A Winning Start
Statistical Analysis Plan

Evaluator (institution): NFER
Principal investigator: Ben Styles

Education Endowment Foundation

PROJECT TITLE	A Winning Start trial
DEVELOPER (INSTITUTION)	Education Endowment Foundation (EEF) as part of the Teacher Choices trials
EVALUATOR (INSTITUTION)	National Foundation for Educational Research (NFER)
PRINCIPAL INVESTIGATOR(S)	Dr Ben Styles
SAP AUTHOR(S)	Dr Joana Andrade, Dr Ben Styles, Palak Roy
TRIAL DESIGN	Two-armed randomised controlled trial (RCT) with a crossover design
TRIAL TYPE	Randomised Control Trial[1]
PUPIL AGE RANGE AND KEY STAGE	Ages 12-13 (Year 8, Key Stage 3)
NUMBER OF SCHOOLS	33
NUMBER OF TEACHER-CLASS UNITS	99
NUMBER OF PUPILS	2260
PRIMARY OUTCOME MEASURE AND SOURCE	Pupils' learning in science as measured by the teachers' end-of-topic tests
SECONDARY OUTCOME MEASURE AND SOURCE	n/a

SAP version history

VERSION	DATE	REASON FOR REVISION
1.0		N/A

SAP Changes from protocol

1. **In the context of NFER's suggested meta-analysis, the analysis of individual topic tests will be completed using OLS with Huber-Wight standard errors, not multilevel models.**

[1] The Randomised Control Trial is being run as part of a feasibility study (see Introduction and Design Overview sections).

The trial design was originally conceived with a small number of topic tests across many schools as a result of recruiting multi-academy trusts that use the same tests. This might have allowed a small number of multilevel models to be run with reasonably precise teacher-level variance estimation. However, in reality, there was very little commonality in test use between schools so running one multilevel model per test would result in poor teacher-level variance estimation. Instead, OLS regression will be used with Huber-Wight standard errors to correct for the teacher-level randomisation.

2. **ICCs will not be calculated in the context of the trial analysis.**
It was stated in the protocol that teacher-level ICCs would be calculated, not only for the purpose of estimating MDES but also because one of the aims of the trial was to present a set of sample size parameters for future trials. However, we will no longer be deriving ICCs in the context of the two proposed analyses. The models to be included in the meta-analysis are no longer multilevel but rather OLS with robust standard errors, from which it is not possible to derive ICCs. Measures other than the ICC are included in the sample size calculations of crossover design trials, with these measures being also dependent on the specificities of the design of each trial. It would not make sense to report ICCs in the specific context of the A Winning Start crossover trial. Instead, we will report teacher-level ICCs across the whole sample from the crossover model, acknowledging that they will not have the same meaning as a teacher-level ICC in a conventional two-level model.

Table of contents

Introduction

This Teacher Choices feasibility RCT aims to answer the research questions centred on the feasibility of the research method as well as the implementation, process and impact evaluation of the teacher choice. Following are research questions taken from the protocol that are relevant to this SAP:

- Is the level of teacher guidance on the 'choices' appropriate to interpret and implement the selected approach and to establish compliance with the intended 'choices'?
- Can the teacher-made tests (such as end of topic tests) be used to measure the impact of teacher choices RCTs? If so, what psychometric properties do these tests need in order to be used as a valid outcome measure for future trials?
- Is it feasible to combine or link attainment data from different teacher-made tests across a number of topics?

- What can we learn from the trials about the expected effect sizes, intra-cluster correlations, pre-post correlations and sample sizes for future teacher choices trials?
- Does a given teacher choice improve pupil attainment outcomes as measured by teacher-made tests?

Design overview

Trial design, including number of arms		Two-armed randomised controlled trial with a crossover design
Unit of randomisation		Teacher[2]
Stratification variables		School and set level
Primary outcome	variable	Pupils' learning in science as measured by the teachers' end-of-topic tests
	measure (instrument, scale, source)	Calibrated/Equated end of topic tests
Secondary outcome(s)	variable(s)	n/a
	measure(s) (instrument, scale, source)	n/a
Baseline for primary outcome	**variable**	Total test score[3]
	measure (instrument, scale, source)	Science attainment at the end of Year 7 (end of Year 7 school science test)
Baseline for secondary outcome	**variable**	n/a
	measure (instrument, scale, source)	n/a

Teacher-class unit level randomisation: For a given topic within a school[4], science teachers were randomly allocated to quiz or discussion starters in the first Spring half term. They then switched to the other condition after half term. In schools where teachers were teaching both topics to more than one Year 8 science class during the Spring term, the classes were aggregated into teacher-class units (TCU) and therefore the randomisation happened at teacher level. The randomisation was stratified by school.

Set/streamed classes were not only identified as such but also ranked top to bottom within each school. Mixed ability classes were given a rank of zero. When more than one class was combined into a TCU, the following rules were applied:

[2] If a teacher is teaching more than one class, these are combined in the same unit.
[3] A baseline is only relevant to the meta-analysis rather than cross-over design analytical solution.
[4] Of the two topics taught at the school during the two Spring terms, we have chosen the topic whose name is first on alphabetical order to be the one that is going to define the random allocation.

- TCUs that were the result of merging at least one mixed ability class were considered to be mixed ability TCUs and given a rank of zero.
- When set/streamed classes with contiguous or equal ranks (e.g. 2,2 or 3,4) were combined into a TCU, the TCU was considered to be set/streamed and given as rank the average rank of the original classes (e.g 2,2 would have a rank of 2; 3,4 would have a rank of 3.5).
- When set/streamed classes with non-contiguous ranks (e.g. 1,3,4 or 2,5) were combined into a TCU, the TCU was considered to be mixed ability and given a rank of zero.

Out of the 99 randomised TCUs in the trial, 54 were set/streamed by ability and 45 mixed ability. To prevent an unbalanced quiz/discussion allocation that might confound the effect of the lesson starters by pupil ability, TCUs were ordered within each school by rank when the stratified randomisation was performed[5]. This procedure guaranteed that near-ability TCUs were paired within each school, as specified in the protocol.

Sample size calculations overview

A conservative approach to sample size calculation was adopted that did not assume a cross-over design due to the uncertainty around calibrating different end-of-topic tests. Instead, it assumed a standard parallel-groups cluster design with the following parameters:

		Protocol		Randomisation	
		OVERALL	FSM	OVERALL	FSM
Minimum Detectable Effect Size (MDES)		0.20	0.27	0.19	0.27
Pre-test/ post-test correlations	level 1 (pupil)	0.70	0.70	0.70	0.70
	level 2 (teacher)				
	level 3 (school)				
Intracluster correlations (ICCs)	level 2 (teacher)	0.17	0.17	0.17	0.17
	level 3 (school)				
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		30	4.23*	22.83	3.22*
Number of TCUs	quiz	80	80	99	99
	discussion	80	80	99	99
	total	80	80**	99	99**
Number of pupils	quiz	1187	170		
	discussion	1187	170		
	total	2374	340	2260	319

5 Please see R code for performing the stratified randomisation in Appendix A.

**14.1% of secondary school pupils were eligible for FSM in January 2019[6]. We collected FSM data directly from schools for this trial so will not be able to use everFSM.*

***Note that we may lose some clusters for the FSM analysis due to the absence of FSM-eligible children in higher science sets.*

Year 8 was chosen over Year 7 as the school routine for Year 8 students is more established and science teachers are likely to pilot new techniques in Year 8 classes before the students move closer to Key Stage 4. Findings from Demack (2019) suggest that 63% of Year 8 science lessons observed by Ofsted are grouped by perceived student ability. Therefore, it is highly likely that at least some trial schools will operate ability grouping in science classes and the class-level ICCs will be as high as 0.4. To mitigate this high class-level ICC, we paired near-ability sets within a school and randomise them to different trial arms (see previous section).

The sample size calculations were run with different scenarios of reduced class-level ICC achieved by pairing the near-ability classes and with the following assumptions: pre-post correlation of 0.7, expected effect size of 0.2 and class size of 30 students. If we achieved a class-level ICC of 0.2, we needed 90 classes in total; with a smaller class-level ICC of 0.17, we needed 80 classes in total and with a class-level ICC of 0.15, we needed 72 classes. Based on these calculations and discussion during the set-up of the trial, it was decided that NFER will recruit 80 classes to take part in this trial and in the event we recruited 99 TCUs. As one of the primary aims of the trial is to present a set of sample size parameters, this design will enable us to investigate whether pairing near-ability science classes helps to reduce the class-level ICCs and/or the assumed ICCs are overly optimistic. This, however, is complicated and potentially mitigated by the cross-over design element.

Analysis

Reliability analysis

As we are using tests chosen, and sometimes written by, teachers it will be important to establish the validity and reliability of the tests. Content validity is being covered in the wider project through expert review of the instruments themselves.Furthermore, relative bias will be explored by including FSM and EAL (and their interaction with topic test score) as covariates in models regressing PTS on each topic test score. Our reliability analysis will constitute four parts:

- Histograms of total test score for each instrument used in the trial. We anticipate around 66 of these as there are 33 schools, each with two topic tests. There may be fewer if there is instrument overlap i.e. where more than one school is using exactly the same instrument.
- Evaluation of local item independence and multidimensionality on a sample of instruments randomly selected for each of the three contrasting topics.
- Cronbach's alpha for each instrument.
- Standard error of measurement from the IRT analysis (see below).

[6] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/812539/Schools_Pupils_and_their_Characteristics_2019_Main_Text.pdf

Any parameters estimated from individual topic test scores are likely to be of low precision, particularly in smaller secondary schools where Year 8 cohort sizes might drop below 100.

Primary outcome analysis

In accordance with the EEF's 2018 statistical analysis guidance (EEF, 2018) the primary outcome analyses in this section will assume intention-to-treat.

Calibration method - linking topic tests within schools

Each student will sit two end-of-topic tests (one after delivery of each of the lesson starters) and the GL Progress Test in Science (PTS). Item-level data is being collected for all three tests. The PTS contains independent single-mark items, whereas topic tests are likely also to contain polytomous items. Our data collection instrument, which science teachers will use to input their item-level scores for each student, therefore includes maximum scores for each item. With a full complement of item-level data and maximum item scores for a given school, we will run a two-parameter IRT model containing items from all three tests. We believe a two-parameter model is more useful for this linking procedure than a Rasch model as we cannot vouch for the psychometric properties of any of the items - either in the end-of-topic or GL tests. The two-parameter model is more likely to fit these diverse items and provide a more meaningful link between the tests. Items that show negative discrimination or that do not fit the model will be removed. Model fit will be determined by a chi-squared test between observed and predicted ability at deciles along the item characteristic curve. This procedure will yield an IRT ability scale for each school i.e. each student will have an ability assigned to them. This ability scale will be used for subsequent analysis of the RCT as it allows the design to be treated as a cross-over with repeated measures within students. Note that as we are potentially linking between three separate domains (knowledge of Topic A, knowledge of the Year 8 science curriculum (PTS) and knowledge of Topic B) this linking process may introduce considerable noise to the outcome measure. How successful the process is depends on the extent to which there is a latent 'scientific ability' that drives performance in all three tests.

Whilst the collation of item-level data by GL Assessment on the PTS test should not be a problem, we are concerned that busy science teachers may not have the time to input item-level data from their tests. If item-level data remains uncollected after some persuasion, we will get back in touch with teachers and ask for a total score for each end-of-topic test. In schools where we only have total scores, it will not be possible to link the tests using IRT. Instead, we will carry out an equipercentile equate between the total score on each topic test and the total score on PTS within such schools. The method we use will follow Kolen and Brennan (1995) Section 3.4: post-smoothed equipercentile equating using cubic splines. Ideally, the extent of smoothing will be decided according to the example in section 3.4.1 of Kolen and Brennan i.e. using as much smoothing as possible subject to remaining within the confidence limits of the unsmoothed equipercentile procedure. However, this may not be possible if we are having to run this procedure across many schools. In practice, it may be that we set the level of smoothing after running, say, three equates and use this level across all schools where we are operating the procedure. The resulting PTS scores will then be standardised to have the same mean and standard deviation as the IRT ability scale.

It is likely that we have a mix of schools, some of which we have been able to link using IRT and some that we have linked through equipercentile equating. The IRT method will probably be better as it relies upon the derivation of a scale that is closer to the notion of a 'true score'. We will therefore retain the mix of linking methods for subsequent analysis rather than reverting to equipercentile equating throughout if just one school fails to supply item-level data.

Analysis of linked data assuming a crossover design

Assuming that it is possible to derive an overall science attainment score by linking the topic tests by the methods proposed in the previous section, we will run a crossover analysis.

As described in section 2a) of Coe and Weidmann (2019), the science attainment score of pupil *i*, under teacher *j* of school *k* at the end of period *p* can be modelled by:

$$Y_{ijkp} = \beta_0 + \delta_k + \alpha_j + u_p + v_{jp} + \tau Z_{jp} + \epsilon_{ijkp}$$
$$\alpha_j \sim N(0, \sigma_\alpha^2)$$
$$v_{jp} \sim N(0, \sigma_v^2)$$
$$\epsilon_{ijp} \sim N(0, \sigma_e^2)$$

where: δ_k are school fixed effects, α_j teacher effects, u_p period effects, v_{jp} a teacher-period interaction, τ the average treatment effect, Z_{jp} identifies if teacher j was allocated to retrieval practice during period *p*, and ϵ_{ijkp} is the measurement error for pupil *i*, in teacher *j* in period *p*.

This proposed model is based on the hierarchical models with cluster effects treated as random analysed in Turner et al (2007). While Turner and colleagues focus on designs which study different sets of subjects in the different periods, in the 'A Winning Start' trial the same cohort of pupils will be assessed in both periods and it is necessary to account for the repeated-measures aspect of the design. This will be done by including an extra pupil level random effect in the model describing the science attainment of pupil *i*, in teacher *j* at the end of period *p*.

We will hence investigate if a teacher's use of quiz as a lesson starter as opposed to a discussion starter had an effect on pupils' mastering of the contents of a science topic as measured by the linked science assessment scores. For this purpose, a multilevel model with three nested levels (period, pupil, and teacher) will be used to account for both cluster randomisation as well as crossover repeated measures design:

$$Y_{ijp} = \beta_0 + \beta_{1i} + \beta_{2j} + \boldsymbol{\beta_3} school_k + \boldsymbol{\beta_4} ability\_set_j + \boldsymbol{\beta_5} period_p + \boldsymbol{\beta_{6j}} period_p + \boldsymbol{\beta_7} starter_{jp} + \epsilon_{ijp}$$

where Y_{ijp} is the score of pupil *i*, in teacher *j*, at the end of Spring period p, $school_k$ is a dummy variable for the school attended by pupil *i* during period p, $ability\_set_j$ is a dummy variable describing if teacher *j* was classified as mixed ability or set/streamed, $period_p$ is a dummy variable for the period at the end of which pupil *i* was tested, and $starter_{jp}$ is a teacher level dummy variable for the starter of the science classes of teacher *j* during period *p*.

The dummy covariates $school_k$ and $ability\_set_j$, were introduced in the model to account for the school-level and TCU ability-level stratified randomisation. and the teacher-level random intercepts term β_{2j} to account for the teacher-level cluster randomisation. The pupil level random intercepts term β_{1i} is included in the model to account for the repeated-measures aspect of this crossover trial..

The crossover design of the trial is reflected by the introduction in the model of the fixed effect term $\boldsymbol{\beta_5} period_p$ and the teacher-level random slopes term $\boldsymbol{\beta_{6j}} period_p$.

Although EEF's statistical analysis guidance (EEF2018) specifies that prior attainment should be controlled for by including a pre-randomisation baseline term in the regression models used in the analyses, this is not a commonly adopted procedure in the context of crossover designs. Conceptually crossover designs can be seen as comparing the elements of a cohort to themselves at two different periods in time, which excludes the necessity for controlling for prior attainment. As such, a pre-randomisation baseline term will not be included in the analysis model. Note that the meta-analytical solution described below does make use of a baseline.

Subgroup analyses

As per protocol, information on FSM eligibility (FSM) and English as an Additional language (EAL) was collected for the Y8 pupils participating in the trial. If the primary outcome analysis is performed assuming the existence of linked data and a crossover design, a post-hoc subgroup analysis will be performed by running the primary outcome model on the subsample corresponding to the groups of interest (FSM-eligible pupils and EAL pupils). This conforms with the 2018 statistical analysis guidelines (EEF 2018).

Missing data

All the information related to the variables in the crossover design three-level model other than outcome (participating pupils, TCU assignation, school, assigned TCU ability, randomised class starter at Spring 1 and Spring 2) has been collected or assigned prior to the beginning of the Spring period. As such, we do not expect missingness in terms of any variables other than linked end of topic assessment scores. In an AB/BA crossover design every subject is, in turn, in the discussion or quiz group depending on period. Bias due to missingness in the outcome variable is only introduced if the information is absent in one of the periods. To prevent the introduction of bias associated with this mechanism, we will only include in the analysis the cases where the pupil has a linked attainment score in both periods.

There is the possibility of class-level or even TCU-level absence from end of topic testing at any given period of the trial. However, it is extremely unlikely that these absences are related to any factors associated with starter allocation, and so, removing the pupils in these classes or TCUs from the analysis is very unlikely to introduce bias.

Given that the number of TCUs recruited for the trial exceeds the planned sample size, plus we anticipate greater power from the crossover element, we do not expect that the decision of excluding from the analysis pupils with no linked attainment score at one of the periods will compromise the power of the analysis.

We will be reporting levels of missingness at the end of both periods as well as an overall level of missingness, in a CONSORT flow diagram, but we will not investigate patterns of missingness.

Compliance

A teacher/TCU-level compliance measure will be considered in the trial.
For each class *k* participating in the trial on each period, we will record the number of lessons where the teacher has used the allocated lesson starter for no more than ten

minutes ($Compliant_k$) and also the total number of lessons taught ($Total_k$). This record will be obtained from teacher logs, which are collected at the end of each period. And from those records we will derive the following teacher/TCU level measures:

$$Compliant_{TCU} = \sum_{k\ in\ TCU} Compliant_k$$
$$Total_{TCU} = \sum_{k\ in\ TCU} Total_k$$

From those we will derive an overall measure of teacher compliance, corresponding to the percentage of lessons taught by a teacher according to the trial specifications:

$$Compliance_{TCU} = \frac{Compliant_{TCU}}{Total_{TCU}}$$

We will calculate the average level of compliance amongst teachers as well as analyse the distribution of the compliance measure in terms of histograms and boxplots.

We will also categorise measures of compliance by which period they occur in and whether they concern using quiz or discussion. We will test if the average level of teacher/TCU compliance differs from Spring period 1 to Spring period 2 and from quiz to discussion by means of paired sample t-tests. Although we can consider teachers as being nested into schools, our analysis will be performed assuming that school is merely a stratifier of the randomisation and not a level of the model, which justifies our choice of paired sample t-tests over multi-level models to enquire on compliance differences over period or lesson starter.

Intra-cluster correlations (ICCs)

Please see SAP changes from protocol section above.

Effect size calculation

If the analysis is performed assuming the existence of linked data and a crossover design we will explore the possibility of computing g_{IG}, a Hedge's g type effect size defined in Madeyski and Kitchenham (2018). According to Madeyski and Kitchenham the standardized g_{IG} effect is comparable to the ones derived for independent group designs. The calculation of g_{IG} is implemented in the R package *reproducer*.

The possibility of calculating g_{IG} is conditional on applicability conditions that we will only be able to verify once we have full access to trial data.

Linear regression models followed by meta-analysis

A multilevel method was originally conceived when we anticipated a sampling design that involved multi-academy trusts, each using the same test. As such a design did not come to fruition, the analysis solution is unlikely to be effective as to carry out a multilevel model on each topic test will likely result in 66 models, in the scenario where there was no test overlap between schools. Whilst possible to automate the running of this many models, in schools

with small numbers of TCUs (in some cases as few as two), multilevel models will not be able to estimate teacher-level variance.

Instead, we propose OLS regression with Huber-White standard errors for each topic test. This approach will avoid the assumptions of a multilevel model. The resulting effect sizes and variances will then be meta-analysed.

Items that were rejected after IRT modelling (see 'calibration method' above) will also be excluded from test scores for this analysis. This should ensure tests of reasonably reliability are being used. We will also be working under the assumption that effects deriving from the topics being taught in the first or second Spring period (period effects) as well as carryover effects[7] at Spring 2 can be regarded as negligible, and, as such, the models are comparable.

In addition to the group variable describing starter allocation, each topic test model will include the score of the end of Year 7 school science test as the baseline covariate, as well as a dummy variable to control for TCU set ability[8].
The OLS regression will be given by the general formula:

$$\mathrm{Y} = \beta_0 + \beta_1 starter + \beta_2 baseline\ Y7 + \beta_3 set\_ability + \beta_4 TCU + \varepsilon$$

Where *Y* is the end of topic test score, β_0 is the intercept, $\beta_4 TCU$ is the teacher fixed effect (β_1 estimated with Huber-White standard errors) and $starter$ and $set\_ability$ are the allocated class starter and TCU ability, respectively. Note that the term $\beta_3 set\_ability$ will only be present in a model if a school has both mixed ability and set/streamed Y8 science classes.
Since we are considering independent subgroups of pupils, the summary effect from the models can be calculated by means of a fixed-effect meta-analysis on individual models. The meta-analysis will follow the methodology prescribed in Chapter 23 of Borenstein et al (2009):

The combined effect will be calculated by creating a weighted mean effect size:

$$M = \sum_{i=1}^{k} W_i Z_i \Big/ \sum_{i=1}^{k} W_i$$

with variance:

$$V_M = 1 \Big/ \sum_{i=1}^{k} W_i$$

where Z_i is the effect size from the ith topic test, $W_i = \frac{1}{V_{Zi}}$ and V_{Zi} in the variance from the i[th] model. The effect size Z_i from each topic test OLS model will be based on the following formula:

$$Z_i = \frac{\beta_1}{\sqrt{(\widehat{var}(Y_i)}}$$

where $\sqrt{(\widehat{var}(Y_i)}$ is from a model without covariates (i.e. unadjusted). The variance V_{Zi} will be the robust standard error of β_1 squared then divided by $\sqrt{(\widehat{var}(Y_i)}$.

[7] Carryover effects correspond to effects that "carry over" from one experimental condition to another, in this case from Spring1 to Spring2.
[8] The TCU ability variables are categorical taking the values "mixed" and "set/streamed".

References:

Borenstein, M., Hedges, L.V., Higgins J.P.T., and Rothstein, H.R. (2009) Introduction to Meta-Analysis, John Wiley & Sons

Coe, R. and Weidmann, B. (2019) Design and Analysis Options, EEF (September 2019)

Demack, S. (2019) Does the Classroom Level Matter in the Design of Educational Trials? A Theoretical and Empirical Review [online]. Available: https://educationendowmentfoundation.org.uk/public/files/Publications/Does_the_classroom_level_matter.pdf [10January, 2020]

EEF (2018) Statistical analysis guidance for EEF evaluations (March 2018)

Kolen, J.M. and Brennan, R.L. (1995) Test Equating, Scaling and Linking; Methods and Practices, Second Edition, Springer

Madeyski, L. and Kitchenham, B. (2018) Effect sizes and their variance for AB/BA crossover design studies, Empirical Software Engineering (2018) 23, 1982-2017

Turner, R.M., White, I.R., and Croudace, T. (2007). Analysis of cluster randomized cross-over trial data: a comparison of methods. Statistics in Medicine, 26(2), 274-289

Appendix A:

```
### The code below performs the stratified randomisation for the crossover trial

rm(list = ls())

setwd("…")

### Load Data a data frame containing the following class level variables:
##  school_id: unique school id
##  class_id: unique class id
##  teacher_id: unique teacher id
##  Spring1.topic and Spring.2 topic: the topics taught in Spring1 and Spring 2
##  grouping: the ability set of the class (mixed, streamed or set)
##  option: NA if grouping=mixed ability, rank (top to bottom within each school) if
##              grouping is either set/streamed

Data<-read.csv(…, header=T, stringsAsFactors = F)
```

```
## Generate class-ranks
Data$option[Data$grouping==mixed]<-0
Data$rank<-parse_number(Data$Option)

### Start a randomisation table
RandTable<-as.data.frame(table(Data$teacher_id,Data$school_id))
RandTable<-RandTable[RandTable$Freq>0,]
colnames(RandTable)[1:2]<-c("teacher_id","school_id")

# Map TCU to teacher_id

### Classes with the same teacher are to be aggregated into the same TCU. In cases of settings and streams that differ by more than one rank unit we consider them to be mixed ability TCUs and give them a rank of 0

# If a teacher is teaching classes that spread over more than 2 contiguous class sets/streams we consider the merged TCU as mixed ability and give it a rank of 0

### If a teacher is teaching set/streamed classes that of the same or contiguous we rank the TCU at their average

###The turank function assigns ranks to TCUs (See Appendix A.1)


### Assign ranks to TCUs (variable TURank)

RandTable$TURank<-
sapply(RandTable$teacher_id,turank,DataFrame=Data,teacherVar="teache_rid",
 rankVar="rank")


###I can now randomise based on School and TURank

firstSeed<-…

dataFrame<-RandTable

randVar<-teacher_id"

stratsList<-c("school_id","TURank")

randGroups<-c("Quiz","Discussion")


###The EETCStratifiedRandomisation function performs a randomisation stratified by the first variable on stratsList while pairing by the second variable on stratsList within the randomisation clusters. (See Appendix A.2)

RandTable<-
EETCStratifiedRandomisation(firstSeed,dataFrame,randVar,stratsList,randGroups)
```

Appendix A.1

```
turank<-function(tu,DataFrame,teacherVar,rankVar){

  aux<-which(colnames(DataFrame)==teacherVar)
  aux<-DataFrame[,aux]==tu
  aux1<-which(colnames(DataFrame)==rankVar)
  tu<-Data[,aux1][aux]

 if (length(tu)>1){

    aux<-max(tu)-min(tu)

    if (aux<2){

      tu<-sign(min(tu))*mean(tu)

    }else{

      tu<-0

    }

  }

  return(tu)

}
```

Appendix A.2

```
EETCstratifiedRandomisation<-
function(firstSeed,dataFrame,randVar,stratsList,randGroups=c("intervention","control
"))){


  ###randVar, the variable corresponding to the randomisation units: can be
specified as a name or a column number

if (typeof(randVar)!="character"){

  randVar<-colnames(dataFrame)[randVar]

}
```

```
#Failsafe check: Duplicated randVar cases and lines with no randvar information

#must removed

dataFrame<-dataFrame[!duplicated(dataFrame[randVar]),]

dataFrame<-dataFrame[!is.na(dataFrame[randVar]),]


###The function can accept the lists of stratifiers either as a list or as a vector

###list will be turned into vectors

stratsList<-unlist(stratsList)


###The list of stratifiers can either be given as a list of names of columns or their
location

if (typeof(stratsList)!="character"){

  stratsList<-colnames(dataFrame)[stratsList]

}


### Failsafe check: if the list of stratifiers contains repeated elements these will be
removed

if(sum(duplicated(stratsList))>0){stratsList<-stratsList[!duplicated(stratsList)]}


### counting the number of stratifiers.

nStrats<-length(stratsList)


set.seed(firstSeed)

seeds<-sample(1:9999,size=(nStrats+1))


#Keep record of the original order of the columns

originalColOrder<-colnames(dataFrame)


###Adding a variable that will allow for the recovery of the original order of the data
##frame rows later on
```

```
dataFrame$originalRowOrd<-1:nrow(dataFrame)

### Ordering the data frame by randVar
dataFrame<-dataFrame[order(dataFrame[randVar]),]

### Assigning a random order to the stratification:
rands<-paste("rand",as.character(1:nStrats),sep="_")



aux<-as.data.frame(sort(unique(dataFrame[,stratsList[1]])))
set.seed(seeds[1])
seeds<-seeds[-1]

aux[rands[1]]<-sample(1:nrow(aux))
colnames(aux)[1]<-stratsList[1]

dataFrame<-merge(dataFrame,aux)

aux<-as.data.frame(sort(unique(dataFrame[,stratsList[2]])))

aux[rands[2]]<-1:nrow(aux)
colnames(aux)[1]<-stratsList[2]

dataFrame<-merge(dataFrame,aux)



###Randomise by randVar variable
```

```
set.seed(seeds[1])

seeds<-seeds[-1]

dataFrame["rand_unit"]<-sample(nrow(dataFrame))


###Reorder the rows of Experiment by rands and rand_unit

rands<-c(rands,"rand_unit")

aux<-do.call(order,dataFrame[rands])

dataFrame<-dataFrame[aux,]



### The randomisation groups can either be specified as a list or as a vector, by default is intervention/control

randGroups<-unlist(randGroups)


###Assigning Control or Intervention Group

aux<-length(randGroups)

aux<-rep(1:aux,times=aux+round(nrow(dataFrame)/aux))

dataFrame$zz_group<-aux[1:nrow(dataFrame)]


rands<-c(rands,"zz_group")


aux<-data.frame(rand_group=randGroups)

set.seed(seeds[1])

aux$zz_group<-sample(1:length(randGroups)


dataFrame<-merge(dataFrame,aux)


##Returning the data frame to its original order

dataFrame<-dataFrame[order(dataFrame$originalRowOrd),]
```

```
###Removing the variables that are no longer necessary
rands<-c("originalRowOrd",rands)
rands<-which(colnames(dataFrame)%in%rands)
dataFrame<-dataFrame[,-rands]
originalColOrder<-c(originalColOrder,"rand_group")
dataFrame<-dataFrame[,originalColOrder]
return(dataFrame)
}
```