100, based on the capacity of the intervention team to deliver training at scale.

For the initial trial calculations we assumed that there were on average 33 pupils per class and that with 100 schools, two year groups per school and two form entry (e.g. around 130

pupils per school split in four classes), we will be collecting data on an estimated 13,000 pupils for this evaluation. Based on EEF guidelines (EEF, 2015),[3] the amount of variation explained by covariates (in this case the pre-test, and also the stratification variables) is assumed to be 0.53 for level 1 (pupils) and 0.00 for level 2 (schools). We also assume an alpha of 5% and an intended 80% power to detect effects. We use two-level clustered designs and base our calculations on an ICC of 0.13 as per EEF guidelines. Power and minimum detectable effect size (MDES) calculations were performed using the PowerUp! tool (Dong and Maynard, 2013).

With the achieved sample of 128 schools the MDES is below that of the original estimation at 0.182 (if the assumptions set out above hold).

With only 128 schools, the study would not be powered for sub-group analysis (see sub-group analyses section), such as different types of SEND or FSM pupils, so we have not presented specific MDES calculations for those groups here.

Table 3: MDES calculations for cluster-randomised trial using whole study sample based on planned and actual numbers of participating schools

***Unit of randomisation*: Schools**	**Initial calculation**	**Actual sample of randomised schools**
Number of schools	100	128
***Assumed* pupils per school**	130	130
***Assumed* ICC (between school variation)**	0.13	0.13
L2 Covariates	N/A	N/A
***Assumed* variance explained by L1 covariates**	0.53	0.53
Minimum Detectable Effect Size	**0.207**	**0.182**

3 sets out the MDES calculations for our analysis. Using the parameters above and with equal allocation to treatment and control the MDES was 0.207 for 100 schools. With the achieved sample of 128 schools the MDES is below that of the original estimation at 0.182 (if the assumptions set out above hold).

With only 128 schools, the study would not be powered for sub-group analysis (see sub-group analyses section), such as different types of SEND or FSM pupils, so we have not presented specific MDES calculations for those groups here.

Table 3: MDES calculations for cluster-randomised trial using whole study sample based on planned and actual numbers of participating schools

***Unit of randomisation*: Schools**	**Initial calculation**	**Actual sample of randomised schools**
Number of schools	100	128
***Assumed* pupils per school**	130	130
***Assumed* ICC (between school variation)**	0.13	0.13
L2 Covariates	N/A	N/A
***Assumed* variance explained by L1 covariates**	0.53	0.53
Minimum Detectable Effect Size	**0.207**	**0.182**

[3] EEF (2015) *Policy on Analysis for EEF evaluations.* London: EEF. (dated 15/12/2015)

Outcome measures

Primary outcome

The **primary outcome** will be reading, measured by (i) pupil attainment in Reading for the year 3 pupils; and (ii) pupil attainment in Reading for year 6 pupils. Outcome testing for the **year 3 pupils** will be done by a third party provider (ACER) using the **Essential Learning Metrics (ELMs) Reading Comprehension** test at the end of summer term 2019. For the purpose of analysis, we propose that these two outcome measures be standardised, with a mean of zero and standard deviation of one, then pooled so that all pupils can be included in the same model. Both outcomes will also be analysed independently as set out in the 'Additional analyses' section.

The ELMs was designed for pupils from Year 2 to Year 10 and reports achievement in reading comprehension on a continuous scale. The scale is represented in both scale scores and described 'bands of achievement', allowing progress to be monitored and mapped over time across the years of schooling. This test covers a range of text types (narrative, persuasive, informational) to assess a pupil's ability in retrieving information, interpreting explicit information, and interpreting by making inferences and reflecting on texts. ELMs is designed with reference to the National Curriculum. ELMs is online, standardised, can be administered to an individual or group (limited by number of computers), is not adaptive, and is untimed (but 40-50 minutes are recommended). For this trial the EEF required a paper version of the test to be used over concerns about how well completed online assessments might be, particularly in schools with older IT infrastructure.

All items in the ELMs tests have been reviewed by English teachers (in England) for linguistic, cultural or curriculum bias, and all items are multiple selected responses that minimise marking bias. External examination ACER staff, who are blind to allocation, will invigilate and mark the tests. ACER will undertake a data processing stage, recoding and checking for integrity with a report documenting recode syntax files, and will also produce a test calibration using ACER ConQuest as well as reviewing psychometric properties. Further, ACER will verify that the non-response rate is below 5%, if this is not the case, the analysis will follow the procedures set out in the 'Missing Data' section. Final data will then be supplied to RAND Europe for analysis. Schools will not be told in advance what the test is (i.e. the name of the test) but will be informed about the general areas it covers.

For the outcome measure for the **year 6 pupils**, Key Stage 2 results on reading and maths will be used, which will reflect performance of these pupils at the end of year 6. (In as much as we assume they such measures will be available at that time – we do not know whether the current KS2 regime will be in place in 2019.) The KS2 test provides the most reliable, efficient and equitable way of obtaining outcome data on attainment for this group. The tests have the advantage of being consistent across the school population and correlate well with KS1 results, which will be use as the baseline measure for Year 6.

The coefficient for the relationship between MITA allocation and the primary outcome measure (pooled for the two cohorts) will represent the main result of the trial.

Secondary outcomes

There are three **secondary outcomes.** i) pupil engagement (pupil level), and ii) change in teacher/TAs practice (teacher/TA level). It is expected that a better deployment of TAs will contribute to improved outcomes in these areas. The third secondary outcome is maths attainment by Year 6 pupils. This will be measured through KS2 maths scores.

Pupil engagement will be measured at the end of the follow-up year (Term 3 during the 2018/19 school year). A previously validated measure, the "Engagement vs. Disaffection with Learning: Student-report" (EvDLS) which has previously been used with the age groups in this trial will be used for this purpose (Skinner et al. 2008). This questionnaire will be undertaken on all treatment and control pupils, and will be invigilated by ACER staff who will be blind to allocation status.

Change in practice measures will be collected through the teacher and TA surveys that will be collected by the MITA team. Questions measuring change in the following aspects of practice will be collected at three times (baseline, end of intervention year and end of follow-up year): *deployment of TAs* (i.e., teachers spending more time, and TAs less time, with lower-attaining and SEND pupils), *increased quality of TA-pupil interactions* (i.e., perceived better TA knowledge and use of effective interactions, and pupils self-scaffolding more often; TAs provide prompts and clues appropriately, not correcting and completing pupils' work), and *increased quality of preparation* (i.e. perceived quality of TAs' pre-lesson preparation and TAs' feedback to teachers, improvements in opportunities for and quality of teacher-TA liaison) (see examples of questions in Appendix). Information about TA practice will also be gathered through classroom observations made in a small number of lessons and interviews. Observations will take place in ten schools at baseline (5 intervention schools and 5 control schools); then at the end of the intervention year in the same 5 intervention schools; and at the end of the follow-up year in the 10 initial control and intervention schools). The selection of lessons for these observations will be made on the lessons having a literacy or numeracy focus for pupils in the relevant cohorts.

Analysis

Primary intention-to-treat (ITT) analysis

The outcome analysis will be on an intention-to-treat (ITT) basis. This method compares outcome means for the treatment and comparison groups, and subjects are analysed according to their randomised group allocation. The ITT approach is inherently conservative as it captures the averaged effect of *offering* the intervention, regardless of whether or not the participants comply with the assignment.

Problems of dropout/non-attendance may be an issue for this trial depending on how motivated school staff are. Our main concern is that new teachers come in or that schools and/or teachers/TAs and/or consultants drop out at some point during the trial. Both risks are mitigated by this being a whole school intervention, although it will be important to assess the levels of staff turnover in both arms of the trial.

The **primary outcome** is pupil-level test scores (both external test and KS2, standardised with mean 0 and standard deviation 1, and then pooled). We will use a two-level multilevel model to account for clustering of data. Multilevel approaches assume that the schools in the study are a random sample of all schools and the multilevel modelling framework can flexibly handle complex variation within/between schools (Snijders and Bosker, 2012).

The main analysis consists of the model for outcomes of pupils nested in schools, which is:

$$Y_{ijpooled} = \beta_0 + \mathrm{MITA}_j\tau + Z_j\beta_1 + X_{ij}\beta_2 + u_j + e_{ij} \qquad (1)$$

where $Y_{ijpooled}$ is the achievement of student *i* in school j; MITA_j is a binary indicator of the school assignment to intervention [1] or control [0]; Z_j are school-level characteristics, here

the two stratifying variables of geographical location and prior KS2 results (as used for randomisation – so a binary measure); X_{ij} represents characteristics at pupil level (pupil $_i$ in school $_j$), specifically standardised baseline pupils scores (ELMS Reading and KS1_READWRITPOINTS, standardised and pooled as for the outcome measure)); u_j are referred to as school-level residuals ($u_j \sim i.i.d\ N(0, \sigma_u^2)$) and e_{ij} are individual-level residuals ($e_{ij} \sim i.i.d\ N(0, \sigma_e^2)$). Equation (1) is known as a 'random intercepts' model because $\beta_{0\mathrm{j}} = \beta_0 + u_j$ is interpreted as the school-specific intercept for school j and $\beta_{0j} \sim i.i.d\ N(\beta_0, \sigma_u^2)$ is random (as in it can take any value). The total residual variance can be partitioned into two components: the between-school variance σ_u^2 and the within-school variance σ_e^2. The between-school variance (ICC) will be calculated in the first instance, using a model with no predictors, but accounting for the clustering of pupils in schools (the so-called empty model).

Our target parameter (i.e. the focal result of the trial) τ is the average effect of the intervention on pupil outcomes compared to control schools. All analyses will be performed in Stata, versions 13.1 onwards.

Interim analyses

No statistical analyses are planned for interim outcomes, but the teacher survey data will be analysed in the late summer/autumn of 2018, to obtain basic descriptive statistics and between-group differences.

Imbalance at baseline for analysed groups

We have taken an active approach to address imbalance by stratifying the randomisation. A well-conducted randomisation will, in expectation, yield groups that are equivalent at baseline (Glennerster and Takavarasha, 2013)[4]. Because schools here are randomly allocated to the control and intervention conditions, any imbalance at baseline will have occurred by chance. To check for, and monitor, imbalance at baseline in the realised randomisation, analyses will be conducted at the school and pupil level. At the school level, the analysis will look at the following variables, by means of cross-tabulations and histograms that assess the distribution of each characteristic within control and treatment groups aggregated from the pupil data in the study sample (rather than publically available school-level statistics):

- Type of school (academy/non-academy).
- OFSTED rating.
- Proportion of pupils eligible for FSM.
- Proportion of pupils with SEND.
- Proportion of pupils speaking English as an additional language (EAL).
- Pupil:teacher ratio. (Including numbers of TAs if possible.)
- School-level average KS1 scores (for sample pupils).

At the pupil level, the initial balance will be assessed for the following characteristics:

- Eligibility for FSM (ever6 FSM).
- SEND status (0 / 1 variable for any SEND status).
- Gender.
- KS1 attainment (expressed as a standardised mean difference).

[4] Glennerster, R. and Takavarasha, K. (2013) *Running randomized evaluations: a practical guide.* London: Princeton University Press.

- ELMs test data (expressed as a standardised mean difference).

Statistical significance tests will not be carried out to assess the balance, as their premise does not hold in randomised control trials[5] (i.e. given appropriate randomisation procedures were followed, any differences between control and treatment groups at baseline will be by definition due to chance, and classical statistical testing is therefore unnecessary). Instead, tables of the means (and standard deviation, where appropriate) for each characteristic will be presented along with distributions. Where differences exist, and in relation to covariates that are deemed to be predictive of the outcome, the magnitude of any differences will be explored[6] and a decision made as to whether they require inclusion in the analysis (see Senn, 1994).[7]

Secondary outcome analyses

The **secondary outcomes** are pupil engagement (at pupil level), improved TA deployment and change in teacher/TA practice (at school level), and maths attainment for Year 6 pupils measured by KS2 Maths Test. Pupil engagement will be measured following a similar specification to equation listed under primary outcome analysis above, whilst the last two secondary outcomes will only be available at school level and captured through staff surveys, interviews and classroom observations. Equation 2 provides the specification for the school level models.

$$Y_j = \beta_0 + \mathrm{MITA}_j\tau + Z_j\beta_1 + X_j\beta_2 + u_j \qquad (2)$$

For the maths outcome for Year 6 pupils, we will use the same specification of the primary analysis using only the Year 6 maths outcome, limiting the sample accordingly:

$$Y_{ijKS2MathsYr6} = \beta_0 + \mathrm{MITA}_j\tau + Z_j\beta_1 + X_{ij}\beta_2 + u_j + e_{ij} \qquad (3)$$

Missing data

Missing data can arise from item non-response or attrition of participants at school, teacher and pupil levels. We will first determine the proportion of missing data in the trial. Our use of administrative data for pupil baseline data should reduce missingness arising from both item non-response and attrition for the older cohort. For the younger cohort we are relying on external testing. Below we set out our missing data strategy.

We will explore attrition across trial arms as a basic step to assess bias (Higgins et al., 2011). We will provide cross-tabulations of the proportions of missing values on all baseline characteristics (as detailed in the previous section, at both pupil and school level), as well as on the primary outcome measures.

---

5 http://www.consort-statement.org/checklists/view/32-consort/510-baseline-data

6 There is a convention in some disciplines that a 10pp (or larger) difference in treatment and control means at baseline constitutes 'imbalance' is thus justification for including those measures in sensitivity analyses, but there are counter-arguments to this idea (see Roberts,C. and Torgerson, D. (1999) 'Baseline imbalance in randomised controlled trials', *BMJ*, 319:185; but also see de Boer et al. (2015) 'Testing for baseline differences in randomized controlled trials: an unhealthy research behavior that is hard to eradicate', *International Journal of Behavioral Nutrition and Physical Activity*, 12:4). .

7 Senn, S. (1994) 'Testing for baseline balance in clinical trials', *Statistics in Medicine*, 13: 1715-1726.

To assess whether there are systematic differences between those who drop out and those who do not – and thus whether these factors should be included in analysis – we will model missingness at follow-up as a function of baseline covariates, including treatment. The analysis model for this approach will mirror the multilevel level model given above (pupils clustered in classes), but the outcome will be a binary variable identifying missingness (yes/no).

For less than 5% missingness overall, a complete-case analysis might suffice (i.e. assuming data are MCAR), but our default will be to check results using approaches that account for missingness but that rely on the weaker MAR assumption. Our preference is to use Full-Information Maximum Likelihood (FIML) over multiple-imputation because FIML can be estimated in a single model and simulation studies show that it can reduce bias as well as MI (for a discussion of FIML vs MI see Allison, 2012). (For missingness on outcome variables only standard statistical packages such as Stata use ML for estimating parameters so FIML would not be necessary (Allison, 2012).)[8]

Exploratory subgroup analyses

With only 128 schools, the study may not be powered for meaningful sub-group analysis, such as different types of SEND and/or FSM pupils.

We will report mean outcomes by sub-categories of SEND/FSM as a basic descriptive step. As an exploratory analysis we will do sub-group analyses for SEND and FSM, acknowledging that this analyses are likely to be underpowered. As an exploratory modelling approach, SEND will be incorporated into the regression analysis as a binary variable [1] if SEND, [0] otherwise (SENprovision_[term][yy]). The SEND indicator will then be interacted with treatment allocation to assess the conditional impact of MITA on SEND pupils. We will follow the same strategy for ever6 FSM pupils [yes/no] (using EverFSM_6_p as the FSM variable).

As there may be differential effects for the two cohorts (Year 3 and Year 6) then we will also conduct an exploratory analysis of the primary outcome for each year group using the same model as specified above in (3), sub-setting the data accordingly.

As these analyses are exploratory and very likely underpowered, we would report point estimates and confidence intervals transformed into effect sizes but would not report significance tests/p-values.

Effect size calculation

With the multilevel models we will use the effect sizes for cluster-randomised trials given in the EEF evaluator guidance; an example, adapted from Hedges (2007) is given below:

$$ES = \frac{(\bar{Y}_T - \bar{Y}_C)_{adjusted}}{\sqrt{\sigma_S^2 + \sigma_{error}^2}}$$

Where $(\bar{Y}_T - \bar{Y}_C)_{adjusted}$ is the mean difference between intervention groups adjusted for baseline characteristics and $\sqrt{\sigma_S^2 + \sigma_{error}^2}$ is an estimate of the population standard deviation (variance). In the multi-level models this variance will be the total variance (across both pupil

8 Allison, P. D. (2012) Why Maximum Likelihood is Better Than Multiple Imputation. Statistical Horizons. http://statisticalhorizons.com/ml-better-than-mi. And the more detailed discussion paper here: http://www.statisticalhorizons.com/wp-content/uploads/MissingDataByML.pdf.

and school levels, without any covariates, as emerging from a 'null' or 'empty' multi-level model with no predictors). The ES therefore represents the proportion of the population standard deviation attributable to the intervention (Hutchison and Styles, 2010). A 95% confidence interval for the ES, that takes into account the clustering of pupils in schools, will also be reported. Effect sizes will be calculated for each of the regressions estimated.

Treatment effect in the presence of Non-compliance

The main framework of analysis for this trial is Intention-to-Treat (ITT). However, we will also be able to explore the effect of the intervention on schools that were allocated to the intervention group and also implemented the intervention.

The analytical approach for non-compliance will consist of substituting the measure of compliance for the treatment allocation variable:

$$Y_{ijpooled} = \beta_0 + \text{MITA\_Compliance}_j\tau + Z_j\beta_1 + X_{ij}\beta_2 + u_j + e_{ij} \quad (4)$$

Our preference is to use a measure of compliance that is set at a threshold so that those schools above the threshold are regarded as 'high' compliers, and those below as 'low' or 'non-compliers', with a reference category of control schools (so a three-category variable). We anticipate that the long lead-in for the implementation will mean that all schools may achieve 'high' compliance eventually. If this is the case, then the compliance analysis will be the same as the main effects because the compliance measure will not vary. If that situation arises, there would be no need for a compliance-based analysis.

The compliance measure we have agreed with the developers involves scoring separate elements of implementation (see Table 4), weighting four elements as more important than others. The primary compliance measures (shaded blue) are given x3 the weight of other measures. If a school scores 70 or above on this measure, the delivery team would consider them to have "complied". This scoring would then inform the creation of the variable described above. We have agreed that missing data on any measure would be scored as zero. We have also agreed that schools would not have sight of the checklist.

As per the protocol, we have proposed a range of measures to capture implementation fidelity and compliance. We realise that we cannot use each measure on its own, so we propose combining the attendance measures collected as a proxy for 'engagement' (i.e. the proportion of all meetings and training scheduled that was attended).

Table 4: Fidelity measures for MITA

Senior Leadership Team buy-in and engagement (responsiveness)				
Measure[1]	**Data source**	**Tasks**	**Compliance score**	**Weighted score**
1. Attendance at all MITA sessions	Register of attendance	a. Attendance at SLT session 1	2	6
		b. Attendance at SLT session 2	2	6
		c. Attendance at SLT session 3	1	3
		d. Attendance at SLT session 4	1	3
			Maximum score	**18**
2. Attendance at school visit meetings	School visit checklist	a. Attendance at Reviewer Visit 1	2	6
		b. Attendance at Reviewer Visit 2	2	6
		c. Attendance at Reviewer Visit 3	2	6
			Maximum score	**18**
3. Attendance at MPTA staff training	MPTA training checklist	a. Attendance at TA training session 1	2	2
		b. Attendance at TA training session 2	2	2
		c. Attendance at teacher training session	1	1
			Maximum score	**5**
Maximum score achievable for SLT engagement				**41**

Adherence to the programme				
Measure	**Data source**	**Tasks**	**Compliance score**	**Weighted score**
4. Development team meetings	School visit checklist	Reviewers record that at least one meeting has taken place (indicative of MITA team having formed)	2	**2**
5. Percentage of Teachers and TAs completing MPTA training	MPTA trainer checklist	a. TA training session 1		
		• 50%-80% of TAs attend	1	3
		• 81%+ of TAs attend	2	6
		b. TA training session 2		
		• 50%-80% of TAs attend	1	3
		• 81%+ of TAs attend	2	6
		c. Teacher training session		
		• 50%-80% of teachers attend	1	3
		• 81%+ of teachers attend	2	6
			Maximum score[2]	**18**
6. Completion of school visits	School visit checklist	a. Reviewer Visit 1 delivered	2	6
		b. Reviewer Visit 2 delivered	2	6
		c. Reviewer Visit 3 delivered	1	3
			Maximum score	**15**
7. Completion of gap tasks	Audit component checklist Returns (e.g. action plan) School visit checklist	a. TA Audit (online) completed	2	2
		b. Completion of staff surveys	2	2
		c. Visioning exercise	2	2
		d. Action plan	2	2
		e. Reflective poster	1	1
		f. % of TAs who have **both:** • created mini-goals **and** • identified & worked on self-scaffolding targets with pupils		
		• 50%-80% of TAs	1	1
		• 81%+ of TAs	2	2
			Maximum score[2]	**11**
Maximum score achievable for adherence				**46**

Maximum score achievable for SLT engagement	**41**
Maximum score achievable for adherence	**46**
Total compliance score achievable	**87**
Score required to be compliant	**70**

Table notes:

[1] Primary compliance measures are shaded blue and given a weighting of x3 relative to the secondary measures (unshaded).

[2] Note that maximum score is not a simple addition for Measures 5 and 7 where points awarded depend on % of teachers/TAs attending training or completing tasks.

Report tables

We will report according to the EEF template.

APPENDIX

The following are examples of questions designed to measure change in practice over time in the teacher and TA surveys. These questions were adapted from existing surveys developed by MITA.

Examples of questions in teacher survey:

Q3: Thinking about what you did in your last three lessons, please order the following five activities by the amount of time spent on each from 1 to 5, where 1 is the activity you spent the MOST time doing in those lessons, and 5 is the activity you spent the LEAST amount of time doing.

- *Working one-to-one with a pupil*
- *Working with a pair or group*
- *Walking around the classroom (monitoring/ briefly supporting pupils)*
- *Delivering lessons*
- *Other (admin, marking)*

Q4: To what extent are the answers you have just provided typical of what you do in other lessons?

Q5: Once again, thinking about what you did in your last three lessons, which two groups of pupils did you spend the MOST time supporting?

- *Higher attaining pupils*
- *Average attaining pupils*
- *Lower attaining pupils (excluding SEND)*
- *Pupils with SEND*
- *Mixed attaining pupils*

Q6: To what extent are the answers you have just provided typical of what you do in other lessons?

Examples of questions in TA survey:

Q8: We would like to know about the opportunities you have to meet and communicate with the teachers you work with. Please select the statement below which best describes your experience.

- *The teacher(s) and I have scheduled time to meet each week*
- *I come into school early and/or stay behind after school. We use this as an opportunity to meet*
- *My communication with teacher(s) is brief and ad hoc (e.g. a couple of minutes before the lesson starts)*
- *There is no opportunity or time to communicate with teacher(s) outside of lessons*

Q9: Thinking about your daily work, for each of the areas listed below please indicate - on average - how prepared do you feel when you come into lessons? Please mark one choice in each row (Always, Often, Sometimes, Rarely, Never).

- *I know which pupil(s) I will support*
- *I am aware of the educational needs of the pupil(s) I will support*
- *I know what topic will be covered in the lessons*
- *I have enough subject knowledge to provide effective support*
- *I have enough pedagogical/ instructional knowledge to provide effective support*
- *I am aware of the expected outcomes for the pupil(s) I will support*
- *I know what feedback I need to give to the teacher at the end of the lesson*