Statistical Analysis Plan
SMART Spaces
Evaluator: UCL Institute of Education
Principal investigator: Jeremy Hodgen

PROJECT TITLE	SMART Spaces: Spaced Learning Revision Programme
DEVELOPER (INSTITUTION)	Queen's University Belfast / Hallam Teaching School Alliance
EVALUATOR (INSTITUTION)	UCL Institute of Education
PRINCIPAL INVESTIGATOR(S)	Jeremy Hodgen
TRIAL (CHIEF) STATISTICIAN	Nicola Bretscher
SAP AUTHOR(S)	Nicola Bretscher, Jeremy Hodgen, Jake Anders
TRIAL REGISTRATION NUMBER	ISRCTN54008927
EVALUATION PROTOCOL URL OR HYPERLINK	Link to evaluation protocol

SAP version history

VERSION	DATE	REASON FOR REVISION
1.0	12.4.2019	

Acknowledgement

We are grateful to Nikki Shure for providing quality assurance.

Table of contents

Introduction

This statistical analysis plan sets out the planned analysis for the evaluation of SMART Spaces: Spaced Learning Revision Programme (SMART Spaces Revision), an efficacy trial funded by the Education Endowment Foundation (EEF), to investigate the effect of the intervention on the chemistry element of the GCSE double award science.

The SMART Spaces revision programme uses spaced learning within chemistry revision for the AQA GCSE double award science examinations. Evidence from neuroscience and cognitive psychology (e.g. Fields, 2009) indicates that including spaces – time intervals - between learning sessions can improve factual recall. It is anticipated that improved factual recall will have a positive impact on the application and analysis as well as knowledge elements of the chemistry score in GCSE double award science. An earlier pilot study (O'Hare, Stark, McGuinness, Biggart & Thurston, 2017), also funded by the EEF, suggested that a combination of short (10 minute) and longer (approximately 24 hour or night-time sleep) spaces provides a promising model of spacing.

The intervention comprises both continuing professional development (CPD) and support for teachers to deliver the SMART Spaces revision programme and teacher implementation of the programme in Year 11 science lessons. The programme consists of six lessons delivered over two weeks and is designed to space the revision of content both between and within lessons. The chemistry topics for AQA Paper 1 are covered in one SMART Spaces lesson. This lesson is repeated three times in the same week, with spaces which allow pupils a night-time sleep between lessons. After at least one further night-time sleep, but ideally the following week, the process is repeated for content associated with AQA Paper 2. Within lessons, chemistry topics are revised using the SMART spaces materials in three short ~12-minute sessions with 10-minute spaces between each topic. During the 10-minute spaces, pupils take part in a sensorimotor activity (such as juggling).

The evaluation is structured as a two-armed school-level cluster randomised controlled trial involving 125 secondary schools. Fifty-four schools were allocated to receive the intervention and 71 to a business as usual control group. Recruitment occurred in Spring-Autumn 2018 with the aim of initiating training for teachers in intervention schools in November 2018. The evaluation will look at the impact of the programme on pupils' performance on the chemistry element of the AQA GCSE double award science.

Design overview

Trial type and number of arms		Cluster randomised, two arms
Unit of randomisation		School
Stratification variables (if applicable)		Randomisation block, School-level prior attainment
Primary outcome	variable	Chemistry attainment
	measure (instrument, scale)	Chemistry sub-scale of AQA GCSE Double Award Science (item-level, continuous)
Secondary outcome(s)	variable(s)	[1] Science attainment [2] Knowledge, application and analysis elements of chemistry attainment
	measure(s)	[1] AQA GCSE Double Award Science (item-level, continuous)

	(instrument, scale)	[2] Knowledge, application and analysis assessment objectives (AO) sub-scales for the Chemistry element of AQA GCSE Double Award Science (item-level, continuous)

This is a cluster randomised controlled trial, with randomisation taking place at the school level. Chemistry teachers often teach across several science classes, so randomising at class-level was not possible. Although pupil-level randomisation would increase power, the disruption that this would cause was judged likely to be unacceptable to many schools, especially in the crucial GCSE year. In addition, in the case of both pupil and class-level randomisation, the potential for within-school contamination of the revision approach would be high.

The developer has limited capacity to deliver the training and coaching to schools. In order to ensure the trial has sufficient power given this limited capacity, allocation to the arms was unequal. The trial successfully recruited 125 secondary schools, with 54 schools randomly allocated to the intervention and 71 to the business as usual control. Schools in the control group will receive £1000 following the completion of all evaluation requirements with staff/school and with the required pupils in 2018 and 2019. After the evaluation has finished, the school may purchase the SMART Spaces programme from QUB/HTSA for use from January 2020.

The eligibility criteria for schools to participate were:

- participating schools must be English state-funded secondary schools and have some of their pupils enrolled in AQA GCSE double award science
- schools had to agree not to participate in another EEF GCSE science randomised trial that would interfere with implementation of the intervention with Year 11 pupils during 2018/19 academic year.
- schools had to return a signed Memorandum of Understanding (MoU), in which they committed to participating fully in the study, including the collection of outcome measures in summer 2019, regardless of which trial arm they are assigned to.

Randomisation took place in two batches following recruitment of schools, including collection of signed MoUs and baseline data, to aid recruitment and delivery. The first randomisation batch of 82 schools took place on 16th October 2018. The second batch of 43 schools took place on 5th December 2018. Each randomisation batch was blocked by school-level prior attainment to ensure sufficient balance between treatment and control groups on this characteristic, with the aim of maximising internal validity. For the purposes of randomisation, prior attainment was measured as pupils' combined KS2 score, calculated by taking the mean KS2 level for mathematics and English.

In the evaluation protocol, we stated an initial preference for a simple randomisation using a four blocked design stratified by:

1. randomisation batch (2 groups), and
2. school-level average prior attainment of pupils included in the analytic sample (2 groups).

The attainment blocks were to be defined by a median split based on sample characteristic i.e. the school-level mean combined KS2 score of pupils included in the analytic sample.

However, randomisation in batches can introduce imbalance. For this reason and prior to the first batch, we simulated the randomisation process based on the actual sample of 82 schools. Based on EEF (2016) guidance on using balance to adjust security ratings for trial evaluations, we set a criterion that imbalance across the final treatment and control groups should not exceed a standardised difference of 0.05 for pupil-level KS2 attainment. Simulation revealed that the initial preferred approach would lead to unacceptable levels of imbalance across the trial arms in this key characteristic of pupil-level attainment. We examined the distribution of imbalance across simulations, as well as relationships between school-level prior attainment, school-level proportion of FSM and school-sample size to gain insight into the causes of this imbalance. As a result, we found the imbalance in pupil-level KS2 attainment appeared to be due to a high degree of variation in the school-sample size, ranging from 6 to 250 pupils, coupled with a relationship between that and school-level prior attainment (see Appendix 6). We also used simulation to check for imbalance in school-level prior attainment and school-level proportion of FSM. We report the imbalance in these measures for the actual randomisation in the batch 1 and 2 randomisation logs, see Appendices 4 and 5. However, we prioritised achieving an acceptable level of balance on pupil-level attainment only, given the difficulties exposed by the simulation results in meeting this key criterion.

Following the simulation results, we modified our randomisation approach to ensure we met the criterion regarding acceptable balance on pupil-level KS2 attainment. We set a trigger for re-randomisation (Morgan & Rubin, 2012) at a standardised difference of 0.05 for pupil-level KS2 attainment for the first randomisation batch (i.e. a lower bound of -0.05 and an upper bound of +0.05). We then calculated appropriate boundaries for triggering re-randomisation in the second batch to achieve an acceptable level of balance between the overall treatment and control groups, taking into account the results of the first batch randomisation. This amended approach and the steps of the process set out below were shared with and agreed by EEF prior to randomisation.

We randomised the first batch of 82 schools according to the following process:

1. The schools were stratified into 16 blocks on the basis of school-level KS2 prior attainment (split across 16 quantiles on school average attainment). This produced clusters of five or six schools, facilitating allocation to treatment and control groups in the ratio 2:3 in a form of pair matching (Imai, King & Nall, 2009).

2. Each school was assigned a randomly generated number (setting a stable seed for the random number generation).

3. The schools were sorted by block and random number.

4. Schools were assigned to the treatment arm and control arm in the ratio 2:3.

5. Re-randomisation (i.e. repetition of steps 1-5) was triggered if the standardised difference between treatment and control groups for pupil-level KS2 attainment exceeded 0.05.

Of the 82 schools in the first batch: 33 were allocated to the intervention, 49 to the control. There was a standardised difference of 0.027 for pupil-level KS2 attainment between treatment and control groups in the first batch randomisation (see Appendix 4).

Simulation was also carried out prior to the second batch randomisation of 43 schools. This supported a blocking strategy with schools grouped again on the basis of school-level KS2 prior attainment in a form of pair matching (Imai, King & Nall, 2009). In the second batch, schools were allocated to treatment and control arms in a close to 1:1 ratio, allowing the total number of schools allocated to the SMART Spaces intervention to rise to 54 (from 50). The change in allocation ratio took advantage of maximum developer capacity in order to increase power by equalising the number of schools allocated to the two trial arms. Taking into account the imbalance between intervention and control groups from the first batch randomisation, we calculated a trigger for re-randomisation be set with a lower bound of -.210 and an upper bound of 0.098 for pupil-level KS2 attainment to meet the criterion regarding acceptable balance on this measure. The uneven boundaries are a result of taking into account the imbalance between intervention and control groups from the first batch randomisation. Simulation results suggested keeping within these boundaries would be achievable. Hence, we randomised the second batch of 43 schools according to the following process:

1. The schools were stratified into 8 blocks on the basis of school-level KS2 prior attainment (split across 8 quantiles on school average attainment). This produced clusters of four to seven schools.

2. Each school was assigned a randomly generated number (setting a stable seed for the random number generation).

3. The schools were sorted by block and random number.

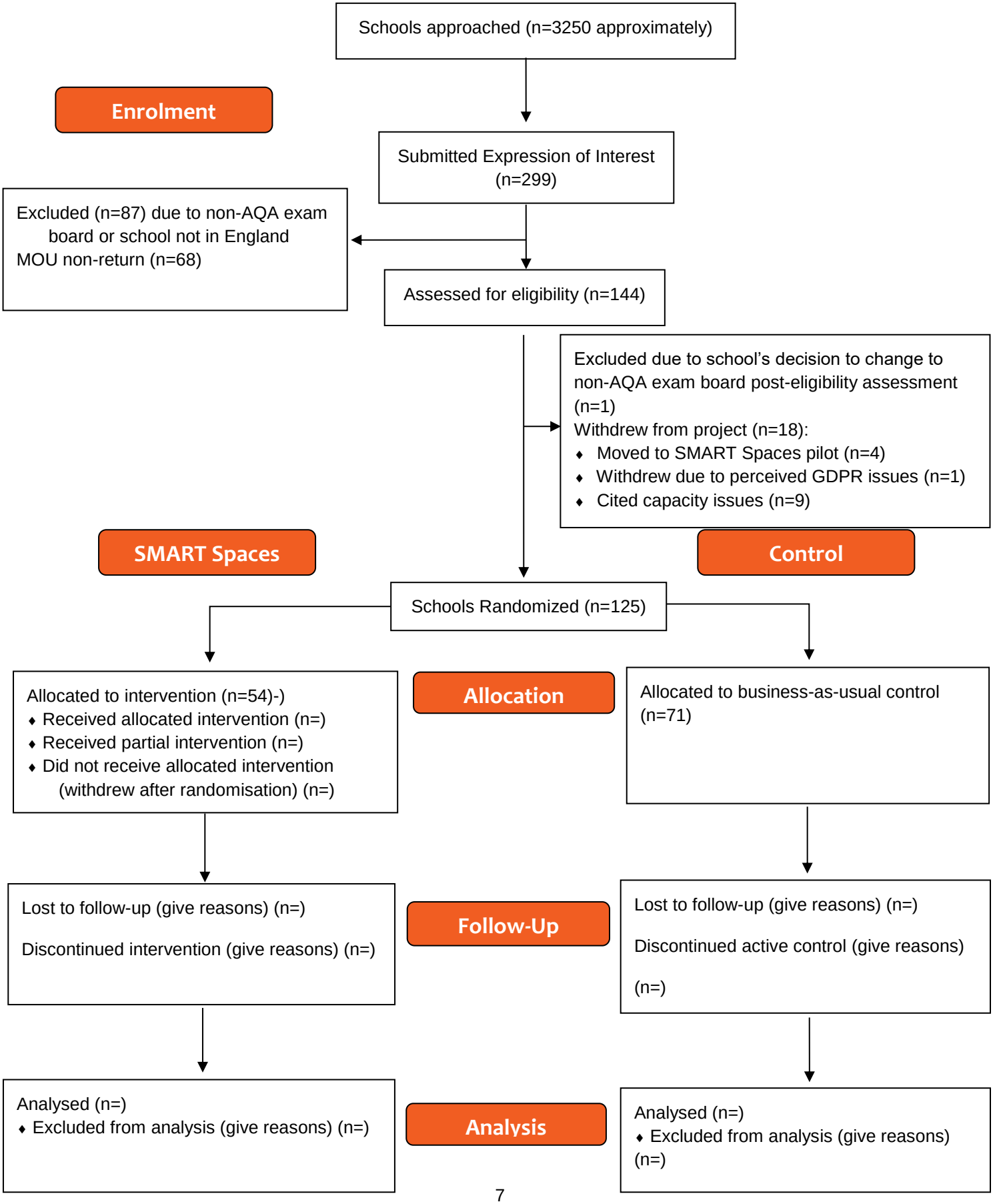
4. Schools were assigned to the treatment arm and control arm in turn.

5. Re-randomisation (i.e. repetition of steps 1-5) was triggered if the standardised difference between treatment and control groups was outside a lower bound of -.210 and an upper bound 0.098 for pupil-level KS2 attainment.

Of the 43 schools in the second batch: 21 were allocated to the intervention, 22 to the control. Overall, 125 schools were randomised, resulting in a total of 54 schools allocated to the treatment group and 71 to the control. It is worth noting at this point, that including pupil-level KS2 attainment as a covariate in the analysis model is the appropriate adjustment for both our re-randomisation and our blocking strategies.

In the protocol, we stated that we would specify an acceptable level of attrition on the basis of simulation of the dataset and amend the protocol when the Statistical Analysis Plan is agreed following randomisation. Due to prioritising recruitment for the SMART Spaces Teaching Pilot, we did not have time to do this by the SAP deadline. We still intend to carry out simulations of the dataset, amending the SAP and protocol as appropriate, and will do so by June 2019.

Follow-up


Schools approached (n=3250 approximately)
Enrolment
Submitted Expression of Interest (n=299)
Excluded (n=87) due to non-AQA exam board or school not in England
MOU non-return (n=68)
Assessed for eligibility (n=144)
Excluded due to school's decision to change to non-AQA exam board post-eligibility assessment (n=1)
Withdrew from project (n=18):
♦ Moved to SMART Spaces pilot (n=4)
♦ Withdrew due to perceived GDPR issues (n=1)
♦ Cited capacity issues (n=9)
SMART Spaces
Control
Schools Randomized (n=125)
Allocated to intervention (n=54)-)
♦ Received allocated intervention (n=)
♦ Received partial intervention (n=)
♦ Did not receive allocated intervention (withdrew after randomisation) (n=)
Allocation
Allocated to business-as-usual control (n=71)
Lost to follow-up (give reasons) (n=)
Discontinued intervention (give reasons) (n=)
Follow-Up
Lost to follow-up (give reasons) (n=)
Discontinued active control (give reasons) (n=)
Analysed (n=)
♦ Excluded from analysis (give reasons) (n=)
Analysis
Analysed (n=)
♦ Excluded from analysis (give reasons) (n=)

Sample size calculations overview

		Protocol		Randomisation	
		OVERALL	FSM	OVERALL	FSM
MDES		0.198	0.210	0.196[1]	0.209
Pre-test/ post-test correlations	level 1 (pupil)	0.50	0.50	0.50	0.50
	level 2 (class)	N/A	N/A	N/A	N/A
	level 3 (school)	0.25	0.25	0.25	0.25
Intracluster correlations (ICCs)	level 2 (class)	N/A	N/A	N/A	N/A
	level 3 (school)	0.15	0.15	0.15	0.15
Alpha		0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8
One-sided or two-sided?		Two-sided	Two-sided	Two-sided	Two-sided
Average cluster size		100	25	113	25
Number of schools	intervention	50	50	54	54
	control	75	75	71	71
	total	125	125	125	125
Number of pupils	intervention	5000	1250	6465	1796
	control	7500	1875	7633	1837
	total	12500	3125	14098	3633

Protocol MDES calculations were carried out using the R package PowerUpR and based on the following assumptions:

- **Randomisation would be at school level stratified using a four blocked design.** A number of options for unequal allocation to the intervention and control groups were considered before settling on the ratio 50:75 intervention to control.
- **Number of children per cluster is 100.** This is equivalent to around four Double Science GCSE classes per school. Since around 75% of pupils take the double award[2] and the average size of a secondary school is around 180 pupils per year group, this is a relatively conservative assumption.
- **Pupil-level pre- to post-test correlation = 0.5.** No data are available for the correlation between the chemistry element of the GCSE double award science and

[1] The overall MDES post-randomisation was calculated using the arithmetic mean of pupils in schools for average cluster size to maintain consistency with the protocol calculations. Due to the level of variation in cluster size in the sample, we also calculated the MDES post-randomisation using the harmonic mean, a more conservative statistic, for average cluster size which produced an overall MDES = 0.198.

[2] https://ffteducationdatalab.org.uk/2017/03/weird-science/

combined Key Stage 2 (KS2) scores. The correlation was estimated on the basis of the correlation between GCSE science and combined KS2 scores of 0.556 (Benton & Sutch, 2014). This was judged a better predictor than the correlation between triple award chemistry and KS2 scores. However, since this correlation is lower (0.427, ibid.), a conservative estimate was judged appropriate.

- **School-level pre- to post-test correlation = 0.25.** This was estimated to be half the pupil-level correlation on the basis of advice from the EEF evaluation advisory panel.
- **An intra-cluster correlation coefficient (ICC) of 0.15**, based on EEF's (2015) guidance.
- **Power: 80%; Significance level: 5%.** These are standard assumptions.

Randomisation MDES calculations were based on the same assumptions, with the following alterations:

- **Randomisation took place in two batches, with schools stratified into 24 blocks** on the basis of school-level KS2 prior attainment in a form of pair matching. Schools were allocated in a ratio 54:71 intervention to control.
- **Number of children per cluster is 113.** This was calculated as the mean number of GCSE Double Award science students for whom data was submitted per school.
- **Number of students eligible for FSM per cluster is 25.** This was calculated as the mean number of students, out of those included in the school-sample, identified as eligibile for FSM per school.
- **In other respects, the assumptions remain the same as those expressed in the evaluation protocol.**

The overall MDES post-randomisation is in line with the effect size g= 0.19 produced from the Optimisation Study (O'Hare et al.,2017).

Analysis

Primary outcome analysis

Our primary analysis will focus on the chemistry sub-scale of AQA GCSE Double Award Science, a continuous numerical variable based on item-by-item mark data. In line with EEF guidance (2018), the outcome variable will be used in its 'raw' form as there is no clear reason to transform the data.

Outcome variables will be modelled on the basis of intention to treat (ITT) using a linear multi-level model. We will fit a 2-level multi-level model of students clustered in schools incorporating the treatment condition, and the pre-test and other stratification variables used for randomisation as covariates. The pre-test measure is a pupil-level continuous numerical variable, comprising a simple aggregation of English (range: 0-50) and mathematics (range: 0-100) KS2 raw scores, noting the importance of literacy and reading comprehension in particular for science attainment at KS2 and 3 (Nunes et al., 2017), and the increased emphasis on mathematics within the science GCSE (OFQUAL, 2015). Stratification variables are the block assignation based on KS2 school-level average attainment and whether the school was randomised as part of the first or second batch. Note that including pupil-level KS2 attainment as a covariate in the analysis model is the appropriate adjustment

for both our re-randomisation and our blocking strategies (Rubin, 2008; Morgan & Rubin, 2012).

We will estimate the following model:

$$y_{ij} = \beta_0 + \beta_1 Treat_j + \beta_2 PreTest_{ij} + \beta_3 X_j + u_j + \varepsilon_{ij}$$

$$u_j \sim N(0, \sigma_u^2)$$

$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$$

where individual i is nested in school j, y_{ij} is the score on the Chemistry sub-scale of AQA GCSE Double Award Science, $PreTest_{ij}$ is the simple aggregation of English and mathematics KS2 scores, $Treat_j$ is our school-level treatment indicator, X_j is a categorical variable indicating block assignation, i.e. a vector of stratification variables, as previously defined in this section. β_0 represents the grand mean for the outcome variable, u_j and ε_{ij} represent a school-level random effect and an error term at the pupil level respectively.

All models will be estimated using Bayesian inference (Gelman et al., 2014) with the software STAN through R using weakly informative and diffuse priors. The primary outcome will be reported using 95% Bayesian credible intervals. We will also report classical confidence intervals to enable comparability with other EEF trials. In a Bayesian framework, there is no direct equivalent to null hypothesis testing. Following Kruschke and Liddell (2018), we will use a ROPE (Region of Practical Equivalence) analysis set at an effect size of ±0.1 around 0 to examine whether the null hypothesis should be accepted as credible or practically distinguishable. This procedure examines the proportion of the Highest Density Interval (HDI) that falls within the ROPE pre-determined effect size. This approach has been selected because it provides comparability with a standard frequentist approach and effect sizes (Kruschke, 2018) used in reporting other EEF trials. We will also report the results of a standard frequentist approach, for consistency with other EEF trials, as set out in the section entitled *Additional analyses* below.

Secondary outcome analysis

We will conduct two secondary outcome analyses:

1. AQA GCSE Double Award Science raw score, a continuous numerical variable based on item-level mark data. Same model as the primary outcome analysis except replace y_{ij} with AQA GCSE Double Award Science raw score.

2. The knowledge, application and analysis assessment objectives (AO) sub-scales for the Chemistry element of AQA GCSE Double Award Science, a continuous numerical variable based on item-by-item mark data. We will model the secondary outcomes as three separate models (rather than through a multivariate multilevel model) given that they consist of questions from the GCSE chemistry sub-scale. Same as the primary outcome analysis except replace y_{ij} variously with each of the three knowledge, application and analysis assessment objectives (AO) sub-scales in turn.

Interim analyses

No interim analyses are planned.

Subgroup analyses

A sub-group analysis will be carried out for everFSM pupils, adding an interaction effect between treatment and everFSM to the primary outcome model. For this everFSM sub-group analysis, we will estimate the following model:

$$y_{ij} = \beta_0 + \beta_1 Treat_j + \beta_2 PreTest_{ij} + \beta_3 X_j + \beta_4 Treat_j * FSMever_{ij} + \beta_5 FSMever_{ij} + u_j + \varepsilon_{ij}$$

$$u_j \sim N(0, \sigma_u^2)$$

$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$$

where the variables are the same as in the primary outcome analysis with the addition of $\boldsymbol{FSMever_{ij}}$ an indicator of FSM eligibility available from the National Pupil Database (NPD) and $\boldsymbol{Treat_j * FSMever_{ij}}$ is an interaction between the school-level treatment indicator and everFSM. This analysis will require the primary outcome data to be matched with the NPD, to provide 'everFSM' data for pupils. NPD data will be matched to original pupil data collected before randomisation.

Secondly and similarly, a sub-group analysis will be carried out for sex, adding an interaction effect between treatment and sex to the primary outcome model. This sub-group analysis is deemed necessary because the under-participation of girls in science is judged to be an important issue for both policy and research (Royal Society, 2014; TISME, 2013). For this sex sub-group analysis, we will estimate the following model:

$$y_{ij} = \beta_0 + \beta_1 Treat_j + \beta_2 PreTest_{ij} + \beta_3 X_j + \beta_4 Treat_j * Sex_{ij} + \beta_5 Sex_{ij} + u_j + \varepsilon_{ij}$$

$$u_j \sim N(0, \sigma_u^2)$$

$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$$

where the variables are the same as in the primary outcome analysis with the addition of $\boldsymbol{Sex_{ij}}$ is a binary indicator of the sex of the pupil and $\boldsymbol{Treat_j * Sex_{ij}}$ is an interaction between the school-level treatment indicator and sex.

Finally, a sub-group analysis will be carried out to investigate differential treatment effects depending on prior attainment at pupil-level by adding an interaction effect between treatment and $\boldsymbol{PreTest_{ij}}$ to the primary outcome model. As noted in the evaluation protocol, one of the findings of the optimisation study was that teachers considered the intervention to have greater benefits for low attaining students and that higher attaining students were less engaged and perceived there to be less benefit (O'Hare, 2017, p.32). For this pupil-level prior-attainment sub-group analysis, we will estimate the following model:

$$y_{ij} = \beta_0 + \beta_1 Treat_j + \beta_2 PreTest_{ij} + \beta_3 X_j + \beta_4 Treat_j * PreTest_{ij} + u_j + \varepsilon_{ij}$$

$$u_j \sim N(0, \sigma_u^2)$$

$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$$

where the variables are the same as in the primary outcome analysis with the addition of $\boldsymbol{Treat_j * PreTest_{ij}}$ an interaction between the school-level treatment indicator and pupil-level prior attainment.

For each sub-group analysis, if a significant interaction is found, we will run a separate model using only the relevant sub-group, using the same model as our primary analysis. For prior-attainment, sub-groups will be defined by tertiles based on prior attainment of the overall sample. The subgroup analysis will be conducted for both the primary and secondary outcomes.

Additional analyses

We will run a sensitivity analysis with cluster size $\boldsymbol{ClusterSize_j}$ included as an additional covariate in the primary and secondary outcome models to take account of variation in cluster size. That is, we will estimate the following model:

$$y_{ij} = \beta_0 + \beta_1 Treat_j + \beta_2 PreTest_{ij} + \beta_3 X_j + \beta_4 ClusterSize_j + u_j + \varepsilon_{ij}$$

$$u_j \sim N(0, \sigma_u^2)$$

$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$$

where $\boldsymbol{y_{ij}}$ is variously the primary and secondary outcome measures.

O'Hare et al (2017) found that pupil engagement in the intervention was a significant implementation factor, with higher engagement score predicting more positive outcome change. As part of the IPE, we will explore the effect of engagement on the intervention, using a pupil-level engagement measure based on survey items from the optimisation study (OS) (O'Hare et al., 2017). We have used factor analysis as an exploratory tool for validating this measure using data from the OS. We have also employed Rasch analysis as a means of confirming the factor analysis results and to construct an interval measure of engagement in spaced learning from these items (for results of validation process see Appendix 7). We will conduct an interaction analysis, by estimating the following model:

$$y_{ij} = \beta_0 + \beta_1 Treat_j + \beta_2 PreTest_{ij} + \beta_3 X_j + \beta_4 Treat_j * Engage_{ij} + \beta_5 Engage_{ij} + u_j + \varepsilon_{ij}$$

$$u_j \sim N(0, \sigma_u^2)$$

$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$$

where the variables are the same as in the primary outcome analysis with the addition of $\boldsymbol{Engage_{ij}}$ is a measure of pupil-level engagement and $\boldsymbol{Treat_j * Engage_{ij}}$ is an interaction between the school-level treatment indicator and pupil-level engagement measure. In the first instance, we will use the raw score on engagement items at the measure of pupil-level engagement, before subsequently performing this analysis with the Rasch-constructed scale.

Additional sensitivity analyses will be conducted i.e. replicating the primary and secondary outcome analyses with different software. As noted above, while the intention is to fit these models using Bayesian inference, we will fit each model classically using lme4 and MLwiN (for consistency with other EEF trials) before we refit the model using linear multilevel/hierarchical regression modelling estimated by Bayesian inference. We will conduct analyses to assess the sensitivity of the model to missingness as laid out in the sub-section on missing data.

Imbalance at baseline

We will check for balance of analysed sample for the following characteristics:

- pre-test simple aggregation of English and mathematics KS2 scores
- proportion ever eligible for Free School Meals.

As per Anders and Shure (2018), we will do this by reporting means and standard deviations for the treatment and control group and calculating absolute standardised differences (Imbens & Rubin, 2015) between the treatment and control groups. We will also present histograms of pre- and post-test data distributions.

Missing data

In this section we set out our strategy for missing data, following the approach described by Anders and Shure (2018). We will describe and summarise the extent of missing data in the primary and secondary outcomes, and in the model associated with the analysis. Reasons for missing data will also be described.

For all models we will implement a missing data strategy if more than 5% of data in the model is missing or if more than 10% of data for a single school is missing. The strategy will be followed separately for each instance of model and variable for which the threshold is exceeded:

- We will first assess whether the missing data is missing at random (MAR), since this is a pre-requisite for missing data modelling to produce meaningful results. To do this we will create an indicator variable for each variable in the impact model specifying whether the data is missing or not. We will then use logistic regression to test whether this missing status can be predicted from the following variables: all variables in the primary outcome analysis model plus eligibility for FSM (and proportion eligible for FSM in the school), sex of the pupil and GCSE science raw score. Where predictability is confirmed we will proceed to the appropriate next step of this strategy.

- For situations for which the MAR assumption appears to hold and only the outcome variable in the model is missing, we will re-estimate the treatment effect using our pre-specified model with the addition of the covariates found to be statistically significantly predictive of missingness of the outcome.

- For situations for which the MAR assumption appears to hold and any variable other than the outcome variable in the model is missing, we will use all variables in the analysis model plus eligibility for FSM (and proportion eligible for FSM in the school), sex of the pupil and GCSE science raw score to estimate a Multiple Imputation (MI) model using a fully conditional specification, implemented using Multiple Imputation by chained equations (mice), an imputation package within R (van Buuren & Groothuis-Oudshoorn, 2011), to create 20 imputed data sets. We will re-estimate the treatment effect using each dataset and take the average and estimate standard error using Rubin's (2004) combination rules.

Analysis using the multiply-imputed dataset will be used as a sensitivity analysis i.e. 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 the multiply-imputed dataset. If the complete case analysis model implies effectiveness but the imputed estimate does not 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.

Compliance

Compliance will be analysed at school-level using an Instrumental Variables (IV) approach with group allocation as the instrumental variable for the compliance indicator. We adapted our approach to compliance as a deeper understanding of the difficulties in delivering an ideal implementation of the SMART Spaces intervention was developed. Capturing these difficulties was rather more complicated than originally envisaged e.g. due to higher variation in school's timetabling of science lessons than expected, hence we adopted a more nuanced approach to compliance. In this section, we set out first our agreed approach before noting our original compliance plan. We will report the results of both sets of analyses specified below.

We agreed a set of compliance indicators with the developer, based on attendance at CPD and coaching sessions and the delivery of SMART Spaces lessons, specified as follows:

- Percentage of teachers attending CPD. This will be calculated using the number of teachers who planned to deliver SMART Spaces, dividing actual CPD and coaching sessions attended by total possible CPD and coaching sessions attended.
- Percentage of SMART Spaces lessons taught. We will calculate this using the actual number of lessons delivered divided by the total lessons expected to be delivered i.e. six SMART Spaces lessons multiplied by the number of double award classes.
- Percentage of delivery with appropriate spacing. This will be calculated as the number of classes receiving the SMART Spaces intervention with appropriate spacing divided by the total number of double award classes.

Attendance registers from training and coaching sessions will be collected from the developer in order to assess attendance. The teacher survey will be used to collect information about number of SMART Spaces lessons taught and the implementation of spacing. We will estimate a (first stage) model of compliance, as follows:

$$Comply_j = \beta_0 + \beta_1 Treat_j + \beta_2 PreTest_{ij} + \beta_3 X_j + u_j + \varepsilon_{ij}$$

$$u_j \sim N(0, \sigma_u^2)$$

$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$$

where $Comply_j$ is a continuous compliance variable based on the indicators defined above. The predicted values of $Comply_j$ from the first stage are used in the estimation of the second stage 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:

$$y_{ij} = \beta_0 + \beta_1 \widehat{Comply}_j + \beta_2 PreTest_{ij} + \beta_3 X_j + u_j + \varepsilon_{ij}$$

$$u_j \sim N(0, \sigma_u^2)$$

$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$$

where $\widehat{Comply}_j$ are the predicted values of treatment receipt derived from the first stage model. Our primary outcome of interest will be β_1, which should recover the effect of the intervention among compliers. Results for the first stage will be reported alongside with i) the correlation between the instrument and the endogenous variable; and ii) an F test as per EEF (2018) statistical analysis guidance.

In our original compliance plan, we agreed a definition of minimum compliance with the developer based on attendance at training and coaching sessions and the delivery of SMART Spaces lessons, specified as:

- All (100%) of teachers delivering SMART must receive training (assessed through QUB/HTSA records of attendance)
- All double award classes to receive both sets of 3 SMART Spaces lessons (6 sessions in total) delivered with appropriate 'spacing' as set out in the protocol (assessed through the teacher and student survey)

As the developer and evaluation teams appreciated the complexity of delivering an ideal implementation, we had concerns that this plan may have led to underestimating the local average treatment effect for minimum compliance. For this reason, just prior to publication of the SAP version 1.0, we adjusted our plan as set out above. Neverthless, for the purposes of transparency, we will also carry out our original compliance plan, reporting the results of the analysis specified in the following paragraphs.

As before, attendance will be assessed using the developer's attendance registers. Teachers and students will be asked about the number of SMART Spaces lesson delivered through the surveys above (allowing us to triangulate these data). Where there is a consensus in student response i.e. a majority agreement, we will assess compliance using the student survey. Otherwise, we will use the teacher survey response.

We will estimate a (first stage) model of compliance, as follows:

$$IdealComply_j = \beta_0 + \beta_1 Treat_j + \beta_2 PreTest_{ij} + \beta_3 X_j + u_j + \varepsilon_{ij}$$

$$u_j \sim N(0, \sigma_u^2)$$

$$\varepsilon_{ij} \sim N(0, \sigma_{\varepsilon}^2)$$

where $IdealComply_j$ is the binary compliance variable defined above. The predicted values of $IdealComply_j$ from the first stage are used in the estimation of the second stage 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:

$$y_{ij} = \beta_0 + \beta_1 \widehat{IdealComply_j} + \beta_2 PreTest_{ij} + \beta_3 X_j + u_j + \varepsilon_{ij}$$

$$u_j \sim N(0, \sigma_u^2)$$

$$\varepsilon_{ij} \sim N(0, \sigma_{\varepsilon}^2)$$

where $\widehat{IdealComply_j}$ are the predicted values of treatment receipt derived from the first stage model. Our primary outcome of interest will be β_1, which should recover the effect of the intervention among compliers. Again, results for the first stage will be reported alongside with i) the correlation between the instrument and the endogenous variable; and ii) an F test as per EEF (2018) statistical analysis guidance.

We will also investigate the effects of “non-compliance” in the control group. The SMART Spaces intervention is not publicly available, so schools in the control group will not have access to the intervention materials. However, there may be some schools, or teachers, in the control group who use a spaced learning approach for revision, and we will attempt to capture these “always compliers" using survey data, and if sufficiently robust data are available, we will investigate control group non-compliance quantitatively.

Intra-cluster correlations (ICCs)

We will employ a random intercept-only multi-level model to estimate the intra-cluster correlation (ICC) of the pre-and post-tests at school-level, as follows:

$$y_{ij} = \beta_0 + u_j + \varepsilon_{ij}$$

$$u_j \sim N(0, \sigma_u^2)$$

$$\varepsilon_{ij} \sim N(0, \sigma_{\varepsilon}^2)$$

where individual $\boldsymbol{i}$ is nested in school $\boldsymbol{j}$, $\boldsymbol{y_{ij}}$ is the score on the Chemistry sub-scale of AQA GCSE Double Award Science, $\boldsymbol{\beta_0}$ represents the grand mean for the outcome variable, $\boldsymbol{u_j}$ and $\boldsymbol{\varepsilon_{ij}}$ represent a school-level random effect and an error term at the pupil level respectively.

The ICC itself will be estimated from this model using the following equation:

$$\boldsymbol{\rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_{\varepsilon}^2}}$$

Effect size calculation

Effect sizes will be calculated using the Cohen's *d* ES for cluster randomised trials as per the current EEF (2018) statistical analysis guidance for evaluations. The formula is specified below:

$$ES = \frac{\bar{Y}_t - \bar{Y}_c}{\sqrt{(\sigma_u^2 + \sigma_\varepsilon^2)}}$$

where $\bar{Y}_t - \bar{Y}_c$ is recovered from β_1 in the primary ITT analysis. σ_u^2 represents the variance of the school level random effects and σ_ε^2 the variance of the pupil level random effects in the primary ITT analysis.

More specifically, we use Cohen's *d* ES for two reasons as follows:

- For multi-level models, such as those used in this evaluation, EEF (2018, p. 4) guidance recommends using Cohen's *d*.
- There is negligible difference between Hedges' *g* and Cohen's *d* in a trial of this size: as per EEF (2018, p. 4, footnote 10) guidance stating that "the difference between Hedges' g and Cohen's d is minimal for samples over 30 so either could be used in practice".

As per Adkins (2017), we will compute effect sizes directly in Stan. For MLwiN, effect sizes will be computed from the saved MCMC simulation values within R (Adkins, 2017). In lme4, the sim() function from the Applied Regression Modelling package (arm) in R will be used to compute the classically derived estimates using the same methodology (Adkins, 2017). As Adkins (2017) notes, credible/confidence intervals can be read off the summary report across all three processes.

References

Adkins, M. (2017). *Statistical analysis plan for Catch Up ® Numeracy*. London: Education Endowment Foundation.

Anders, J. & Shure, N. (2018). *Statistical analysis plan: Craft of Writing.* London: Education Endowment Foundation.

Benton, T., & Sutch, T. (2014). *Analysis of use of Key Stage 2 data in GCSE predictions. Report commissoned by Ofqual Ref:14/5471*. Coventry: Office of Qualifications and Examinations Regulation.

EEF. (2015). *Intra-cluster correlation coefficients* Retrieved from https://educationendowmentfoundation.org.uk/public/files/Evaluation/Writing_a_Protocol_or_SAP/ICC_2015.pdf

EEF (2016) Classification of the security of findings from EEF evaluations. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Evaluation/Carrying_out_a_Peer_Review/2016_Classifying_the_security_of_EEF_findings.pdf

EEF. (2018). *Statistical analysis guidance*. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Grantee_guide_and_EEF_policies/Evaluation/Writing_a_Protocol_or_SAP/EEF_statistical_analysis_guidance_2018.pdf

Fields, R. D. (2009). Making memories stick. *Scientific American*, *120* (February), 5. http://doi.org/10.1126/scisignal.2009ec285

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., & Rubin, D.B. (2014) *Bayesian Data Analysis, 3rd edition,* Boca Raton, FL: CRC Press.

Imai, K., King, G., & Nall, C. (2009). The essential role of pair matching in cluster-randomized experiments, with application to the Mexican Universal Health Insurance Evaluation. *Statistical Science*, 24(1), 29–53.

Imbens, G. M. & D. B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* New York, NY, Cambridge University Press.

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review, 25*(1), 178-206. doi:10.3758/s13423-016-1221-4

Kruschke, J. K. (2018). Rejecting or Accepting Parameter Values in Bayesian Estimation. *Advances in Methods and Practices in Psychological Science, 1*(2), 270–280. doi:10.1177/2515245918771304

Morgan, K. L., & Rubin, D. B. (2012). Re-randomisation to improve covariate balance in experiments. *The Annals of Statistics*, 40(2), 1263-1282.

Nunes, T., Bryant, P., Strand, S., Hillier, J., Barros, R., & Miller-Friedman, J. (2017). *Review of SES and Science Learning in Formal Educational Settings: A Report Prepared for the EEF and the Royal Society.* London: Educational Endowment Foundation.

OFQUAL. (2015). *GCSE Subject Level Conditions and Requirements for Combined Science*. Coventry: OFQUAL.

O'Hare, L., Stark, P., McGuinness, C., Biggart, A., & Thurston, A. (2017). *Spaced Learning: The Design, Feasibility and Optimisation of SMART Spaces: Evaluation report and executive summary*. London: Education Endowment Foundation.

Royal Society. (2014). *Vision for science and mathematics education.* London: The Royal Society.

Rubin, D. (2004). *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley and Sons.

Rubin, D. (2008). Comment: The design and analysis of gold standard randomized experiments. *Journal of the American Statistical Association*, 103(484), 1350–1353.

TISME. (2013). *What influences participation in science and mathematics? A briefing paper from the Targeted Initiative on Science and Mathematics Education (TISME)*. London: TISME / King's College London.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3), 67. doi:10.18637/jss.v045.i03

Appendix 1

Simulation code in Stata set out below is essentially the same as the Batch 1 randomisation code except that, after defining the randomisation program, there is simulation code rather than a direct call on the program:

```
set more off
cap log close
clear
cd "S:\SMART_Spaces_Evaluation\Data Confirmed"
log using pairedrandomisation.log, replace

// Set a random seed.
set seed 99823


// Import the list of schools from the Data Return Record list
import excel using "SMART_Spaces_Data_Return_Record.xlsx", ///
        sheet("Sheet1") cellrange(A1:AF115) firstrow

// Drop unless it is confirmed that we've received the dataset
keep if Readyforrandomisation=="YES"
// Drop unless School SMART ID and LAESTAB are present and in right format
keep if SchoolSMARTID!=""
keep if LAEstabDfENumber!=.


// Loop through all observations grabbing the KS2 mean and school size data
//from individual school data spreadsheets
local N = _N
quietly generate ks2_schoolmean = .
quietly generate school_size = .
quietly generate sheet_laestab = .
quietly generate sheet_doubleaward = .
quietly generate sheet_withdrawn = .
quietly generate fsmprop = .


forvalues i=1/`N' {
        local schoolid = SchoolSMARTID[`i']

        preserve // Preserve the overall data file
        clear
        capture import excel using "School Data - ready for randomisation/`schoolid'.xlsx",
        /// Open the spreadsheet provided by the school identified with their URN
                sheet(School and teacher information) cellrange(B6:B23)
        if _rc!=0 {
                di "Import failed for `schoolid'"
                exit
                }
        di "Currently processing data for `schoolid'"
        local sheet_laestab = B[1] //Grab LAESTAB
        local sheet_doubleaward = B[16] //Grab number of double award students
        local sheet_withdrawn = B[17] //Grab number of withdrawn students
        local sheet_confirm = ""
        local sheet_confirm = B[18] // Grab Confirmation

        if "`sheet_confirm'" != "Yes" & "`sheet_confirm'" != "YES" & "`sheet_confirm'" != "yes" {
                di "Withdrawal procedures not confirmed for school `schoolid'"
                // Check the resulting log for any schools where withdrawal procedures have not been
confirmed
```

```
		}

	restore

	quietly replace sheet_doubleaward = `sheet_doubleaward' if _n==`i' // Put no. of DA students into main dataset
	quietly replace sheet_withdrawn = `sheet_withdrawn' if _n==`i' // Put no of withdrawn students into main dataset
	quietly replace sheet_laestab = `sheet_laestab' if _n==`i' // Put LAESTAB into main dataset

	preserve
	clear
	capture import excel using "School Data - ready for randomisation/`schoolid'.xlsx", /// Import pupil data from the same school spreadsheets
		sheet(Pupil information) firstrow

	drop if UniquePupilNumberUPN=="" // Only keep lines with UPNs (drops lines that are not people)

	local pupilnum = .
		local pupilnum = _N //number of data rows i.e. number of pupils with data submitted in that school
	di _N

	keep KS2PupilAverage EvereligibleforFSMYN //only keep pupils average KS2 level and ever FSM

	cap tostring EvereligibleforFSMYN, replace
	quietly gen FSM = 0 // Lots of blanks for FSM and have verified that these are intended as meaning they are not FSM
	quietly replace FSM = 1 if EvereligibleforFSMYN=="Y"
	quietly replace FSM = 1 if EvereligibleforFSMYN=="y"
	quietly replace FSM = 1 if EvereligibleforFSMYN=="Yes"
	quietly replace FSM = 1 if EvereligibleforFSMYN=="YES"
	quietly replace FSM = 1 if EvereligibleforFSMYN=="yes"
	quietly replace FSM = 1 if EvereligibleforFSMYN=="T"
	quietly replace FSM = 1 if EvereligibleforFSMYN=="t"
	quietly replace FSM = 1 if EvereligibleforFSMYN=="True"
	quietly replace FSM = 1 if EvereligibleforFSMYN=="TRUE"
	quietly replace FSM = 1 if EvereligibleforFSMYN=="true"
	quietly replace FSM = 1 if EvereligibleforFSMYN=="1"
	local fsmprop ""
	quietly sum FSM // Work out the proportion flagged as FSM
	local fsmprop = r(mean) // Save it as a macro to put back into main spreadsheet

	local ks2_schoolmean = .
	quietly summarize KS2PupilAverage //work out school average KS2 level
	local ks2_schoolmean = r(mean) //save it as a macro to put back into main spreadsheet

	restore
	quietly replace school_size = `pupilnum' if _n==`i' // Put no of data rows into main spreadsheet
	quietly replace ks2_schoolmean = `ks2_schoolmean' if _n==`i' // Put school average KS2 into main spreadsheet
	quietly replace fsmprop = `fsmprop' if _n==`i' // Put proportion FSM into main spreadsheet
	}

assert Numberofstudents==sheet_doubleaward - sheet_withdrawn // Verification checks on the no of pupils we have recorded on our sheets and their sheets to force manual verification if there are anomalies
assert Numberofstudents==school_size //verification that data rows is equal to the number of students
//assert LAEstabDfENumber==sheet_laestab // Verification checks on the LAEstabs we have recorded on our sheets and their sheets to force manual verification if there are anomalies
```

```
assert school_size<. //Check school size has been successfully produced for all schools
assert ks2_schoolmean<. //Chack KS2 school mean has been successfully produced for all schools
assert fsmprop<. // Check that an EAL proportion has been successfully produced for all schools

xtile ks2_schoolmean16 = ks2_schoolmean, nq(16)
xtile ks2_schoolmean16_wt = ks2_schoolmean [fw=school_size], nq(16)

*** STANDARDISE KS2 AND FSM VARS
cap sum ks2_schoolmean, de
gen std_ks2_schoolmean = (ks2_schoolmean - r(mean))/r(sd)

cap sum ks2_schoolmean [fw=school_size], de
gen stdwt_ks2_schoolmean = (ks2_schoolmean - r(mean))/r(sd)

cap sum fsmprop, de
gen std_fsmprop = (fsmprop - r(mean))/r(sd)

cap sum fsmprop [fw=school_size], de
gen stdwt_fsmprop = (fsmprop - r(mean))/r(sd)

*** DEFINE RANDOMISATION PROGRAMME
cap program drop randomise
program define randomise, rclass

        cap drop random
        cap drop treatment

        gen double random = runiform()

        sort ks2_schoolmean16 random
        egen treatment = fill(1 0 1 0 0 1 0 1 0 0)

        regress stdwt_ks2_schoolmean treatment [aw=school_size]
        return scalar balance_ks2_weight = _b[treatment]

        regress std_ks2_schoolmean treatment
        return scalar balance_ks2_unweight = _b[treatment]

        regress stdwt_fsmprop treatment [aw=school_size]
        return scalar balance_fsm_weight = _b[treatment]

        regress std_fsmprop treatment
        return scalar balance_fsm_unweight = _b[treatment]

        regress school_size treatment
        return scalar balance_school_size_unweight = _b[treatment]

        sum treatment
        return scalar treat_prop = r(mean)

        end

*** RUN SIMULATIONS
preserve
simulate treat_prop = r(treat_prop) balance_ks2_unweight=r(balance_ks2_unweight)
balance_ks2_weight=r(balance_ks2_weight) balance_fsm_unweight = r(balance_fsm_unweight)
balance_fsm_weight=r(balance_fsm_weight) balance_school_size_unweight =
r(balance_school_size_unweight), reps(1000): randomise
sum balance_ks2_weight balance_ks2_unweight balance_fsm_weight balance_fsm_unweight
balance_school_size_unweight, de
restore
exit
```

Appendix 2

Batch 1 randomisation code in Stata, showing seed incrementation for re-randomisation 1-6:

```
set more off
cap log close

clear
cd "S:\SMART_Spaces_Evaluation\Data Confirmed"
log using pairedrandomisation.log, replace

// Set a random seed. Never run more than once without restarting Stata or risk it won't be replicable
//set seed 8148 // Value of 1 GBP to Thai Baht at 6.21pm 16-10-18
//set sortseed 52048 //Value of 1 GBP to Turkish Lira at 6.21pm 16-10-18

//Randomisation 2
//set seed 8149 // Add 1 to seed above
//set sortseed 52049 //Add 1 to sortseed above

//Randomisation 3
//set seed 8150 // Add 1 to seed above
//set sortseed 52050 //Add 1 to sortseed above

//Randomisation 4
//set seed 8151 // Add 1 to seed above
//set sortseed 52051 //Add 1 to sortseed above

//Randomisation 5
//set seed 8152 // Add 1 to seed above
//set sortseed 52052 //Add 1 to sortseed above

//Randomisation 6
set seed 8153 // Add 1 to seed above
set sortseed 52053 //Add 1 to sortseed above

// Import the list of schools from the Data Return Record list
import excel using "SMART_Spaces_Data_Return_Record.xlsx", ///
        sheet("Sheet1") cellrange(A1:AF115) firstrow

// Drop unless it is confirmed that we've received the dataset
keep if Readyforrandomisation=="YES"
// Drop unless School SMART ID and LAESTAB are present and in right format
keep if SchoolSMARTID!=""
keep if LAEstabDfENumber!=.


// Loop through all observations grabbing the KS2 mean and school size data from individual school
data spreadsheets
local N = _N
quietly generate ks2_schoolmean = .
quietly generate school_size = .
quietly generate sheet_laestab = .
quietly generate sheet_doubleaward = .
quietly generate sheet_withdrawn = .
quietly generate fsmprop = .


forvalues i=1/`N' {
        local schoolid = SchoolSMARTID[`i']

        preserve // Preserve the overall data file
        clear
```

```
	capture import excel using "School Data - ready for randomisation/`schoolid'.xlsx",
	/// Open the spreadsheet provided by the school identified with their URN
		sheet(School and teacher information) cellrange(B6:B23)
	if _rc!=0 {
		di "Import failed for `schoolid'"
		exit
		}
	di "Currently processing data for `schoolid'"
	local sheet_laestab = B[1] //Grab LAESTAB
	local sheet_doubleaward = B[16] //Grab number of double award students
	local sheet_withdrawn = B[17] //Grab number of withdrawn students
	local sheet_confirm = ""
	local sheet_confirm = B[18] // Grab Confirmation

	if "`sheet_confirm'" != "Yes" & "`sheet_confirm'" != "YES" & "`sheet_confirm'" != "yes" {
		di "Withdrawal procedures not confirmed for school `schoolid'"
		// Check the resulting log for any schools where withdrawal procedures have not been
confirmed
		}

	restore


	quietly replace sheet_doubleaward = `sheet_doubleaward' if _n==`i' // Put no. of DA students
into main dataset
	quietly replace sheet_withdrawn = `sheet_withdrawn' if _n==`i' // Put no of withdrawn students
into main dataset
	quietly replace sheet_laestab = `sheet_laestab' if _n==`i' // Put LAESTAB into main dataset

	preserve
	clear
	capture import excel using "School Data - ready for randomisation/`schoolid'.xlsx",
	/// Import pupil data from the same school spreadsheets
		sheet(Pupil information) firstrow

	drop if UniquePupilNumberUPN=="" // Only keep lines with UPNs (drops lines that are not
people)

	local pupilnum = .
	local pupilnum = _N //number of data rows i.e. number of pupils with data submitted in that
school
	di _N

	keep KS2PupilAverage EvereligibleforFSMYN //only keep pupils average KS2 level and ever
FSM

	cap tostring EvereligibleforFSMYN, replace
	quietly gen FSM = 0 // Lots of blanks for FSM and have verified that these are intended as
meaning they are not FSM
	quietly replace FSM = 1 if EvereligibleforFSMYN=="Y"
	quietly replace FSM = 1 if EvereligibleforFSMYN=="y"
	quietly replace FSM = 1 if EvereligibleforFSMYN=="Yes"
	quietly replace FSM = 1 if EvereligibleforFSMYN=="YES"
	quietly replace FSM = 1 if EvereligibleforFSMYN=="yes"
	quietly replace FSM = 1 if EvereligibleforFSMYN=="T"
	quietly replace FSM = 1 if EvereligibleforFSMYN=="t"
	quietly replace FSM = 1 if EvereligibleforFSMYN=="True"
	quietly replace FSM = 1 if EvereligibleforFSMYN=="TRUE"
	quietly replace FSM = 1 if EvereligibleforFSMYN=="true"
	quietly replace FSM = 1 if EvereligibleforFSMYN=="1"
	local fsmprop ""
	quietly sum FSM // Work out the proportion flagged as FSM
	local fsmprop = r(mean) // Save it as a macro to put back into main spreadsheet
```

```
		local ks2_schoolmean = .
		quietly summarize KS2PupilAverage //work out school average KS2 level
		local ks2_schoolmean = r(mean) //save it as a macro to put back into main spreadsheet

		restore
		quietly replace school_size = `pupilnum' if _n==`i' // Put no of data rows into main
spreadsheet
		quietly replace ks2_schoolmean = `ks2_schoolmean' if _n==`i' // Put school average KS2 into
main spreadsheet
		quietly replace fsmprop = `fsmprop' if _n==`i' // Put proportion FSM into main spreadsheet

		}

assert Numberofstudents==sheet_doubleaward - sheet_withdrawn // Verification checks on the no of
pupils we have recorded on our sheets and their sheets to force manual verification if there are
anomalies
assert Numberofstudents==school_size //verification that data rows is equal to the number of students
//assert LAEstabDfENumber==sheet_laestab // Verification checks on the LAEstabs we have
recorded on our sheets and their sheets to force manual verification if there are anomalies
assert school_size<. //Check school size has been successfully produced for all schools
assert ks2_schoolmean<. //Chack KS2 school mean has been successfully produced for all schools
assert fsmprop<. // Check that an EAL proportion has been successfully produced for all schools


xtile ks2_schoolmean16 = ks2_schoolmean, nq(16)
xtile ks2_schoolmean16_wt = ks2_schoolmean [fw=school_size], nq(16)

*** STANDARDISE KS2 AND FSM VARS
cap sum ks2_schoolmean, de
gen std_ks2_schoolmean = (ks2_schoolmean - r(mean))/r(sd)

cap sum ks2_schoolmean [fw=school_size], de
gen stdwt_ks2_schoolmean = (ks2_schoolmean - r(mean))/r(sd)

cap sum fsmprop, de
gen std_fsmprop = (fsmprop - r(mean))/r(sd)

cap sum fsmprop [fw=school_size], de
gen stdwt_fsmprop = (fsmprop - r(mean))/r(sd)


*** DEFINE RANDOMISATION PROGRAMME
cap program drop randomise
program define randomise, rclass

		cap drop random
		cap drop treatment

		gen double random = runiform()

		sort ks2_schoolmean16 random
		egen treatment = fill(1 0 1 0 0 1 0 1 0 0)

		regress stdwt_ks2_schoolmean treatment [aw=school_size]
		return scalar balance_ks2_weight = _b[treatment]

		regress std_ks2_schoolmean treatment
		return scalar balance_ks2_unweight = _b[treatment]

		regress stdwt_fsmprop treatment [aw=school_size]
		return scalar balance_fsm_weight = _b[treatment]
```

```
        regress std_fsmprop treatment
        return scalar balance_fsm_unweight = _b[treatment]

        regress school_size treatment
        return scalar balance_school_size_unweight = _b[treatment]

        sum treatment
        return scalar treat_prop = r(mean)

        end


randomise

// Just check that has worked and we have intended treatment allocation
label define treatment 0 "Control" 1 "Treatment", replace
label val treatment treatment
tab treatment

rename ks2_schoolmean16 KS2Blocks
rename treatment Treatment
rename random Random

// Export a spreadsheet for internal records
export excel using "Randomisation Outcome.xlsx", ///
        replace firstrow(variables) cell(A1) sheet("Allocation")

// Remove some extraneous detail that doesn't need to be in the spreadsheet shared with project grop
keep SchoolSMARTID SchoolName Headteacheremail Schoolcontactname Schoolcontactemail
DataManagername DataManageremail LAEstabDfENumber Treatment

// Export a spreadsheet to share with the project team
export excel using "Randomisation Outcome to Project Team.xlsx", ///
        replace firstrow(variables) cell(A1) sheet("Allocation")

log close
exit
```

Appendix 3

Batch 2 randomisation code in Stata. The code is the same as for Batch 1 randomisation except that randomisation was successful at the first attempt (so no need for seed incrementation), eight blocks are specified based on KS2 school average and allocation to treatment and control was in turn:

```
set more off

cap log close

clear
cd "S:\SMART_Spaces_Evaluation\Data Confirmed"
log using pairedrandomisation_batch2_1.log, replace

//Set a random seed. Never run more than once without restarting Stata or risk it won't be replicable
set seed 7568 // Value of 1 GBP to Thai Baht at 10.28am on 5-12-18
set sortseed 85416 //Value of 1 GBP to Turkish Lira at 10.28am on 5-12-18


// Import the list of schools from the Data Return Record list
import excel using "SMART_Spaces_Data_Return_Record.xlsx", ///
        sheet("Sheet1") cellrange(A1:AG145) firstrow

// Drop unless it is confirmed that we've received the dataset
keep if Readyforrandomisation=="YES"
// Drop unless school is allocated to randomisation batch 2
keep if Randomisationbatch==2
// Drop unless School SMART ID and LAESTAB are present and in right format
keep if SchoolSMARTID!=""
keep if LAEstabDfENumber!=.


// Loop through all observations grabbing the KS2 mean and school size data from individual school data spreadsheets
local N = _N
quietly generate ks2_schoolmean = .
quietly generate school_size = .
quietly generate sheet_laestab = .
quietly generate sheet_doubleaward = .
quietly generate sheet_withdrawn = .
quietly generate fsmprop = .


forvalues i=1/`N' {
        local schoolid = SchoolSMARTID[`i']

        preserve // Preserve the overall data file
        clear
        capture import excel using "School Data - ready for randomisation/`schoolid'.xlsx",
        /// Open the spreadsheet provided by the school identified with their URN
                sheet(School and teacher information) cellrange(B6:B23)
        if _rc!=0 {
                di "Import failed for `schoolid'"
                exit
                }
        di "Currently processing data for `schoolid'"
        local sheet_laestab = B[1] //Grab LAESTAB
        local sheet_doubleaward = B[16] //Grab number of double award students
        local sheet_withdrawn = B[17] //Grab number of withdrawn students
        local sheet_confirm = ""
```

```
	local sheet_confirm = B[18] // Grab Confirmation

	if "`sheet_confirm'" != "Yes" & "`sheet_confirm'" != "YES" & "`sheet_confirm'" != "yes" {
		di "Withdrawal procedures not confirmed for school `schoolid'"
		// Check the resulting log for any schools where withdrawal procedures have not been confirmed
		}

	restore


	quietly replace sheet_doubleaward = `sheet_doubleaward' if _n==`i' // Put no. of DA students into main dataset
	quietly replace sheet_withdrawn = `sheet_withdrawn' if _n==`i' // Put no of withdrawn students into main dataset
	quietly replace sheet_laestab = `sheet_laestab' if _n==`i' // Put LAESTAB into main dataset

	preserve
	clear
	capture import excel using "School Data - ready for randomisation/`schoolid'.xlsx", /// Import pupil data from the same school spreadsheets
		sheet(Pupil information) firstrow

	drop if UniquePupilNumberUPN=="" // Only keep lines with UPNs (drops lines that are not people)

	local pupilnum = .
	local pupilnum = _N //number of data rows i.e. number of pupils with data submitted in that school
	di _N

	keep KS2PupilAverage EvereligibleforFSMYN //only keep pupils average KS2 level and ever FSM

	cap tostring EvereligibleforFSMYN, replace
	quietly gen FSM = 0 // Lots of blanks for FSM and have verified that these are intended as meaning they are not FSM
	quietly replace FSM = 1 if EvereligibleforFSMYN=="Y"
	quietly replace FSM = 1 if EvereligibleforFSMYN=="y"
	quietly replace FSM = 1 if EvereligibleforFSMYN=="Yes"
	quietly replace FSM = 1 if EvereligibleforFSMYN=="YES"
	quietly replace FSM = 1 if EvereligibleforFSMYN=="yes"
	quietly replace FSM = 1 if EvereligibleforFSMYN=="T"
	quietly replace FSM = 1 if EvereligibleforFSMYN=="t"
	quietly replace FSM = 1 if EvereligibleforFSMYN=="True"
	quietly replace FSM = 1 if EvereligibleforFSMYN=="TRUE"
	quietly replace FSM = 1 if EvereligibleforFSMYN=="true"
	quietly replace FSM = 1 if EvereligibleforFSMYN=="1"
	local fsmprop ""
	quietly sum FSM // Work out the proportion flagged as FSM
	local fsmprop = r(mean) // Save it as a macro to put back into main spreadsheet

	local ks2_schoolmean = .
	quietly summarize KS2PupilAverage //work out school average KS2 level
	local ks2_schoolmean = r(mean) //save it as a macro to put back into main spreadsheet

	restore
	quietly replace school_size = `pupilnum' if _n==`i' // Put no of data rows into main spreadsheet
	quietly replace ks2_schoolmean = `ks2_schoolmean' if _n==`i' // Put school average KS2 into main spreadsheet
	quietly replace fsmprop = `fsmprop' if _n==`i' // Put proportion FSM into main spreadsheet
```

```
	}

assert Numberofstudents==sheet_doubleaward - sheet_withdrawn // Verification checks on the no of pupils we have recorded on our sheets and their sheets to force manual verification if there are anomalies
assert Numberofstudents==school_size //verification that data rows is equal to the number of students
assert LAEstabDfENumber==sheet_laestab // Verification checks on the LAEstabs we have recorded on our sheets and their sheets to force manual verification if there are anomalies
assert school_size<. //Check school size has been successfully produced for all schools
assert ks2_schoolmean<. //Chack KS2 school mean has been successfully produced for all schools
assert fsmprop<. // Check that an FSM proportion has been successfully produced for all schools


xtile ks2_schoolmean8 = ks2_schoolmean, nq(8)

*** STANDARDISE KS2 AND FSM VARS
cap sum ks2_schoolmean, de
gen std_ks2_schoolmean = (ks2_schoolmean - r(mean))/r(sd)

cap sum ks2_schoolmean [fw=school_size], de
gen stdwt_ks2_schoolmean = (ks2_schoolmean - r(mean))/r(sd)

cap sum fsmprop, de
gen std_fsmprop = (fsmprop - r(mean))/r(sd)

cap sum fsmprop [fw=school_size], de
gen stdwt_fsmprop = (fsmprop - r(mean))/r(sd)


*** DEFINE RANDOMISATION PROGRAMME
cap program drop randomise
program define randomise, rclass

	cap drop random
	cap drop treatment

	gen double random = runiform()

	sort ks2_schoolmean8 random
	egen treatment = fill(0 1 0 1)

	regress stdwt_ks2_schoolmean treatment [aw=school_size]
	return scalar balance_ks2_weight = _b[treatment]

	regress std_ks2_schoolmean treatment
	return scalar balance_ks2_unweight = _b[treatment]

	regress stdwt_fsmprop treatment [aw=school_size]
	return scalar balance_fsm_weight = _b[treatment]

	regress std_fsmprop treatment
	return scalar balance_fsm_unweight = _b[treatment]

	regress school_size treatment
	return scalar balance_school_size_unweight = _b[treatment]

	sum treatment
	return scalar treat_prop = r(mean)

	end

randomise
```

```
// Just check that has worked and we have intended treatment allocation
label define treatment 0 "Control" 1 "Treatment", replace
label val treatment treatment
tab treatment

rename ks2_schoolmean8 KS2Blocks
rename treatment Treatment
rename random Random

// Export a spreadsheet for internal records
export excel using "Batch 2 Randomisation Outcome.xlsx", ///
        replace firstrow(variables) cell(A1) sheet("Allocation")

// Remove some extraneous detail that doesn't need to be in the spreadsheet shared with project grop
keep SchoolSMARTID SchoolName Headteacheremail Schoolcontactname Schoolcontactemail
DataManagername DataManageremail LAEstabDfENumber Treatment

// Export a spreadsheet to share with the project team
export excel using "Batch 2 Randomisation Outcome to Project Team.xlsx", ///
        replace firstrow(variables) cell(A1) sheet("Allocation")

log close
exit
```

Appendix 4

Batch 1 Randomisation record: Carried out by NB & JH 16/10/18

Procedure:
1. Initial Stata status set using the decimal places of the Thai Bhat and Turkish Lira at 6.23pm 16/10/18 (XE Corporation) for the seed (8148) and sortseed (52048).
2. Simulations suggest that re-randomisation will be triggered by the threshold of 0.05 (for standardised differences), so we will attempt to achieve balance on pupil-level KS2, but not on FSM.
3. If re-randomisation. triggered, we will inform Jake Anders by email, then save the log for each attempt, then shut down and reopen Stata, then rerun the randomisation code with seed and sortseed increased by 1.

Remember, for the SAP, include a statement that our existing plan to include KS2 attainment as a covariate is the appropriate adjustment for both our re-randomisation and our blocking strategies.

Record of Randomisation attempts

Appropriate balance was achieved on 6th Attempt

Randomisation attempt 1: Triggered re-randomisation.
Pupil-level KS2: 0.074 (p=.748)
School-level KS2: 0.008 (p=.971)
School-level FSM: -0.425 (p=.058)

Randomisation attempt 2: Triggered re-randomisation.
Pupil-level KS2: -0.084 (p=.710)
School-level KS2: -0.217 (p=.339)
School-level FSM: -0.243 (p=.284)

Randomisation attempt 3: Triggered re-randomisation.
Pupil-level KS2: 0.197 (p=.386)
School-level KS2: -0.088 (p=.700)
School-level FSM: 0.150 (p=.508)

Randomisation attempt 4: Triggered re-randomisation.
Pupil-level KS2: -0.182 (p=.428)
School-level KS2: -0.177 (p=.436)
School-level FSM: -0.061 (p=.790)

Randomisation attempt 5: Triggered re-randomisation.
Pupil-level KS2: -0.233 (p=.313)
School-level KS2: 0.038 (p=.868)
School-level FSM: 0.198 (p=.383)

Randomisation attempt 6: SUCCESSFUL.
Pupil-level KS2: 0.027 (p=.906)
School-level KS2: -0.063 (p=.783)
School-level FSM: 0.397 (p=.078)

Of the 82 schools in the batch: 33 allocated to the intervention, 49 to the control.

Spreadsheet outputs checked

Appendix 5

Batch 2 Randomisation record: carried out by NB & JH 5-12-18

Initial runs of the code resulted in the following data checks:

- 2 filenames corrected to SMART username
- 5 schools LAEstab number corrected

Simulation results

- Simple randomisation (without grouping), 1000 simulations, produced (balance ks2 weight) mean = -0.0180 ; sd = 0.310 . The sd is high.
- Randomisation with grouping nq(8), 1000 simulations, produced (balance ks2 weight) mean = -0.043 ; sd = 0.143

Procedure:

1. Initial Stata status set using the decimal places of the Thai Bhat and Turkish Lira at 10.28am 5/12/18 (XE Corporation) for the seed (7568) and sortseed (85416).
2. Calculations suggest that re-randomisation will be triggered by lower bound -.210 and upper bound 0.098 .
3. If re-randomisation. triggered, we will inform Jake Anders by email, then save the log for each attempt, then shut down and reopen Stata, then rerun the randomisation code with seed and sortseed increased by 1.

Remember, for the SAP, include a statement that our existing plan to include KS2 attainment as a covariate is the appropriate adjustment for both our re-randomisation and our blocking strategies.

Record of Randomisation attempts

Appropriate balance was achieved on 1st Attempt

Randomisation attempt 1
Pupil-level KS2: -0.100 (p=.750)
School-level KS2: -0.004 (p=.990)
School-level FSM: 0.328 (p=.288)

Of the 43 schools in batch 2: 21 allocated to the intervention, 22 to the control.

Spreadsheet outputs checked

Appendix 6

Scatter-plots and correlation for:

(i) school-level KS2 mean against school sample-size

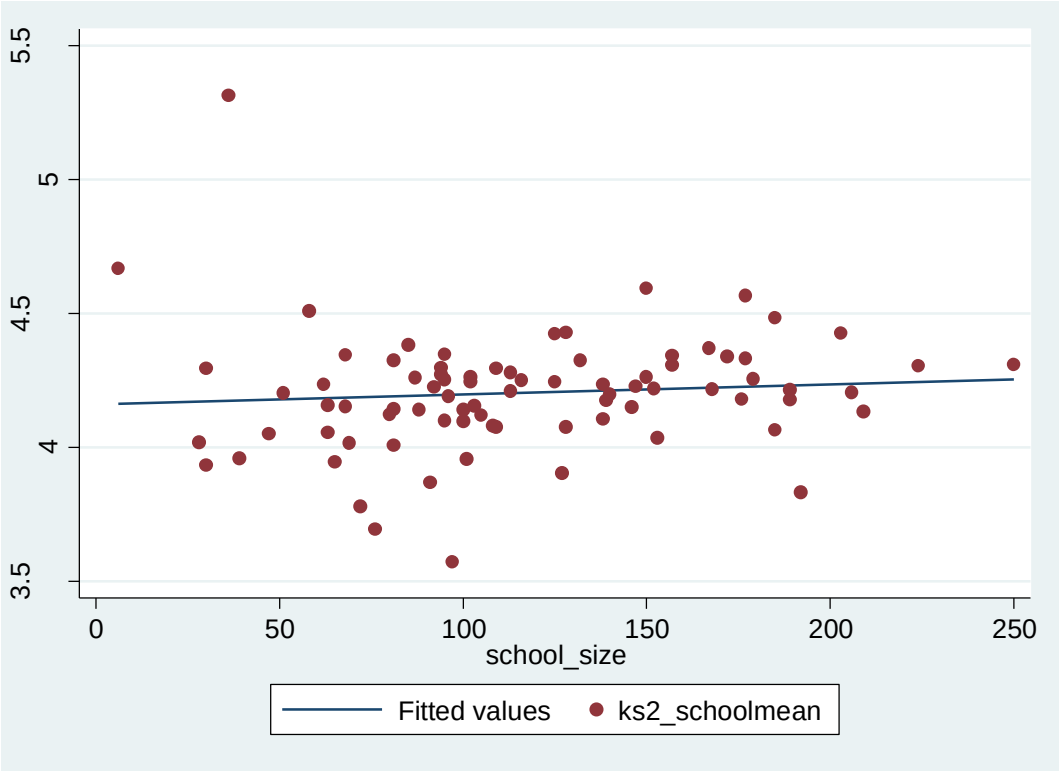
(ii) school-level KS2 mean against FSM proportion

(iii) school sample-size against FSM proportion

Observations: 82 schools

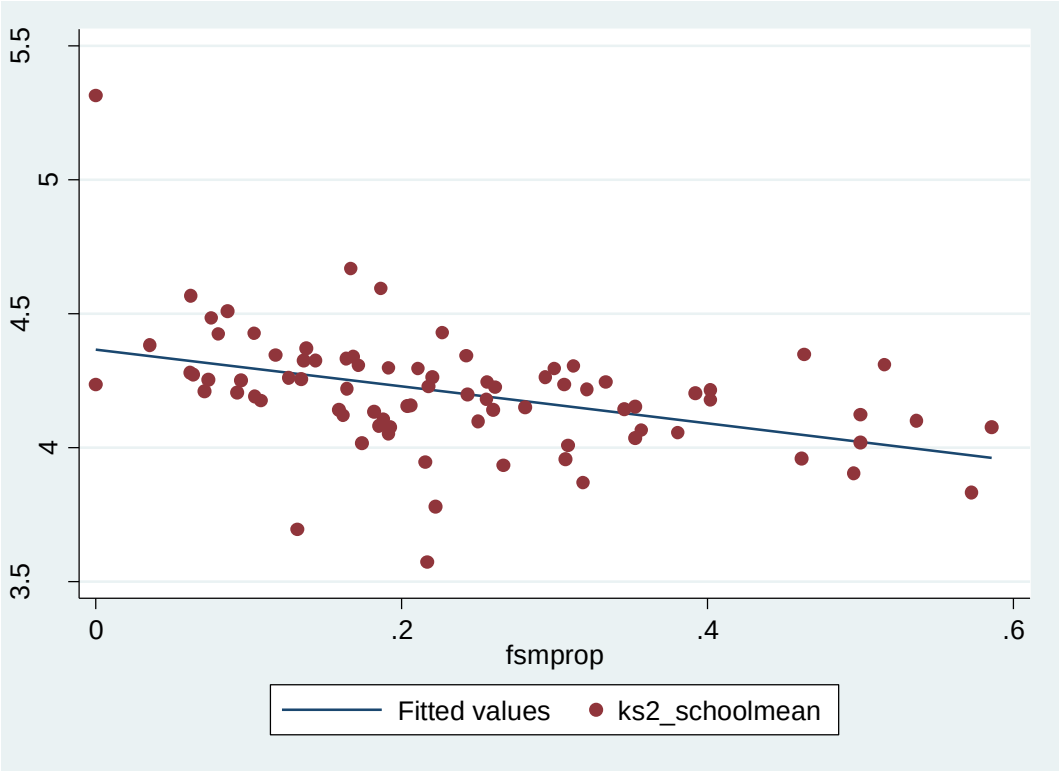
(i) School-level KS2 mean against school sample-size

Correlation: 0.0853



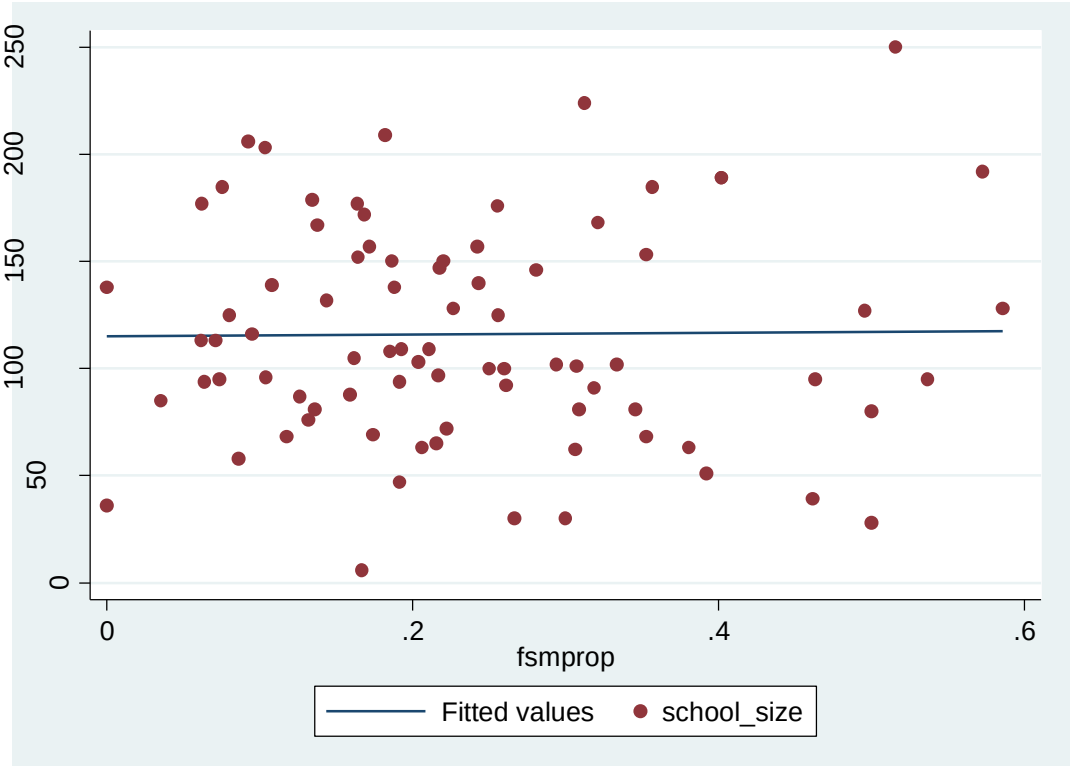
(ii) School-level KS2 mean against FSM proportion

Correlation: -0.4198



(iii) **School sample-size against FSM proportion**

Correlation: 0.0108



Appendix 7

SMART Spaces – Main Trial IPE pupil paper-based survey

Draft notes on analysis of QUB engagement items, NB 15-3-19

17 engagement items, Likert scale 1= strongly disagree; 4 strongly agree
3 items highlighted in yellow seem to be negatively worded.
-999.0 = missing

1. I was enthusiastic to try Spaced Learning
2. The Spaced learning lessons helped me learn more easily than normal lessons
3. I think Spaced Learning works well for revision
4. The spaced Learning lessons ran smoothly
5. The space activity activities (e.g. juggling) were helpful
6. Spaced learning would work for all classes, not just revision
7. I think 3 lessons was enough time for Spaced Learning.
8. The class as a whole enjoyed Spaced learning
9. The teacher was confident at delivering Spaced Learning
10. I felt more motivated to learn during Spaced Learning than in normal classes
11. I would be happy to try Spaced learning again in the future
12. I think Spaced Learning would also be useful for other subjects
13. I found the spaced learning lessons tiring
14. I found the spaced learning lessons fun
15. I found the spaced learning lessons helpful for revision
16. I found Spaced Learning too repetitive
17. I did not enjoy the space activities (e.g. juggling)

In SPSS, checked frequencies – categories %s for valid responses correspond with those in the OS report.

Results of exploratory factor analysis in SPSS.

Interpretation based on support notes at:

https://www.ibm.com/support/knowledgecenter/pt-br/SSLVMB_24.0.0/spss/tutorials/fac_telco_communalities.html#fac_telco_communalities

Principal components analysis, no rotation

KMO measure of sampling adequacy = .903

Bartlett's sphericity test: approx. chi sq = 1688.870, df = 136, sig. < .0005

Interpretation: high KMO close to 1 suggests factor analysis might be useful; similarly sig of Bartlett small, close to 0 suggests factor analysis useful

Initial communalities: for correlation analyses, the proportion of variance accounted for in each variable by the rest of the variables.

Extraction communalities: are estimates of the variance in each variable accounted for by the factors in the factor solution. Smaller values indicate items that don't fit well.

I was enthusiastic to try Spaced Learning	.638
The Spaced learning lessons helped me learn more easily than normal lessons	.598
I think Spaced Learning works well for revision	.699
The spaced Learning lessons ran smoothly	.378
The space activity activities (e.g. juggling) were helpful	.481
Spaced learning would work for all classes, not just revision	.606
I think 3 lessons was enough time for Spaced Learning.	.585
The class as a whole enjoyed Spaced learning	.609

The teacher was confident at delivering Spaced Learning	.603
I felt more motivated to learn during Spaced Learning than in normal classes	.664
I would be happy to try Spaced learning again in the future	.677
I think Spaced Learning would also be useful for other subjects	.661
I found the spaced learning lessons tiring	.578
I found the spaced learning lessons fun	.657
I found the spaced learning lessons helpful for revision	.644
I found Spaced Learning too repetitive	.536
I did not enjoy the space activities (e.g. juggling)	.812

Loaded mostly onto one factor, see below.
Total variance explained by 1st four factors = 61.3%
Note: initial eigenvalues = extraction sums of squared loadings

Component	total	% of variance
1	6.57	38.6
2	1.64	9.67
3	1.18	6.94
4	1.03	6.06

Absolute factor loadings > 0.2

I was enthusiastic to try Spaced Learning	**.693**		.382	
The Spaced learning lessons helped me learn more easily than normal lessons	**.750**			
I think Spaced Learning works well for revision	**.712**	.351	-.253	
The spaced Learning lessons ran smoothly	**.573**			
The space activity activities (e.g. juggling) were helpful	**.606**			-.295
Spaced learning would work for all classes, not just revision	**.631**		-.436	
I think 3 lessons was enough time for Spaced Learning.		**.541**		**-.518**
The class as a whole enjoyed Spaced learning	**.717**			
The teacher was confident at delivering Spaced Learning	**.497**		.431	.412
I felt more motivated to learn during Spaced Learning than in normal classes	**.812**			
I would be happy to try Spaced learning again in the future	**.817**			
I think Spaced Learning would also be useful for other subjects	**.742**		-.271	
I found the spaced learning lessons tiring		**.559**	.483	
I found the spaced learning lessons fun	**.780**		.205	
I found the spaced learning lessons helpful for revision	**.654**	.337	-.316	
I found Spaced Learning too repetitive	-.366	**.632**		
I did not enjoy the space activities (e.g. juggling)	-.292	**.562**	-.211	**.606**

Note items in grey also had a loading of >0.4 on another factor

Items with main loading on factor 1:

I was enthusiastic to try Spaced Learning
The Spaced learning lessons helped me learn more easily than normal lessons
I think Spaced Learning works well for revision
The spaced Learning lessons ran smoothly
The space activity activities (e.g. juggling) were helpful
Spaced learning would work for all classes, not just revision
The class as a whole enjoyed Spaced learning
The teacher was confident at delivering Spaced Learning
I felt more motivated to learn during Spaced Learning than in normal classes
I would be happy to try Spaced learning again in the future
I think Spaced Learning would also be useful for other subjects

I found the spaced learning lessons fun
I found the spaced learning lessons helpful for revision

Items with main loading and/or loading >0.5 on factor 2:

I think 3 lessons was enough time for Spaced Learning.
I found the spaced learning lessons tiring
I found Spaced Learning too repetitive .
I did not enjoy the space activities (e.g. juggling)

Note no items loaded mainly or >0.5 on factor 3, though several were >0.4

Items with loading >0.4 on factor 3:

Spaced learning would work for all classes, not just revision
The teacher was confident at delivering Spaced Learning
I found the spaced learning lessons tiring

Items with main loading and/or loading >0.5 on factor 4:

I think 3 lessons was enough time for Spaced Learning.
I did not enjoy the space activities (e.g. juggling)

Interpretation of factors

Most of the items loaded strongly on the first factor, which suggests they could operate reasonably well as a scale. Factor 2 seems to consist of the negatively phrased items. Note that 'I think 3 lessons was enough time for Spaced learning' might be interpreted as a euphemistic way of saying I don't want any more SMART Spaces lessons. In this sense, it could be interpreted as a negatively phrased item. Factors 3 & 4 are less easy to interpret.

Thoughts on item wording

The results of the factor analysis prompted a consideration of the wording of items, focussing primarily on those items that did not load mainly onto the first factor and those that did not fit the factor solution well.

The estimated variance in Items 4 and 5 (highlighted in blue) explained by the factor solution was relatively low. This indicates that these items do not fit well into the factors produced by PCA. This may be explained by analysing the wording of these items. For example, Item 4 'the spaced learning lessons ran smoothly' seems to tap fidelity issues, e.g. about whether the lesson went according to plan, rather than pupils' engagement necessarily. With Item 5, the 'The space activity activities (e.g. juggling) were helpful' seems clunky in its repetition of 'activity'; 'e.g. juggling' seems confusing if the children had done some alternative spacing activity other than juggling and it is unclear what 'helpful' means in this context. Hence it may be that students simply found this item confusing and as a result their responses do not fit well with other items. Phrasing as enjoyment rather than being helpful for some undefined purpose might improve this item e.g. 'In Spaced Learning lessons, I enjoyed the spacing activity between blocks of chemistry revision.'

Factor 2 seems to consist of negatively phrased items. In item 13, the word 'tiring' seems potentially to have multiple interpretations: is it the fast pace of the lesson that is tiring, or the physicality of spacing activities? Boring might be better. Item 16 asks for agreement with 'I found Spaced Learning too repetitive'. Of course, the lessons are designed to be repetitive, hence it makes more sense to whether there is too much repetition but again boring might be better here. The wording of item 17 is in terms of enjoyment which connects well with other items e.g. 1, 8, 10, 14. However, 'space

activities (e.g. juggling) may again cause confusion if it is not clear what these refer to. Note that 'I think 3 lessons was enough time for Spaced learning' might be interpreted as a euphemistic way of saying I don't want any more SMART Spaces lessons. In this sense, it could be interpreted as a negatively phrased item. On the other had it might be interpreted as saying the Spaced learning was sufficient for revision. In any case, there are 6 SMART Spaces lessons spread over two weeks so this item seems inappropriate for the main trial. In addition to these issues, negative wording of items can increase response bias through inattention and confusion (van Sonderen, Sanderman, and Coyne, 2013)[3].

Item 9 'The teacher was confident at delivering Spaced Learning' did not load highly on any factor and seemed to load fairly evenly across three factors (1,3,4). Again this item does not seem to tap pupil engagement but rather pupils' perception of teacher engagement – again rather more related to fidelity perhaps.

Item 6 'Spaced learning would work for all classes, not just revision' has a loading >0.4 on factor 3. Classes is an odd one here, it would make more sense if it said 'for all chemistry lessons' instead. Classes could easily refer to other science lessons or even subjects. For example, contrast this item with Item 12 'I think Spaced Learning would also be useful for other subjects', which loaded better on factor 1.

Factor analysis: second attempt

Removing the problematic items highlighted above, leaves the following 9 items.

1. I was enthusiastic to try Spaced Learning
2. The Spaced learning lessons helped me learn more easily than normal lessons
3. I think Spaced Learning works well for revision
4. The class as a whole enjoyed Spaced learning
5. I felt more motivated to learn during Spaced Learning than in normal classes
6. I would be happy to try Spaced learning again in the future
7. I think Spaced Learning would also be useful for other subjects
8. I found the spaced learning lessons fun
9. I found the spaced learning lessons helpful for revision

KMO measure of sampling adequacy = .898

Bartlett's sphericity test: approx. chi sq = 1156.982, df = 36, sig. < .0005

Interpretation: high KMO close to 1 suggests factor analysis might be useful; similarly sig of Bartlett small, close to 0 suggests factor analysis useful. So this looks fine.

Extraction communalities: are estimates of the variance in each variable accounted for by the factors in the factor solution. Smaller values indicate items that don't fit well. (only factors with Eigenvalues > 1 extracted i.e. only first factor extracted)

I was enthusiastic to try Spaced Learning	.511
The Spaced learning lessons helped me learn more easily than normal lessons	.511
I think Spaced Learning works well for revision	.534
The class as a whole enjoyed Spaced learning	.528
I felt more motivated to learn during Spaced Learning than in normal classes	.664

[3] Sonderen, E. v., Sanderman, R., & Coyne, J. C. (2013). Ineffectiveness of Reverse Wording of Questionnaire Items: Let's Learn from Cows in the Rain. PLoS ONE, 8(7), e68967. doi:10.1371/journal.pone.0068967

I would be happy to try Spaced learning again in the future	.690
I think Spaced Learning would also be useful for other subjects	.537
I found the spaced learning lessons fun	.621
I found the spaced learning lessons helpful for revision	.470

Factor 1 is the only one with eigenvalue > 1, and explains 56.7% of the variance.
Factor 2 has eigenvalue = .928, explaining 10.3% of the variance.
Factor loadings below > 0.2:

I was enthusiastic to try Spaced Learning	**.715**	
The Spaced learning lessons helped me learn more easily than normal lessons	**.742**	
I think Spaced Learning works well for revision	**.731**	.522
The class as a whole enjoyed Spaced learning	**.727**	-.474
I felt more motivated to learn during Spaced Learning than in normal classes	**.815**	
I would be happy to try Spaced learning again in the future	**.831**	
I think Spaced Learning would also be useful for other subjects	**.733**	
I found the spaced learning lessons fun	**.788**	-.302
I found the spaced learning lessons helpful for revision	**.686**	.517

So all 9 items load highly on the first factor.

The second factor loadings are mainly very small. The two items that load >0.5 on this factor are very similar and very specific to revision.

Summary: these 9 items look as though they would operate ok as a scale.

Cronbach's alpha

For all 17 items, **Cronbach's alpha = 0.822** , based on 236 valid cases.

SPSS output suggests Cronbach alpha would increase if the negative items were deleted.

I found the spaced learning lessons tiring
I found Spaced Learning too repetitive
I did not enjoy the space activities (e.g. juggling)
Also if item 'I think 3 lessons was enough time for Spaced Learning' was deleted.

Removing these four items, the remaining 13 'positive' items have **Cronbach's alpha = 0.911**, based on 239 cases.

I've just been scanning the Optimisation study and read their section on the engagement items. They say they used 13 items with Cronbach-alpha of 0.91. Although the OS refers to Appendix 2 for the engagement items, it does not specify which 13 of the 17 items listed they used to construct the scale. However, the 13 items with C-a=0.91 fits with my results from SPSS after removing the 4 negative items.

SPSS output suggests that deleting 'The teacher was confident at delivering Spaced Learning' wouldn't change Cronbach alpha. And it doesn't, remaining at 0.911.

With just the 9 items used in the second attempt factor analysis, Cronbach's alpha was 0.904, so not much reduced.

Factor analysis: third attempt

This time with only four negative items removed. So 13 items remaining.

KMO measure of sampling adequacy = .924
Bartlett's sphericity test: approx. chi sq = 1506.723, df = 78, sig. < .0005

Factor 1 has eigenvalue = 6.36, and explains 48.95% of the variance.
Factor 2 has eigenvalue = 1.12, explaining 8.640% of the variance.
Factor loadings below > 0.2:

I was enthusiastic to try Spaced Learning	**.696**	.291
The Spaced learning lessons helped me learn more easily than normal lessons	**.748**	
I think Spaced Learning works well for revision	**.731**	-.369
The spaced Learning lessons ran smoothly	**.585**	
The space activity activities (e.g. juggling) were helpful	**.602**	
Spaced learning would work for all classes, not just revision	**.630**	-.456
The class as a whole enjoyed Spaced learning	**.707**	.366
The teacher was confident at delivering Spaced Learning	.504	**.521**
I felt more motivated to learn during Spaced Learning than in normal classes	**.807**	
I would be happy to try Spaced learning again in the future	**.816**	
I think Spaced Learning would also be useful for other subjects	**.744**	
I found the spaced learning lessons fun	**.678**	-.352
I found the spaced learning lessons helpful for revision	**.775**	.291

Conclusion: nine items used in the second attempt factor analysis will do for an engagement scale.

Rasch analysis on QUB items in R using the eRm package

Note Participant code 1-416 but there are 2 participants for each of p_codes 10, 39, 173, 206, 252.

Check for duplicates in R reveals the results for the same p_codes are duplicated. (data in rows 80, 115, 147,224, 356)

Looking at how each individual item correlates with the total score of the rest of the item set, the negative items (Time, Tiring, Repetitive, NotEnjoy) all have particularly low correlations. This corresponds with the results from the analysis in SPSS.

Looking at distribution across categories, there doesn't appear to be any items with too low counts or irregular distributions e.g. disordered categories.

Full item set (17 items)

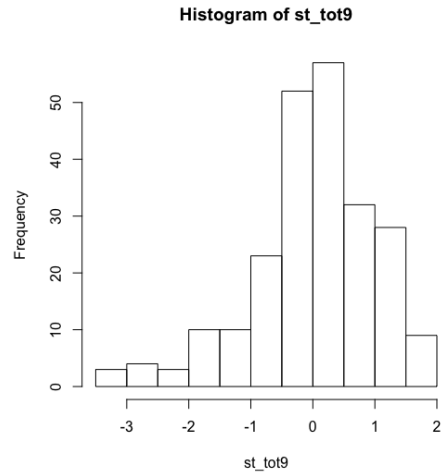
Analysis fails to run on the full item set. Omitting either Smoothly or Teacher allows the analysis to run. But including both these items in the analysis e.g. whilst omitting other items causes problems (NaNs produced). It is not clear why this should be the case, although we do know from the SPSS analysis that Teacher appears to load highest on a second factor, separate to the other items. Also Smoothly we consider to be an oddly phrased item. It may be that eRm is just a flaky package.

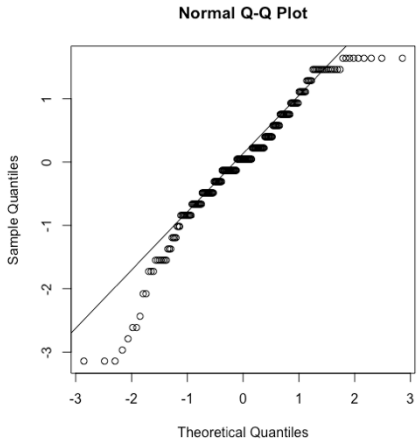
Note also Rasch analysis won't run on the 13 item set, with negative items removed because this set still includes both Smoothly and Teacher.

Nine item set

3 students answered 0 for all items; 9 answered 3 for all items.

Histogram of total score and standardised total score is somewhat negatively skewed.





Note in RSM, one item is set to 0. So this can be considered as a baseline difficulty. Also category parameters w_0 and w_1 are set to 0.

Note "When estimates for the person parameters are of interest some care has to be taken if the cml method is used since person parameters cancel from the estimation equations" p11 (Mair, Hatzinger, Maier, on eRm).

Eta, n - 'basic' parameter estimates, item 'easiness'; just do -eta for 'difficulty'

-Beta – 'ordinary' item parameter estimates, item difficulty, B=Wn, note $B_1 = 0$

Theta – person parameter