Evaluation of Children's University, effectiveness trial
Statistical Analysis Plan

Education Endowment Foundation

Evaluator (institution): National Foundation for Educational Research (NFER)
Principal investigator(s): Pippa Lord

Template last updated: August 2019

PROJECT TITLE	Evaluation of Children's University, effectiveness trial
DEVELOPER (INSTITUTION)	Children's University Trust
EVALUATOR (INSTITUTION)	National Foundation for Educational Research
PRINCIPAL INVESTIGATOR(S)	Pippa Lord
PROTOCOL AUTHOR(S)	Chris Morton, Afrah Dirie and Palak Roy
TRIAL DESIGN	Two-arm cluster randomised controlled trial with random allocation at school-level
TRIAL TYPE	Effectiveness
PUPIL AGE RANGE AND KEY STAGE	Age 9/10 (Year 5) (at baseline) and age 10/11 (Year 6) at follow-up, Key stage 2 (KS2)
NUMBER OF SCHOOLS	164
NUMBER OF PUPILS	5,588
PRIMARY OUTCOME MEASURE AND SOURCE	KS2 maths and reading scaled scores as two separate outcomes (accessed via National Pupil Database (NPD)); Romano-Wolf corrected
SECONDARY OUTCOME MEASURE AND SOURCE	NFER survey that includes following established scales[1]: • Engagement scale from Panorama SEL Survey, Panorama Education • Valuing of school scale from Panorama SEL Survey, Panorama Education

[1] These measures are slightly different to what was proposed at protocol stage, as we have since analysed the baseline data for proposed measures and these two reach the desired reliability - see secondary outcomes section more details.

SAP version history

VERSION	DATE	REASON FOR REVISION
1.0 [*original*]		*N/A*

Table of contents

Introduction

Children's University is a charity that works in partnership with schools to develop a love of learning in children aged 5 - 14. The Children's University network is made up of over 60 delivery partners in the UK who work with local communities, local authorities, national partnerships, schools and parents to deliver these opportunities to children. Children's University (CU) aims to improve the aspirations and attainment of pupils by encouraging participation in learning activities beyond the normal school day. Children's University centres support schools to provide a range of learning opportunities, such as after-school clubs, projects and enrichment activities, and visits to destinations such as libraries, sports clubs, historic centres, museums, or anywhere that offers structured learning activities for children.

Pupils use a 'Passport to Learning' to record activities and hours, and these are rewarded by the collection of credits, certificates, and graduations (for further detail on the intervention see the compliance section). This gives pupils the opportunity to develop character, resilience, and life skills within and beyond the school curriculum. The development of these traits and skills is tied in with improving life chances including good health and wellbeing, avoiding behavioural and social difficulties, and employability (Cullinane and Monticute, 2017; Clarke, et al., 2015).

Children's University has been tested through a previous EEF efficacy trial. This trial found positive impacts on Key Stage 2 maths and reading results equivalent to about 2 months additional progress. Small improvements were also seen for a range of non-cognitive outcomes, such as teamwork, social responsibility, and aspirations. These results had moderate security and provided initial evidence that well-supported enrichment activities can improve children's academic and non-cognitive outcomes.

This effectiveness trial is funded by the EEF to test whether Children's University is effective in raising attainment and non-cognitive outcomes at scale over two academic years. Children's University Trust and the local CUs were responsible to recruit schools. 208 primary schools were recruited between October 2020 and May 2021. Once the recruitment was completed, NFER collected pupil data and asked schools to administer baseline pupil surveys and parent expressions of interest. Completion of these baseline activities were condition to randomisation. 165 schools were randomised with an equal allocation into two arms: intervention and control. Of these, one school had returned empty pupil surveys, so they were removed from the trial. This left 164 schools that completed baseline requirements. The randomisation was stratified by 11 CU localities so that local CUs have a fixed number of intervention schools to support. This resulted in 81 intervention schools and 83 control schools.

Pupils volunteer themselves to take part in CU activities rather than the intervention being delivered to every pupil in a cohort. The opportunity to volunteer was therefore built into the trial design and offered to all pupils in the cohort. In order for us to identify such 'eligible volunteers' from all participating schools, we asked children and parents to express their interest in taking part in CU programme prior to randomisation. We sought this volunteering via parent expression of interest forms and pupil surveys at baseline. The group established this way has become the eligible volunteers to take part in CU programme and constitute the cohort for primary analysis. Across the two randomisation groups, we have 7,073 pupils, of which 5,588 volunteered to take part in CU programme and 1,485 did not volunteer.

Design overview

Table 1: overview of the project's design

Trial design, including number of arms		Two arm, cluster randomised effectiveness trial
Unit of randomisation		School
Stratification variables (if applicable)		CU Locality
Primary outcome	**variable**	1. Maths attainment 2. Reading attainment
	measure (instrument, scale, source)	1. KS2_MATSCORE, 0-999, NPD 2022/2023 KS2 attainment data 2. KS2_READSCORE, 0-999, NPD 2022/2023 KS2 attainment data These outcomes are Romano-Wolf corrected.
Secondary outcome(s)	**variable(s)**	1. Engagement 2. Valuing of school
	measure(s) (instrument, scale, source)	NFER survey administered at follow-up that includes the following established scales: 1. Engagement scale from Panorama SEL Survey, 5-25, Panorama Education 2. Valuing of school scale from Panorama SEL Survey, 4-20, Panorama Education
Baseline for primary outcome	**variable**	1. Maths attainment 2. Reading attainment
	measure (instrument, scale, source)	1. Whether working at or above expected standard for KS1_MATH_OUTCOME, categorical, NPD 2018/2019 KS1 attainment data 2. Whether working at or above expected standard for KS1_READ_OUTCOME, categorical, NPD 2018/2019 KS1 attainment data
Baseline for secondary outcome	**variable**	1. Engagement 2. Valuing of school
	measure (instrument, scale, source)	NFER survey administered at baseline that includes the following established scales: 1. Engagement scale from Panorama SEL Survey, 5-25, Panorama Education 2. Valuing of school scale from Panorama SEL Survey, 4-20, Panorama Education

Randomisation

A cluster randomisation was performed, with schools randomised into the intervention and control arms using a 1:1 ratio. Randomisation was stratified by 11 CU Localities, which were: Bexley, Devon & Cornwall, East London, Elevate a), Enrich, Essex & Suffolk, Peterborough, Rotherham, Wakefield, Westminster and Wolverhampton. The number of schools randomised to the control and intervention in each strata is given in Table 2 below. The intention for stratifying by locality was to aid delivery of the intervention, ensuring that each local CU manager supports half the number of schools recruited in their local area.

Randomisation was carried out by an NFER Statistician using R code. Code was stored for reproducibility and transparency (included in this document as Appendix A). The statistician was not blinded to group allocation. Randomisation allocation data was then passed to NFER's Research and Product Operations team and Children's University Trust who liaised with schools.

Table 2: number of schools randomised in each Children's University Locality

CU Locality	Control	Intervention	Total
Bexley	3	4	7
Devon and Cornwall	4	5	9
East London	10	9	19
Elevate	18	18	36
Enrich	12	11	23
Essex and Suffolk	11	12	23
Peterborough	6	7	13
Rotherham	3	2	5
Wakefield	2	1	3
Westminster	6	6	12
Wolverhampton	8	7	15
Total	83	82	165[2]

[2] 165 schools were randomised with an equal allocation into two arms: intervention and control. Of these, one school had returned empty pupil surveys, so they were removed from the trial. This left 164 schools that completed baseline requirements.

Sample size calculations overview

Table 3: figures used to calculate the MDES at the protocol and randomisation stages

		Protocol		Randomisation	
		OVERALL	FSM	OVERALL	FSM[3]
Minimum Detectable Effect Size (MDES)		0.19	0.26	0.17	0.22
Pre-test/ post-test correlations	level 1 (pupil)	0.65	0.65	0.65	0.65
Intracluster correlations (ICCs)	level 2 (school)	0.18	0.18	0.18	0.18
Alpha[4]		0.025	0.025	0.025	0.025
Power		0.8	0.8	0.8	0.8
One-sided or two-sided?		2	2	2	2
Average cluster size (volunteers)		20	3.14	34.07	5.35
Number of schools	intervention	75	75	81[5]	81
	control	75	75	83	83
	total	150	150	164	164
Number of pupils	intervention	1500	235	2763	434
	control	1500	235	2825	444
	total	3000	470	5588	878

For the sample size calculations at the protocol stage, the target number of schools in each arm of the trial was 75. Correlation between KS1 and KS2 was assumed to be between 0.60 and 0.70 and ICC was assumed to be between 0.16 and 0.20 points which were based on our review of EEF funded studies. The cohort size would be the number of volunteers before randomisation. To calculate this, we used the percentage of pupils who volunteered from the efficacy trial which was 29 from a cohort of 42. Since we refined volunteering for this trial, we assumed that fewer pupils would volunteer and at protocol stage, we expected that between 15 and 20 pupils would volunteer. Using these parameters, Table 3 above presents one such scenario with a correlation between KS1 and KS2 scores estimated at 0.65[6] and the ICC was estimated to be 0.18. It was also assumed that approximately 20 pupils per school would

[3] As free school meal (FSM) eligibility data is not available at the randomisation stage the number of FSM eligible pupils randomised is estimated by multiplying the total pupils randomised by the estimated proportion of FSM-eligible pupils in the population (15.7%).
[4] This alpha value was assumed at the protocol and randomisation stages when a Bonferroni correction was planned to account for multiple outcomes. It does not apply to the final analysis, as a Romano-Wolf correction will now be used instead.
[5] 82 schools were randomised to the intervention, but one was withdrawn shortly afterwards, so it is known for the purpose of sample size calculations that they won't be providing data.
[6] This Pearson correlation was estimated at the protocol and randomisation stages, when it was thought the KS1 baseline would be binary. The KS1 baseline now has four levels, so the Pearson correlation cannot be calculated, but it is expected that the proportion of model variance explained should be as high or higher compared to the binary baseline.

volunteer to take part in CU activities. At the study planning and randomisation stages, it was intended that a Bonferroni correction would be used to adjust for having two primary outcomes, which is reflected in an alpha value of 0.025, rather than the usual 0.05. With the given parameters, an MDES of 0.19 was detected with a power of 0.8.

Although the trial is not powered to detect differences for the free school meals eligible pupils (FSM), the focus was for the trial to recruit schools from disadvantaged areas and to support schools to encourage FSM children to volunteer. With the same parameters and with an assumption that 15.7%[7] of the volunteers have ever been eligible for FSM, the MDES increased to 0.26 for this subgroup.

There was minimal dropout up to the point of randomisation and the number of schools recruited was slightly higher than expected. Holding other parameters involved in the sample size calculation constant, this resulted in slightly lower MDES values at the point of randomisation: 0.17 overall and 0.25 for FSM-eligible pupils only.

In this analysis plan, we propose to use the Romano-Wolf correction to account for multiple testing instead of the Bonferroni correction, as the latter does not account for the dependence between the test statistics of the two outcomes and so is overly conservative. The Romano-Wolf correction accounts for this dependence by resampling from the original data, producing corrected p-values using an algorithm described by Romano and Wolf (2016). As a Bonferroni-adjusted alpha value was assumed when designing and recruiting for the study, this value remains in the power calculations. Since the Romano-Wolf correction leads to greater power than the Bonferroni correction, the actual MDES at analysis is expected to be smaller than that given.

The efficacy trial demonstrated an effect size of 0.23 for reading and 0.20 for maths (Higgins et al archive analysis[8]) so this MDES is considered adequate for these outcomes. A range of other potential scenarios, varying the MDES, ICC and correlation between KS1 and KS2 scores, are presented in the protocol for this study.

Outcome measures

Primary outcomes

1. Maths attainment

The maths attainment primary outcome will be measured using scaled maths scores from National curriculum assessments taken at the end of KS2. These range between 0 and 999 and can be obtained via the 'KS2_MATSCORE' variable in the National Pupil Database (NPD). The corresponding baseline measure will be the 'KS1_MATH_OUTCOME' variable from the NPD, which will be treated as a categorical variable with four levels: 'below expected standard' (NPD codes 'BLW', 'PK1', 'PK2, 'PK3' and 'PK4'[9]), 'working towards expected standard' (code 'WTS'), 'at expected standard' (code 'EXS') and 'above expected standard' (code 'GDS'). This will be modelled in regressions with 'below expected standard' as the

[7] The source of this estimate is: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/826252/Schools_Pupils_and_their_Characteristics_2019_Accompanying_Tables.xlsx.
[8] Note that these figures were obtained from an analysis of archived data from the previous evaluation, not from the evaluation report itself.
[9] The reason for merging these codes is that only a small percentage of pupils received each one (https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/851305/KS1_tables.xlsx), which could cause model estimation problems.

reference level and three indicator variables for the remaining levels. Codes indicating no result (codes ‘A’, ‘D’ and ‘Q’) will be treated as missing data.

2. Reading attainment

The reading attainment primary outcome will be measured similarly to the maths outcome above, based on scaled reading scores in the 0-999 range and using the ‘KS2_READSCORE’ variable in the NPD. The baseline measure will use the ‘KS1_READ_OUTCOME’ variable from the NPD, which will be converted to a categorical variable using the same codes as for the maths baseline measure above.

Secondary outcomes

All secondary and exploratory outcome measures described below were collected via a bespoke pupil survey created by NFER. This was completed by pupils at baseline (prior to randomisation) and pupils will answer the same questions at follow-up in July 2023. The baseline measure in each case therefore uses the same scale as the outcome measure. All secondary and exploratory outcomes will be obtained by summing their constituent items. If some items are missing from a measure, then the total score for that measure will be treated as missing (consistent with the complete case analysis used in the secondary analysis models).

It was specified in the study protocol that of four potential secondary outcome measures (Engagement, Valuing of school, Self-esteem and Goals and aspirations), only those demonstrating a Cronbach’s Alpha value of 0.7 or above in the baseline pupil data would be included in the secondary analysis. Engagement and Valuing of School met this criterion (Table 4) and so were included as secondary outcomes. To add further confidence in these measures, a confirmatory analysis was performed for Engagement and Valuing of School in R using the package ‘lavaan’ (Rosseel, 2012). Results from the factor analysis overall supported the use of these measures: the root mean square error was 0.063, the Comparative Fit Index was 0.993 and the Tucker-Lewis Index was 0.990. The other two candidate measures (Self-esteem and Goals and aspirations) did not meet the $\alpha > 0.7$ eligibility criterion for inclusion in the secondary analysis (see Table 5). However, as these measures are an important part of the theory of change for this study, options other than completely removing them were considered. The options included merging their constituent items into a single measure (with a Cronbach’s Alpha of 0.66) and considering them for further exploratory analyses rather than secondary outcomes. To support this decision-making an exploratory factor analysis (varimax rotation) with four factors was performed in R using the ‘psych’ package (Revelle, 2022). One of the factors that emerged from this analysis primarily loaded onto the five items that comprise the Self-esteem and Goals and aspirations measures (loadings between 0.375 and 0.477).

While the factor analyses described above provided some limited support for combining the two measures, it was decided that the resulting measure did not have a clear and meaningful interpretation. Additionally, the combined measure still did not reach the pre-specified Cronbach’s Alpha threshold of 0.7. Thus, we decided that the Self-esteem and Goals and aspirations measures will instead be included in exploratory analyses, with the appropriate caveats about their interpretation made clear in the final report. Details of these and the other two exploratory measures in this trial can be seen in Table 5 below.

Table 4: Properties of secondary outcome measures

Measure (min-max value)	Constituent items	Response categories (scores)	Cronbach's alpha
Engagement (5-25)	In general, how excited are you about going to your lessons?	Not at all excited, Slightly excited, Somewhat excited, Quite excited, Extremely excited (1-5)	0.77
	How focused are you on the activities in your lessons?	Not at all focused, Slightly focused, Somewhat focused, Quite focused, Extremely focused (1-5)	
	In your classes, how excited are you to join in?	Not at all excited, Slightly excited, Somewhat excited, Quite excited, Extremely excited (1-5)	
	When you are not in school, how often do you talk about ideas from your lessons?	Almost never, Once in a while, Sometimes, Frequently, Almost always (1-5)	
	How interested are you in your lessons?	Not at all interested, Slightly interested, Somewhat interested, Quite interested, Extremely interested (1-5)	
Valuing of school (4-20)	How interesting do you find the things you learn in your lessons?	Not at all interesting, Slightly interesting, Somewhat interesting, Quite interesting, Extremely interesting (1-5)	0.70
	How often do you use ideas from school in your everyday life?	Almost never, Once in a while, Sometimes, Frequently, Almost always (1-5)	
	How important is it to you to do well in your lessons?	Not at all, A little bit, Somewhat, Quite a bit, A tremendous amount (1-5)	
	How useful do you think school will be to you in the future?	Not at all useful, Slightly useful, Somewhat useful, Quite useful, Extremely useful (1-5)	

Table 5: Properties of exploratory outcome measures

Measure (min-max value)	Constituent items	Response categories (scores)	Cronbach's alpha
Self-esteem (3-15)	• I can work out my problems • I can do most things if I try • There are many things that I do well	Never, Not very often, Sometimes, Very often, Always (1-5)	0.56
Goals and aspirations (2-10)	• I have goals and plans for the future • I think I will be successful when I grow up	Never, Not very often, Sometimes, Very often, Always (1-5)	0.50
Problem-solving (3-15)	• When I need help, I find someone to talk to • I know where to go for help when I have a problem • I try to work out problems by talking about them	Never, Not very often, Sometimes, Very often, Always (1-5)	0.60
Fear of communication (8-40)	• Talking to someone new worries me.* • I like to talk when the whole class listens. • Talking to teachers makes me feel uncomfortable.* • I like talking in front of a group of people • I like it when I don't have to talk.* • When someone asks me a question, it makes me nervous.* • I like to talk to people I haven't met before. • I look forward to talking in class (e.g. to contribute to discussions). Items marked with a '*' are reverse-scored.	Strongly disagree, Disagree, Not sure, Agree, Strongly agree (1-5)	0.63

The secondary outcome measures of this study are:

1. Engagement

To measure the engagement and valuing of school outcomes, the Panorama Social-Emotional Learning (SEL) survey (Panorama Education, n.d.) was used. A section of the SEL survey focuses on the pupil's learning environment and the degree to which this influences their academic success and social-emotional development, which is broken down into 10 subscales. The engagement subscale consists of five Likert-type items which can be rated between 1 and 5. The Likert scale labels vary from item to item, but higher numbers are positive. Summing these item scores produce an outcome measure in the range 5-25, where higher values indicate greater engagement with school classes.

The fact that the Likert scale labels for Engagement vary between its constituent items (see Table 4) is worth further comment. As the 1-5 scores of the constituent items are summed to obtain the overall measure, there is an assumption that each increase of one point represents the same 'amount' of the underlying variable (Engagement). It could be argued that this is less likely to hold with inconsistent scale labels: for example, the increase from 'not at all excited' to 'slightly excited' may not be the same as the increase from 'almost never' to 'once in a while'. However, the results of the confirmatory factor analysis reported above provide some reassurance as to the validity of the measure, as does the fact that it is an established scale. Similar comments apply to 'Valuing of school' below, which also has inconsistent Likert scale labels.

2. Valuing of school

Another subscale of the Panorama SEL survey is 'Valuing of school': the degree to which pupils feel that school is interesting and important to them. The subscale consists of four Likert-type items, rated between 1 and 5, where higher ratings are positive. The overall score for the subscale ranges between 4 and 20, with higher values indicating a greater value placed on schooling.

Exploratory outcomes

As described above, Self-esteem and Goals and aspirations had lower reliability (<0.7 Cronbach's Alpha) and hence will not be considered for secondary analyses. These outcomes, along with the other two additional outcomes, Problem-solving and Fear of communication, will be considered for exploratory analyses. While writing the protocol, it was decided Problem-solving and Fear of communication would be exploratory rather than secondary outcomes (irrespective of their reliability) because these scales measure a pupil's perception of their Problem-solving abilities or Fear of communication, rather than directly measuring the concepts themselves.

1. Self-esteem

The self-esteem of pupils will be measured via three items, which comprise the self-esteem subscale of the Student Resilience Scale (SRS) (Sun and Stewart, 2007). The SRS is a 47-item measure, consisting of 12 subscales that quantify child resilience and protective factors in their family, school and community. For each item of the self-esteem subscale, pupils indicate the frequency of positive occurrences (e.g., "I can work out my problems") with a rating between 1 ("never") and 5 ("always"). The sum of these ratings forms an outcome with range 3-15, where higher scores indicate higher self-esteem.

2. Goals and aspirations

The ability of pupils to form goals and aspire towards a successful future will be measured via two items, which form the goals and aspirations subscale of the SRS. These items again relate to positive occurrences ("I have goals and plans for future" and "I think I will be successful when I grow up"), which are rated between 1 ("never") and 5 ("always"). The resulting outcome therefore takes a value between 2 and 10, where higher ratings are considered positive.

3. Problem-solving

Pupils' perception of their ability to work through their problems with the support of others will be measured via three items, which form the problem-solving subscale of the SRS. These items relate to positive attitudes (e.g., "I try to work out problems by talking about them"), which are rated between 1 ("never") and 5 ("always"). The resulting outcome therefore takes a value between 3 and 15, where higher ratings indicating that the pupil is more able to utilise the support of others to help solve their problems.

4. Fear of communication

The Personal Fear of Communication scale concerns feelings about communicating with other people (McCroskey, 1981). Eight of the fourteen items from this scale are used to measure Fear of Communication in this trial. Both positively and negatively worded statements are included in these items, which are rated between 1 ("strongly disagree") and 5 ("strongly agree"). This reverses the original instrument (where 1 was "strongly agree" and so on), allowing higher scores to be positive, in line with other outcomes in this trial. Negatively worded items are reverse-scored, so that the overall total of the items will take a value between 8 and 40, with higher values indicating less fear of communication.

Analysis

Both the primary and secondary analysis will be performed within an intention-to-treat framework and will follow the EEF's 2018 statistical analysis guidelines. All analysis will be performed on the sample of volunteers: pupils that volunteered to take part in CU activities pre-randomisation at both control and intervention schools. As this is an intention-to-treat approach, if NPD records show a pupil has left their current school during the trial period, they will still be included in the analysis. All primary and secondary modelling will be 'complete-case' analysis: if a pupil is missing any variables in the model specified, they will not be included in that part of the analysis. The impact of using a complete case analysis will be explored in the missing data section.

Primary outcome analysis

The purpose of the primary analysis is to investigate whether CU has an effect on pupil's maths and reading attainment at KS2. Two models will be run, with KS2 maths and reading scores as the respective outcomes. These will be taken from the NPD and they are described in the outcomes section.

To account for clustering of KS2 scores within schools, a two-level (pupil and school) linear mixed effects model will be used for each outcome. Fixed effects models can also account for this clustering but would not be appropriate here, as they do not allow for inference around school-level variables, in particular the intervention indicator. The model for the maths primary outcome will be:

$$\mathrm{KS2\_math}_{ij} = \beta_0 + u_{0j} + \beta_1 \mathrm{intervention}_j + \boldsymbol{\beta_2 \mathrm{KS1\_math}_{ij}} + \boldsymbol{\beta_3 \mathrm{CU\_locality}_j} + \epsilon_{ij}$$

Where $\mathrm{KS2\_math}_{ij}$ is the KS2 maths score of pupil i in school j, and $\mathbf{KS1\_math}_{ij}$ is a vector of indicator variables for three levels of KS1 maths attainment (plus 'below expected standard' as the reference level, see Outcome measures section). In this equation $\mathrm{intervention}_j$ is the random allocation (intervention or control) indicator for school j, and u_{0j} is the random intercept for school j. $\mathbf{CU\_locality}_j$ is a vector of 10 indicator variables for the 11 strata of CU locality (the reference level for this is the Elevate locality). For the reading primary outcome, the right side of the equation above will be the same, except $\mathrm{KS1\_math}_{ij}$ will be replaced by the KS1 reading baseline measure.

All multi-level models in the primary and secondary analyses will be run using the package 'lme4' (Bates, et al., 2015) in the software R.

Correction for multiple testing

As the number of outcomes in a study increases, the probability of at least one false positive finding (known as the 'family-wise error rate' or FWER) also increases. In this trial there are two primary outcomes, so a Romano-Wolf correction will be applied to p-values from the primary analyses, restricting the FWER to 0.05. The reason for choosing the Romano-Wolf correction over the Bonferroni correction is that the latter does not account for the dependence between the test statistics of the two outcomes and so is overly conservative[10]. The Romano-Wolf correction accounts for this dependence by resampling from the original data, producing corrected p-values using an algorithm described by Romano and Wolf (2016). The resampling itself can be performed in several ways; for this study a permutation-based method designed for cluster-randomised trials will be used, as described by Watson et al. (2021) and implemented using the R package 'crctStepdown' (Watson, 2021).

One difficulty with using the Romano-Wolf correction is that there is no way to apply it to the construction of confidence intervals. That means that the usual 95% confidence intervals will be presented for this trial. Where these confidence intervals 'disagree' with the corrected p-values (that is, the interval does not contain zero, but the p-value is above 0.05) there will be additional discussion in the final report about interpreting the uncertainty surrounding the point estimate in the context of multiple testing.

Secondary outcome analysis

The purpose of the secondary analysis is to investigate the effect of participation in CU activities on two non-cognitive pupil outcomes. For each of these outcomes the analysis will closely mirror that described in the primary analysis section. A linear model with two levels (pupil and school) will be calculated to account for clustering of pupil outcomes within schools.

For example, in the case of the Engagement secondary outcome:

$$\mathrm{Engagement}_{ij} = \beta_0 + u_{0j} + \beta_1 \mathrm{intervention}_j + \beta_2 \mathrm{baseline\_engagement}_{ij} + \boldsymbol{\beta_3} \mathbf{CU\_locality}_j + \epsilon_{ij}$$

Where $\mathrm{engagement}_{ij}$ is the SEL Engagement score for pupil i in school j at follow-up and $\mathrm{baseline\_engagement}_{ij}$ is the same score measured by the baseline survey. As before, $\mathrm{intervention}_j$ is the intervention or control indicator, $\mathbf{CU\_locality}_j$ is a vector of indicators for the CU locality variable and u_{0j} is the random intercept for each school.

[10] Conservative in the sense that the actual FWER will be lower than the nominal rate of 0.05 and a higher power could be achieved while maintaining a FWER of 0.05.

For the other secondary outcome Valuing of school, the right-hand side of the above equation remains the same, except $\mathrm{baseline\_engagement}_{ij}$ is replaced with the baseline measurement of Valuing of school.

Subgroup analyses

We will perform subgroup analyses for each of the two primary outcomes. This section illustrates the model structures using the KS2 maths outcome. The same analyses will be performed for the outcome of KS2 reading scores, except the KS1 maths baseline measure will be replaced by the KS1 reading measure. In each analysis, p-values will be corrected for multiple testing using the Romano-Wolf correction, as described for the primary analysis.

To investigate whether the effect of the intervention is different for pupils from disadvantaged backgrounds, subgroup analysis will be performed on pupils who are eligible for FSM. This will use the 'EVERFSM_6_P[11]' variable collected from the NPD for the 2022/2023 academic year to define FSM eligibility. Two approaches will be used: (i) rerunning both primary outcome models for FSM-eligible pupils only and (ii) adding an interaction between FSM-eligibility and intervention assignment to the primary outcome models. Specifically, the two level (pupil and school) linear model used for (ii) in the case of the maths primary outcome will be:

$$\begin{aligned}\mathrm{KS2\_math}_{ij} = {} & \beta_0 + u_{0j} + \beta_1 \mathrm{intervention}_j + \beta_2 \mathrm{FSM}_{ij} + \beta_3 \mathrm{intervention}_j \times \mathrm{FSM}_{ij} \\ & + \boldsymbol{\beta_4 \mathrm{KS1\_math}_{ij}} + \boldsymbol{\beta_5 \mathrm{CU\_locality}_j} + \epsilon_{ij}\end{aligned}$$

where again $\mathrm{KS2\_math}_{ij}$ is the KS2 maths score of pupil i at school j, and $\mathbf{KS1\_math_{ij}}$ is a vector of indicator variables for three levels of KS1 maths attainment (plus 'below expected standard' as the reference level). $\mathrm{CU\_locality}_j$ is a vector of indicators for the CU locality stratification variable and u_{0j} is the random intercept for school j. In this equation $\mathrm{intervention}_j$ is the intervention or control indicator for school j, FSM_{ij} is an indicator variable for whether the pupil is eligible for FSM and $\mathrm{intervention}_j \times \mathrm{FSM}_{ij}$ is the interaction between the two. This means that the statistical significance of β_3 indicates whether there is a differential effect of the CU programme on maths attainment for FSM-eligible pupils. However, it should be emphasised that this trial is not powered to detect an effect for the FSM-eligible subgroup.

Analysis of exploratory outcomes

There are four exploratory non-cognitive measures in this trial: Self-esteem, Goals and aspirations, Problem-solving and Fear of communication. Each of these will be the outcome in a multilevel linear regression, similar to the secondary analysis models. Modelling will be conducted in exactly the same manner as for the secondary outcomes, with the baseline measure of the respective exploratory outcome being included as a covariate in each regression.

Longitudinal follow-up analyses

No longitudinal follow-up is planned for the study at this time.

[11] This variable indicates whether a pupil is known to have been eligible for FSM for any period in the last 6 years.

Imbalance at baseline

To assess imbalance in baseline characteristics a table will be produced describing characteristics of the control and intervention groups after randomisation. The following variables will be included in the table:

School level

- Proportion of pupils eligible for FSM in 2022/23
- CU locality
- Whether the school is urban or rural
- School type (academy, maintained or independent) in 2022/23
- Most recent overall Ofsted rating in 2022/23
- Proportion of pupils meeting the expected standard in their KS2 maths exam in 2022/23
- Proportion of pupils meeting the expected standard in their KS2 reading exam in 2022/23

Pupil level

- FSM eligibility in 2022/23 (‘EVERFSM_6_P’ from NPD)
- Baseline (KS1) maths outcome
- Baseline (KS1) reading outcome

Categorical variables will be described in terms of counts and proportions, while means and standard deviations will be given for continuous variables. School-level variables will be obtained from NFER’s own database of school characteristics, pupil-level variables will be obtained from the NPD. As KS1 maths and reading ratings will be treated as categorical variables, the difference between the intervention and control groups will be expressed as odds ratios, with 95 percent confidence intervals.

Missing data

The missing data analysis will only be performed for the primary outcomes. The analysis described below is for the KS2 maths outcome, but the equivalent process will occur for KS2 reading outcome. These outcomes will be considered separately, and one may go through more stages of missing data analysis than the other. For example, this could occur if less than five percent of cases were missing for just one of the two outcomes.

The number and proportion of missing cases will be reported for the maths primary analysis model. If this is less than five percent then the potential for bias in a complete case analysis will be considered minimal and no further missing data analysis will take place, following EEF’s statistical guidance. Otherwise, the number and proportion of missing values for each variable included in the primary analysis models will be reported. As the intervention group status and CU locality variable were known at randomisation, only the baseline and follow-up maths scores will be missing.

Assuming both baseline (KS1) and follow-up (KS2) maths results have missing data, a multilevel logistic regression with two levels (pupil and school) will be run for each; the outcomes will be the logit probability of the respective maths scores being missing. All other variables from the maths primary analysis model will be included as covariates, together with these additional variables that may be associated with missingness:

- Pupil FSM eligibility in 2022/23 (‘EVERFSM_6_P’ from NPD)

- Pupil's school type (academy, maintained or independent) in 2022/23
- Proportion of pupils eligible for FSM in 2022/23 at the pupil's school
- Missingness of the pupil's baseline (KS1) reading result
- Missingness of the pupil's KS2 reading outcome

The additional variables which demonstrate an association with missingness in the KS2 maths score, as indicated by a p-value below 0.05, will be included as covariates in the primary analysis model. If rerunning the primary analysis with these extra covariates included alters the substantive interpretation of the intervention effect, then the KS2 maths outcome may be 'missing at random' conditional on the inclusion of those covariates. This would then need to be discussed in the report.

If missingness of the baseline (KS1) maths covariate is associated with any of the additional variables from the above list, a sensitivity analysis will be performed using multiple imputation. Missing data will be imputed using predictive mean matching: the imputation model will include all variables from the primary analysis, together with any additional variables associated with missingness of baseline or follow-up maths scores. Ten datasets will be generated, each using ten iterations, and the primary analysis model will then be re-run on each. Estimates from each model will be pooled into a single set of estimates and standard errors that will be compared to the results of the original analysis.

The multiple imputation will be run using the packages 'mice' (van Buuren and Groothuis-Oudshoorn, 2011) and 'smcfcs' (Bartlett, et al., 2022) in R.

The analysis described above relates to missing variables in the primary analysis, but there are also other potential sources of missing data in this trial. The compliance analysis and modelling of participation in different CU activities (see sections below) relies on accurate and complete data being recorded on the CU online digital platform. If pupil participation in an activity is not recorded by their school or activity provider this could introduce bias into estimates from any resulting analysis. As it won't be possible to confirm whether CU online data is missing for a pupil, it won't be possible to assess the impact of missing data on analysis estimates accurately.

Association between secondary outcomes and attainment

The logic model for this trial theorises that participation in CU activities increases the short-term outcome of positive identification with the school (Valuing of school measure) and the intermediate outcome of motivation to learn (Engagement measure), amongst other factors. These in turn improve KS2 maths and reading scores. The possibility of investigating these causal relationships using mediation analysis was considered. However, as KS2 exams are sat before the follow-up measurement of the secondary outcomes (the 'effect' precedes the 'cause'), it was decided that the causal interpretation implied by a mediation analysis was not sufficiently credible to proceed. Instead, a simple analysis will be conducted in which the association of Engagement and Valuing of school with KS2 maths scores will be investigated. Linear mixed effects models will be calculated for the KS2 maths outcome, using intervention pupils only:

$$\mathrm{KS2\_math}_{ij} = \beta_0 + u_{0j} + \beta_2 \mathrm{Engagement}_{ij} + \boldsymbol{\beta_4 \mathrm{KS1\_math}_{ij}} + \boldsymbol{\beta_5 \mathrm{CU\_locality}_{j}} + \epsilon_{ij}$$

$$\mathrm{KS2\_math}_{ij} = \beta_0 + u_{0j} + \beta_2 \mathrm{Value\_school}_{ij} + \boldsymbol{\beta_4 \mathrm{KS1\_math}_{ij}} + \boldsymbol{\beta_5 \mathrm{CU\_locality}_{j}} + \epsilon_{ij}$$

In these equations terms are defined as for previous models. $\mathrm{Engagement}_{ij}$ and $\mathrm{Value\_school}_{ij}$ are the follow-up Engagement and Valuing of school measures for pupil i at school j. If β_2 is larger than zero, this indicates that pupils with high Engagement (or Valuing of school) close to the time of their KS2 exams had higher average attainment. One causal explanation for this (amongst others, recalling the exams precede follow-up Engagement measurement) is that higher Engagement causes higher KS2 attainment, corroborating one aspect of the logic model.

Both models will be repeated using KS2 reading score as the outcome and including the baseline KS1 reading measure instead of KS1 maths as a covariate, so that there will be four models in total. Romano-Wolf adjusted p-values will be calculated for each pair of maths and reading outcomes; for example, a correction for two hypotheses will be applied to the two models with Engagement as a predictor and the outcomes of KS2 maths and reading scores respectively.

Compliance

Compliance data will be collected from the digital platform- CU online[12]. Pupils receive awards at set thresholds to celebrate their participation in extra-curricular activities. The first of these is the bronze award, which is gained at 30 hours of CU activities.

A challenge when investigating the causal effect of variables derived from the CU Online data on KS2 attainment (as in this compliance analysis) is that KS2 exams will be sat in May, before the final CU Online data extract is received. This means the final extract will overestimate pupil's participation in CU activities before KS2 attainment is measured. It is not possible to simply filter out activities with post-exam dates, as activity hours are often uploaded onto the platform in one go (e.g., all hours for a pupil uploaded at the end of a term), lagging behind the actual dates the activities were completed. As a solution, the evaluators will be working with Children's University to obtain an additional extract of CU Online data shortly before the May KS2 exams. Schools will be encouraged to upload all pupil activity data in preparation for this additional extract, to make it as accurate a reflection of pupil CU activity participation up to that point as possible. However, it is still possible that some activity data will not be uploaded in time for the May CU Online extract. This possibility will be investigated using descriptive statistics and graphical methods. For example, the average amount of CU activity hours uploaded each day between the May extract and final extract will be compared to the period leading up to the May extract and to average across the entire study period.

There will be two compliance measures analysed in this trial:

1. Whether a pupil participated in at least 30 hours of CU activities. This is a dichotomous measure of compliance.
2. The total number of hours of participation in CU activities. This is a continuous measure of compliance.

Pupils at intervention schools will have complied with the programme if their CU online data shows that they took part in at least 30 hours of CU activities. The main compliance measure for this study is therefore binary and defined at pupil level. However, while 30 is the minimum required to receive an award in the CU programme, the benefits of the programme for pupils might be expected to gradually accrue up to and beyond the 30 hours threshold. To investigate this 'dosage' relationship, an additional compliance measure will be created that is equal to the number of hours recorded in the CU online data, or zero for pupils in control schools. Compliance analysis is described for the maths primary outcome below, but the equivalent models will also be run for the reading primary outcome. This means there will be four models in total, investigating the effect of the two compliance measures on the maths and reading outcomes.

An instrumental variable (IV) analysis will be performed, using two-stage least squares methods (Angrist and Imbens, 1995) to estimate the effect of binary compliance with the intervention on maths attainment. For the first stage the compliance indicator will be regressed on treatment assignment, together with covariates from the maths primary analysis model (the baseline KS1 maths indicator and CU locality). For the second stage KS2 maths scores are regressed on each pupil's predicted compliance value from the first stage, together with covariates from the maths primary analysis model. The coefficient for predicted compliance in this second stage is the CACE (complier average causal effect) estimate for the effect of compliance (30 hours of CU activities) on KS2 maths scores. Results from both stages will be reported.

The IV two-stage least squares methods described above will be repeated for compliance measure (2), which will be treated as a continuous variable. The modelling will be the same as that described above, except that the outcome at the first stage is continuous rather than binary.

All IV analysis will be performed using the 'ivreg' (Fox, et al., 2021) and 'ivpack' (Jiang and Small, 2014) packages in the R software.

Graduation ceremony attendance

As described in the protocol, pupils are presented awards, certificates, diplomas, or degrees depending on the number of CU participation hours they have amassed. Pupils that received a bronze award or greater are invited to attend a graduation ceremony celebrating their achievements. Participation at graduation ceremony is an important output in the logic model as it provides a sense recognition for pupils' achievements where families and schools celebrate their commitment to learn beyond the classroom. At the protocol writing stage it was planned that the effect of graduation ceremony attendance on KS2 maths and English scores would be investigated in a separate piece of analysis. However, while writing this analysis plan it was decided that this analysis would not be possible due to limitations of the data available. This is firstly because some graduation ceremonies occur after KS2 exams are sat, preventing the causal inference. Graduation ceremony attendance data is not recorded for individual pupils. While ceremony attendance may contribute to the effect of receiving a bronze award (30+ hours of CU activities, the main compliance analysis measure) on KS2 scores, it will not be possible to separate out and measure this contribution.

Participation in different types of CU activities

As an additional piece of analysis, we will investigate whether the amount of participation in certain types of CU activities is associated with each primary outcome. This analysis will use CU participation data collected via the CU Online platform. When an activity is registered for CU validation by its provider (the pupil's school or an outside organisation) they are asked to classify the activity using up to four of the following categories:

- Arts, culture and music
- Careers and enterprise
- Citizenship
- Family learning
- History and heritage

- Languages
- Literacy
- Mental health and wellbeing
- Nature and the environment
- Online
- Outdoor learning
- Practical life skills
- Science, technology, engineering and maths
- Social and community action
- Sports and physical
- Uniformed groups

For example, performative dance might be assigned the category ‘Arts, culture and music’, as well as ‘Sports and physical’. In order to understand the relationship between participating in different categories of activities and the effect on outcomes, a two-level (pupil and school) linear regression model will be used to explore the association between amount of time pupils spent in undertaking specific types of activities and KS2 maths score, specified as follows:

$$\text{KS2\_math}_{ij} = \beta_0 + u_{0j} + \boldsymbol{\beta_2}\mathbf{KS1\_math_{ij}} + \boldsymbol{\beta_3}\mathbf{CU\_locality_j} + \boldsymbol{\beta_4}\mathbf{activity\_hours_{ij}} + \epsilon_{ij}$$

The entries of vector $\mathbf{activity\_hours_{ij}}$ are the number of hours spent on each type of CU activity listed above; it is a vector of length 16, with zero entries where a pupil has not undertaken activities of that category. Following on from the performative dance example above, if a pupil’s record shown they had spent 9 hours on this activity, but no time on any others, then $\mathbf{activity\_hours_{ij}}$ would have entries of ‘9’ for each of ‘Arts, culture and music’ and ‘Sports and physical’, with zeros elsewhere. Other covariates are defined as described in the primary analysis section. This analysis will include volunteer pupils at intervention schools only.

By default, the intention is to include the number of hours spent on every activity category given above as covariates in the regressions. However, in practice there may be very small numbers of pupils taking up certain types of activities and/or high levels of correlation between the hours spent on different activities, which could lead to model estimation problems. Upon viewing the CU online data at follow-up, the research team will make a final decision on how the activity types will be parametrised in the analysis.

To investigate the impact of the range of activities on outcomes, further analysis will be performed for intervention volunteers that took part in at least one CU activity. The following two-level (school and pupil) linear regression will be used:

$$\begin{aligned}\text{KS2\_math}_{ij} = {} & \beta_0 + u_{0j} + \boldsymbol{\beta_2}\mathbf{KS1\_math_{ij}} + \boldsymbol{\beta_3}\mathbf{CU\_locality_j} + \beta_4\text{many\_activities}_{ij} \\ & + \beta_5\text{diverse\_activity\_categories}_{ij} \\ & + \beta_6\text{many\_activities}_{ij} \times \text{diverse\_activity\_categories}_{ij} + \epsilon_{ij}\end{aligned}$$

Here $\text{many\_activities}_{ij}$ is an indicator variable, taking the value 1 if the pupil took part in more than the median number of CU activities observed in the data[13] and 0 otherwise. Similarly, $\text{diverse\_activity\_categories}_{ij}$ takes the value 1 if the pupil took part in more than the median number of activity categories observed in the data and 0 otherwise. The interaction between

[13] The calculation of median number of activities will not include values of zero, as this analysis is only for pupils that took part in one or more activities.

these represents a possible additional effect associated with taking part in many activities of diverse types.

It is also important to note that because participation in certain types of CU activities is not randomised[14], it will not be possible to draw causal conclusions from both these analyses. If, for example, pupils that spend more hours on maths-related CU activities have better KS2 maths scores, this could be due to the nature of the pupils who are interested in such activities (e.g., more engaged by maths), rather than as a direct consequence of activity participation. This analysis will therefore be considered exploratory, focusing on associations rather than causal inference.

Items on future life and aspirations

The follow-up pupil survey contains two items relating to the pupil's future life and aspirations. The first of these asks pupils to choose what they would most like to do when they are 18: one of five potential options (e.g., 'Study at university') may be chosen. The number and proportion of pupils selecting each option in the control and intervention groups will be reported, and chi-squared tests will be performed to test for a difference between the groups.

The second item asks pupils to order seven statements from most (rank 1) to least important (rank 7) in their future (e.g., 'Finding a job'). Each statement's median rank in the control and intervention groups will be reported and the distribution of ranks in the control and intervention groups will be compared using Mann-Whitney U tests.

Intra-cluster correlations (ICCs)

The ICC for the maths primary outcome model will be calculated as the proportion of maths score variance attributable to level 2 (between-school) variation. The ICC will be calculated for the maths primary analysis model and for an empty model (i.e., one with no covariates). This will then be repeated for the reading primary outcome model, so there will be four ICCs calculated in total.

Effect size calculation

For all primary and secondary models the difference in the outcome between the intervention and control groups will be presented as an effect size (Hedges, 2007). As these are all linear multilevel models the effect size will be calculated as:

$$\mathrm{ES} = \frac{(\overline{\mathrm{Y}}_{\mathrm{I}} - \overline{\mathrm{Y}}_{\mathrm{C}})_{\mathrm{adjusted}}}{\sqrt{\sigma_S^2 + \sigma_{\mathrm{error}}^2}}$$

$(\overline{\mathrm{Y}}_{\mathrm{I}} - \overline{\mathrm{Y}}_{\mathrm{C}})_{\mathrm{adjusted}}$ is the intervention coefficient from the model, while σ_S^2 and $\sigma_{\mathrm{error}}^2$ are the between-school and within-school variance from an empty model. In order to obtain a 95% confidence interval for the effect size, a confidence interval for $(\overline{\mathrm{Y}}_{\mathrm{I}} - \overline{\mathrm{Y}}_{\mathrm{C}})_{\mathrm{adjusted}}$ will first be calculated by adding or subtracting 1.96 times its standard error. The end points of this confidence interval will then be divided by the denominator in the formula given above.

[14] Nor is it amenable to quasi-experimental methods that seek to mimic randomisation such as instrumental variable analysis. Instrumental variable analysis would require no causal pathway between the instrumental variable (randomisation) and the outcome (KS2 maths score) other than through the variable of interest (e.g., hours spent on sports clubs), which is not the case here.

References

Angrist, J., D. and Imbens, G., D. (1995) 'Two-stage least squares estimation of average causal effects in models with variable treatment intensity', *Journal of the American Statistical Association*, 90(430), pp. 431–442. Available at: https://doi.org/10.1080/01621459.1995.10476535.

Bartlett, J., Keogh, R., Bonneville, E.F. and Ekstrøm, C.T. (2022) 'smcfcs: Multiple Imputation of Covariates by Substantive Model Compatible Fully Conditional Specification'. Available at: https://CRAN.R-project.org/package=smcfcs (Accessed: 19 October 2022).

Bates, D., Mächler, M., Bolker, B. and Walker, S. (2015) 'Fitting linear mixed-effects models using lme4', *Journal of Statistical Software*, 67(1), pp. 1–48. Available at: https://doi.org/10.18637/jss.v067.i01.

Clarke, A.M., Morreale, S., Field, C.-A., Hussein, Y. and Barry, M.M. (2015) *What works in enhancing social and emotional skills development during childhood and adolescence?* Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/411492/What_works_in_enhancing_social_and_emotional_skills_development_during_childhood_and_adolescence.pdf (Accessed: 19 October 2022).

Cullinane, C. and Montacute, R. (2017) *Life lessons: improving essential life skills for young people*. Available at: https://www.suttontrust.com/wp-content/uploads/2017/10/Life-Lessons-Report_FINAL.pdf (Accessed: 19 October 2022).

Fox, J., Kleiber, C., Zeileis, A. and Kuschnig, N. (2021) 'ivreg: Instrumental-Variables Regression by "2SLS", "2SM", or "2SMM", with Diagnostics'. Available at: https://CRAN.R-project.org/package=ivreg (Accessed: 19 October 2022).

Hedges, L.V. (2007) 'Effect sizes in cluster-randomized designs', *Journal of Educational and Behavioral Statistics*, 32(4), pp. 341–370. Available at: https://doi.org/10.3102/1076998606298043.

Jiang, Y. and Small, D. (2014) 'ivpack: Instrumental Variable Estimation'. Available at: https://CRAN.R-project.org/package=ivpack (Accessed: 19 October 2022).

McCroskey, J.C., Andersen, J.F., Richmond, V.P. and Wheeless, L.R. (1981) 'Communication apprehension of elementary and secondary students and teachers', *Communication Education*, 30(2), pp. 122–132. Available at: https://doi.org/10.1080/03634528109378461.

Panorama Education (no date) *Social emotional learning assessment | Panorama Education*, *Panorama Education*. Available at: https://www.panoramaed.com/social-emotional-learning-sel (Accessed: 24 April 2023).

Revelle, W. (2022) 'psych: Procedures for Psychological, Psychometric, and Personality Research'. Available at: https://CRAN.R-project.org/package=psych (Accessed: 19 October 2022).

Romano, J.P. and Wolf, M. (2016) 'Efficient computation of adjusted p-values for resampling-based stepdown multiple testing', *Statistics & Probability Letters*, 113, pp. 38–40. Available at: https://doi.org/10.1016/j.spl.2016.02.012.

Rosseel, Y. (2012) 'lavaan: An R package for structural equation modeling', *Journal of Statistical Software*, 48(2), pp. 1–36. Available at: https://doi.org/10.18637/jss.v048.i02.

Sun, J. and Stewart, D. (2007) 'Development of population-based resilience measures in the primary school setting', *Health Education*, 107(6), pp. 575–599. Available at: https://doi.org/10.1108/09654280710827957.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011) 'Mice: Multivariate imputation by chained equations in R', *Journal of Statistical Software*, 45(3), pp. 1–67. Available at: https://doi.org/10.18637/jss.v045.i03.

Watson, S. (2021) 'crctStepdown: Univariate Analysis of Cluster Trials with Multiple Outcomes'. Available at: https://CRAN.R-project.org/package=crctStepdown (Accessed: 19 October 2022).

Watson, S.I., Akinyemi, J. and Hemming, K. (2022) 'Multiple testing corrections for p-values and confidence intervals from generalised linear mixed models', *Mathematics*, 1(3). Available at: https://doi.org/10.48550/arXiv.2107.10017.

Appendix A

```
## Stratified Randomisation code

##1. Set work directory
setwd("…/Randomisation")
dir<-dir()

## 2. Load data, a file consisting of all schools to be randomised, as well as their CU locality and
## NFER number (a school-level identifier, meaningless outside this project)
Experiment<-read.xlsx(paste(getwd(),"/","Randomisation school file.xlsx",sep=""))

names(Experiment)[c(1,3)]=c("NFERNo", "CU")

## Identify stratification and unique identifier variables

## 3.list the stratification variables (is this case just one stratifier, CU locality)
stratification<-list("CU")
n_strats<-length(stratification)

## 4.specify unique identifier variable, in this case NFER number
ui<-"NFERNo"

## 5. What time is now? (hh.mm)(e.g: 11.35), used to create a random seed
time_now<-14.29

aux<-100*trunc(time_now)+100*(time_now-trunc(time_now))
set.seed(aux)
seeds<-sample(1:9999,size=(n_strats+2))

## check for duplicated or missing unique identifier information
## must remove
Experiment<-Experiment[!duplicated(Experiment[ui]),]
Experiment<-Experiment[!is.na(Experiment[ui]),]
```

```
## store the original order of the columns
originalColOrder<-colnames(Experiment)

## Adding a variable that will allow for the recovery
## of the original order of the data frame rows later on
Experiment$originalRowOrd<-1:nrow(Experiment)

### Ordering Experiment by unique identifier
Experiment<-Experiment[order(Experiment[ui]),]

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

for (i in 1:n_strats){
  aux<-as.data.frame(sort(unique(Experiment[,stratification[[i]]])))
  set.seed(seeds[1])
  seeds<-seeds[-1]
  aux[rands[i]]<-sample(1:nrow(aux)
  Experiment<-merge(Experiment,aux,by.x=stratification[[i]],by.y=colnames(aux)[1])
}

## Randomise by unique identifier
set.seed(seeds[1])
seeds<-seeds[-1]
Experiment["rand_ui"]<-sample(nrow(Experiment))

## Reorder the rows of Experiment by rands and rancluster
rands<-c(rands,"rand_ui")
aux<-do.call(order,Experiment[rands])
Experiment<-Experiment[aux,]

## Assigning Control or Intervention Group

Experiment$grp<-(1:nrow(Experiment))%%2+1

rands<-c(rands,"grp")

aux<-data.frame(group=c("Control", "Intervention"))
set.seed(seeds[1])
aux$randgroup<-sample(1:2)

Experiment<-merge(Experiment,aux,by.x="grp",by.y="randgroup")

## Returning the data frame to its original order
Experiment<-Experiment[order(Experiment$originalRowOrd),]

## Removing the variables that are no longer necessary
originalColOrder<-c(originalColOrder,"group")
Experiment<-Experiment[,originalColOrder]
```