Trial Evaluation Protocol: KEEP Teaching

Evaluator (institution): UCL Institute of Education

Principal investigator(s): David Wilkinson

PROJECT TITLE	KEEP Teaching (KEep Early-career Physicists Teaching)
DEVELOPER (INSTITUTION)	Institute of Physics
EVALUATOR (INSTITUTION)	UCL Institute of Education
PRINCIPAL INVESTIGATOR(S)	David Wilkinson
PROTOCOL AUTHOR(S)	David Wilkinson, Mark Hardman, Marian Mulcahy, Sam Sims
TRIAL DESIGN	Two-arm randomised controlled trial with random allocation at school(teacher) level
TEACHER CAREER STAGE AND SCHOOL TYPE	Early career teachers in secondary schools
NUMBER OF SCHOOLS	200
NUMBER OF TEACHERS	200
PRIMARY OUTCOME	Teacher Job satisfaction
SECONDARY OUTCOME	Teacher Retention in the profession Teacher Retention in school

Protocol Version History

VERSION	DATE	REASON FOR REVISION
2.0	07 February 2022	A number of changes were required as a result of Covid 19: The primary outcome was changed from teacher retention in the profession to teacher job satisfaction. This reflected the lower number of Physics NQTs, which meant that the trial would be underpowered in terms of detecting an impact on teacher retention, but not on detecting an impact on teacher job satisfaction. Teacher retention will remain as a secondary outcome. The IPE was strengthened to explore the impact of Covid-19 on timetabling and on the validity of the trial. Timetabler and Head of Department surveys were delayed for the first cohort, to avoid additional stress for teachers during Covid-19 lockdown. This was found to support direct comparisons with Newly Qualified Teacher (NQT) surveys. Additionally, the implementation of all surveys will now be at the end of the school year. GCSEs did not take place in summer 2020 and 2021, hence it was decided to drop school-level attainment as a secondary outcome. Additional analysis is included that will explore the relationship between outcomes and the 'matchedness' of timetables. This will not exploit the experimental design of the project, but will provide further information on how 'matchedness' of timetables is associated with teacher outcomes.
1.0	7th May 2019	Original version

Contents

Intervention

The short description that follows is based on the *Template for Intervention Description and Replication* (TIDieR) checklist, which should be read in conjunction with the provisional logic model (see Figures 1 and 2).

1. Brief name

KEEP Teaching.

2. Why (rationale/theory)

Science teachers are more likely to leave the profession than non-science teachers, particularly within their first five years of teaching. One distinctive characteristic of science teaching in England is that individuals usually have to provide instruction in three subjects: biology, chemistry and physics. This means that science teachers are usually providing instruction in at least two subjects in which they do not have a degree.

Research suggests that early-career teachers who were asked to teach multiple subjects were more likely to leave the profession. For example, Donaldson and Johnson (2010) study three cohorts of Teach for America teachers and find that those with multiple-subject assignments were more likely to leave the teaching profession than those with single subject assignments. Sims (2018) also finds that teachers who feel unprepared for the subjects they are assigned to teach have higher turnover intentions. Olmo (2010) comes to similar conclusions but also provides evidence that this relationship is explained by teachers feeling less capable when teaching outside of their specialism. Sims and Jerrim (2000) show that teachers in England who report lower job satisfaction are also much more likely to leave their jobs, or profession. A one standard deviation increase in job satisfaction is associated with a 40% reduction in the odds of leaving the school and a 57% reduction in the odds of leaving the profession. MetLife (2006) explored the reasons behind teachers in America leaving the profession using interviews with a nationally representative sample of US public school teachers. The findings showed the teachers being assigned to classes which they did not feel qualified to teach was amongst the most important reason for them leaving the profession. Qualitative evidence comes from the Future Physics Leaders – Newly Qualified Teacher (NQT) support programme by the Institute of Physics (IOP). Case study interviewees reported the positive effect of matched timetabling (a greater proportion of lessons within subject specialism) in supporting the development of physics NQTs[1].

In addition to impacts upon retention, there is some evidence that pupils being taught by specialists also leads to improved attainment outcomes. A 2005 study from the US explored the effects of a teacher's mathematical knowledge on students' mathematical achievement (Hill, Rowan & Ball, 2005). This study examined student achievement data from 1,190 first grade students, 1,773 third grade students, 334 first-grade teachers and 365 third-grade teachers in 115 elementary schools, with an over-representation of high-poverty schools. The study assessed students twice in each academic year and found students' growth in mathematics scores translated to one-half to two-thirds of a month of additional growth per

[1] Please note that the term NQT: Newly Qualified Teacher is still widely used in schools, although it is increasingly being replaced by reference to Early Career Teachers (ECTs), in their first two years of teaching. This change is due to the introduction of new policy around supporting ECTs from 2021, after this trial started. We therefore continue to use the term NQT to refer to teachers in their first year of teaching after qualification.

standard deviation difference in a teacher's mathematical content knowledge. Cook and Mansfield (2016) also find that teachers who accumulate more experience teaching a specific subject are more effective. Unfortunately, due to the cancellation of GCSEs in 2020 and 2021, we do not have the data to assess pupil attainment as an outcome.

Alongside the evidence supporting subject specialisation at secondary level, it is important to note that there is also evidence that students benefit from increased teacher-student familiarity. A recent study using administrative data in North Carolina (Hill and Jones, 2018) found that where students were assigned the same elementary school teacher as they had been taught by in a previous year, they made small but significant test score gains compared to their counterparts. These effects appeared strongest for students from ethnic minority backgrounds and students taught by less effective teachers. If students are taught by multiple specialist science teachers, those teachers are likely to be less familiar with them. The overall effect of specialist physics teachers therefore remains to be established.

This study aims to understand the impact of increasing the proportion of time Physics NQTs spend teaching Physics (through better aligned timetabling) on their job satisfaction. Figure 1 highlights that this timetabling approach is expected to work from reduced workload and improved pedagogical content knowledge. Improved job satisfaction may result in better retention of Physics NQTs. The attainment of pupils may also be enhanced as a result of being taught by teachers with enhanced pedagogical content knowledge.

3. Who (recipients)

Small-scale scoping research for this trial found that the proportion of lessons within specialism for physics NQTs varied between 20 and 80%. The small sample size means this should be treated as a lower bound on the extent of variation. The unit of analysis for this trial will be non-selective state schools within England who have recruited a physics NQT. For the purposes of this project, a physics NQT is defined as someone with a physics or 'maths with physics' degree or mechanical, civil or electrical engineering degree; and/or someone with a physics or physics and maths Postgraduate Certificate in Education (PGCE) (or Qualified Teacher Status (QTS)-equivalent teaching certificate labelled as specialising in physics or physics and maths).

4. What (materials)

Tailored guidance materials are provided to schools based upon an initial assessment of their timetabling processes. These materials may include guidance, example timetables and direct feedback upon timetables.

5. What (procedures)

Eligible schools may be engaged in the trial prior to them recruiting a physics NQT. Likewise, physics NQTs may be engaged in the trial prior to gaining employment at an eligible school. Crucially however, the unit of randomisation in this trial is a physics NQT and eligible school pairing. As a result, no eligible school or physics NQT will be deemed to be a participant in the trial until they have formed such a pairing and both the school Senior Leadership Team (SLT) and the physics NQT have jointly signed a Memorandum of Understanding (MoU). The MoU will outline the processes and expectations within the project.

Once a matched pair of physics NQT and eligible school are allocated to a trial condition, the Institute of Physics (IOP) team will then offer guidance to promote 'matchedness' of the timetable for the physics NQT, if they are a treatment school. A timetable is more matched if the physics NQT has a greater proportion of their lessons in physics. Once a draft timetable is produced, this will be shared with the IOP who may repeat the cycle of guidance and provide feedback to maximise 'matchedness'.

6. Who (implementers)

A project manager at the IOP (0.4 Full-Time Equivalent - FTE). The project has a 0.7 FTE administrator and a dedicated Marketing Officer during the recruitment stages of the project (0.7 FTE)

7. How (mode of delivery)

Tailored guidance will be delivered through e-mail exchange, phone calls and some face-to-face liaison.

8. Where (setting)

Non-selective state secondary schools in England.

9. When and how much (dosage)

Once a physics NQT is recruited to an eligible school the matched pair will then be randomised into the intervention and control condition. Those within the intervention will receive tailored guidance as soon as possible, ideally before May each year (since much of the timetabling takes place May-July). Since the guidance is tailored there will be variation in the specific amount of guidance provided. Some schools may also receive additional cycles of guidance and feedback in order to further improve timetables. Dosage is considered to relate to the level of 'matchedness' which is achieved in implemented timetables, which will be monitored within the trial.

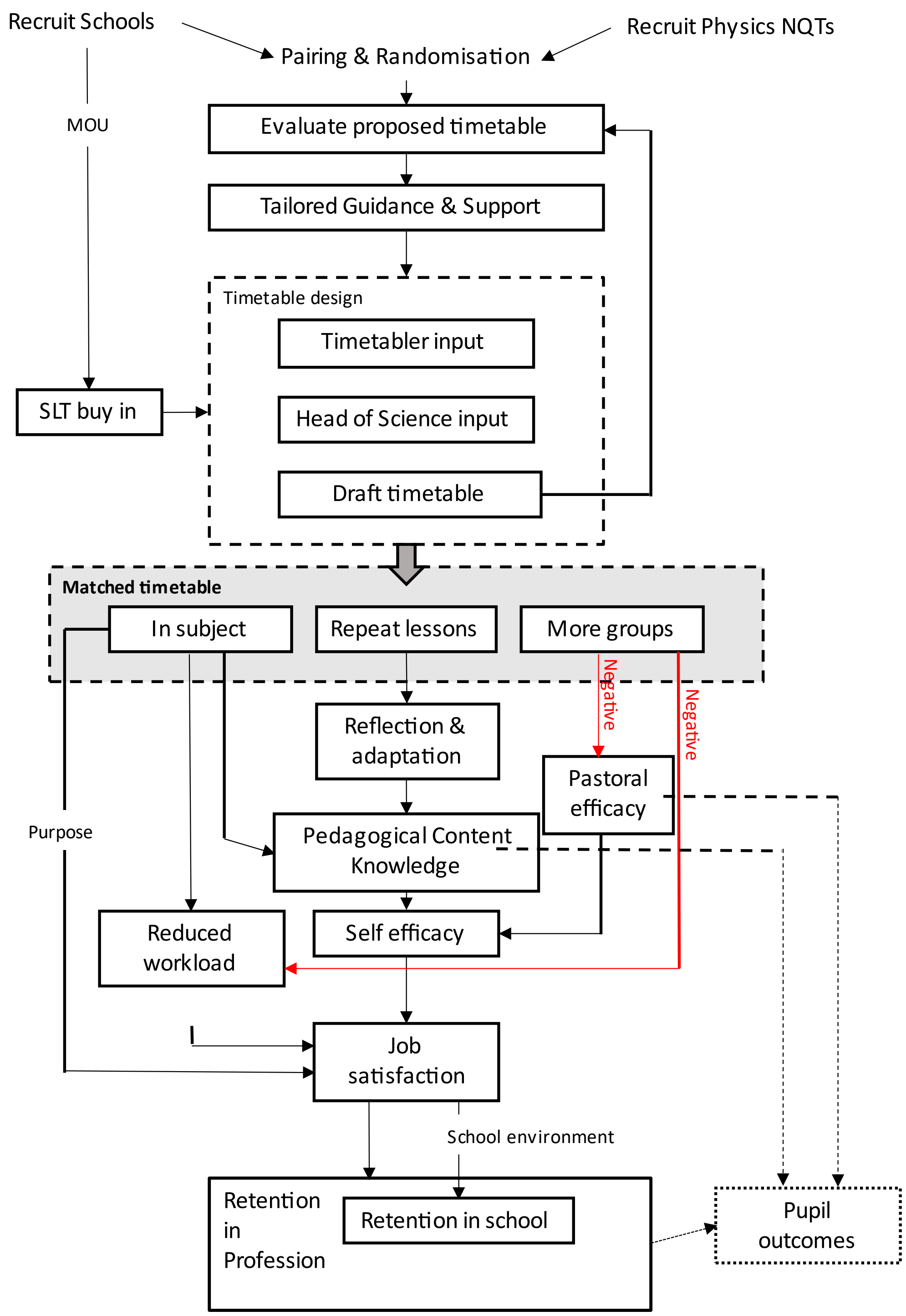
10. Tailoring

A high level of tailoring is anticipated in the nature of the IOP guidance to schools. Schools may also vary in the way that they achieve 'matchedness' of timetables. However, the ultimate aim is to achieve greater 'matchedness' of timetables.

11. How well (planned)

The intervention itself is relatively simple to implement as it requires input only during the period over which timetables are set. However, effective implementation requires the support of senior leaders, timetablers and heads of science in schools. There is an expectation of the need to repeat and escalate guidance and support in some schools.

Figure 1 – KEEP Teaching Logic Model



Study Rationale and Background

Teacher shortages are a persistent and widespread problem in public school systems and are particularly severe among science teachers (Dolton, 2006). While other subjects generally see shortages eliminated during economic downturns, shortages of science teachers tend to persist between economic cycles (Goldhaber et al., 2014; Smithers & Robinson, 2008). Where no appropriately qualified teachers are available, research shows that school leaders tend to lower recruitment standards, make increased use of temporary teachers or increase class sizes (Moor et al., 2006; Smithers & Robinson, 2000), all of which have been linked with reduced pupil attainment (Fredriksson et al., 2012; Mocetti, 2012; Schanzenbach, 2006). Finding ways to reduce the shortage of appropriately qualified science teachers is therefore important.

The shortage of these teachers is due in part to the higher rates at which science teachers leave the profession (Kelly, 2004; Worth & De Lazzari, 2017). This in turn reflects the fact that teachers with Science, Technology, Engineering and Maths (STEM) degrees face a higher outside pay ratio than other teachers (MAC, 2016), providing them with monetary incentives to move to jobs in other sectors.

Science teaching is also in some ways more demanding than teaching in other subjects, in that science teachers generally have to teach one or two subjects in which they do not have a degree. For example, a chemistry graduate will generally be expected to teach biology and physics, as well as chemistry. The high numbers of physicists choosing to teach mathematics instead of mixed science (Smithers & Robinson, 2008), suggests perhaps that this is seen as undesirable by teachers. Teaching multiple subjects also increases workload because teachers have to teach a higher number of unique lessons per year, each of which comes with its own amount of preparation time. This is likely to be why early-career teachers who are given multiple subjects to teach are also more likely to leave the profession (Donaldson & Johnson, 2010). Teachers who report lower job satisfaction are also much more likely to leave their jobs and the profession (Sims and Jerrim, 2020), hence the demands of teaching science therefore also help to explain the higher levels of wastage (leaving the profession) among science teachers.

To support our understanding of workloads for science NQTs we carried out some scoping work to ascertain (i) when timetables are set in schools, (ii) what is the scope for change in timetables in terms of changing the extent to which NQTs teach within their specialism and (iii) would changing NQT timetables effect other members of the science department.

Our findings indicated that, in general, timetabling happens when staffing is fixed for the next year, typically after the May half-term notice window. This is after pupils have chosen their subjects for the next academic year and once Key Stage 4 groups are determined and often when the exam period is over.

We modelled timetables in terms of factors that would reduce teacher workload, such as: the percentage of lessons within specialism, the percentage of a teacher's groups that are within the same year group and the number of unique pupil groups. We found considerable variation in NQT timetables across all three dimensions suggesting some scope for change in timetables. However, in smaller schools (less than 1,000 pupils) there was less variation, suggesting a possible school size constraint in setting timetables. We also found no obvious trade-offs in terms of these indicators between NQTs extent of specialization and the extent of specialization of other teachers.

This suggested that there is scope to influence the ‘matchedness’ of timetables without affecting other teachers.

Impact Evaluation

Research questions

The evaluation will address the following primary research question:

> **RQ1.** What is the size of the effect of the KEEP Teaching intervention on the job satisfaction of physics NQTs towards the end of their NQT year?

In addition, the evaluation will address the following secondary research questions:

> **RQ2.** What is the size of the effect of the KEEP Teaching intervention on the retention within the teaching profession of physics NQTs three years after starting their NQT year at that school, compared to a business-as-usual control?
>
> **RQ3.** What is the size of the effect of the KEEP Teaching intervention on the retention within a school of physics NQTs three years after starting their NQT year at that school, compared to a business-as-usual control?
>
> **RQ4.** What is the association[2] between the extent of 'matchedness' and job satisfaction, as well as 'matchedness' and teacher retention?

Focusing on both within profession and within school retention allows us to focus on whether KEEP Teaching reduces wastage from the profession, as well as whether KEEP Teaching improves retention rates within the school where the NQT spent their NQT year. The former is the parameter of interest from a public policy perspective while the latter is more important for each specific school.

Additional questions relating to the Implementation and Process Evaluation (IPE) are discussed below.

Design

The trial was originally designed as a school-level randomised controlled trial involving 300 schools with a Physics NQT. However, due to Covid-19, recruitment is likely to be around 200 schools. It will remain a two-arm efficacy trial: KEEP Teaching compared to a business as usual control. The 200 schools with NQTs will be recruited over a three-year period with an expectation of recruiting 60 to 70 each year, with equal allocation to treatment and control in each of the three cohorts. We expect to allocate 100 schools to the intervention and 100 to the business as usual control. See the timeline below for further details.

In addition, for the third year of the recruitment period, the recruitment window was extended to provide a bigger sample of matched schools and NQTs. These pairings will not be randomly assigned to an intervention and control group. This is because these schools and NQTS will have been recruited after timetabling has taken place. However, they will provide an additional group of schools and NQTS that will allow us to examine the associations between 'matchedness' and outcomes as highlighted in Research Question 4 above.

The primary outcome will be the job satisfaction of the NQT, with data collected via an online survey sent to NQTs towards the end of the year. Secondary outcomes will be (i) the retention of the NQT in the teaching profession using data from the School Workforce Census (SWC);

[2] In this analysis we will not exploit the randomised control trial design, instead we will consider associations between the indicators using the full sample or teachers from both the treatment and control groups.

and (ii) retention of the NQT in the same schools they were working at during their NQT year, also using data from the SWC. To allow time for retention (job exits) to be observed we will examine retention up to three years after the NQT year.

<table>
<tr><th colspan="2">Trial type and number of arms</th><td>Two-arm randomised</td></tr>
<tr><th colspan="2">Unit of randomisation</th><td>School / NQT pairing</td></tr>
<tr><th colspan="2">Stratification variables
(if applicable)</th><td>None</td></tr>
<tr><th rowspan="2">Primary outcome</th><td>variable</td><td>NQT job satisfaction</td></tr>
<tr><td>measure
(instrument, scale)</td><td>This data will be collected from a short online survey in the summer term of the NQT year</td></tr>
<tr><th rowspan="2">Secondary outcome(s)</th><td>variable(s)</td><td>NQT retention in profession
NQT retention in school</td></tr>
<tr><td>measure(s)
(instrument, scale)</td><td>Whether the NQT remains in the teaching profession in the state sector up to 3 years after starting their NQT year.

Whether the NQT remains in the same school as they were for their NQT year, up to 3 years after starting their NQT year.

This data will be matched to NQTs from the SWC for the three years following their NQT year.</td></tr>
</table>

Randomisation

Schools will be recruited into the trial over the course of three years with a target of 60 to 70 schools per year. We want to randomly allocate schools to treatment or control as soon as they join the trial, sometimes known as sequential treatment allocation. This is because we want to maximise the amount of time the implementation team have to work with schools prior to schools finalising their timetable.

One way to achieve this would be to simply flip a coin (or generate a random 0 or 1 using a computer) every time a school is recruited. However, this may result in an uneven number of schools in treatment or control groups. This is undesirable because there will be a small reduction in power associated with uneven allocation to treatment and control.

In order to allow sequential treatment allocation while also ensuring equal size treatment and control groups, we adopt the following method for randomisation:

1. We generate a dataset with two hundred rows labelled 1, 2, 3... 200.
2. We flip a coin (using the computer) for row 1. We assign row 1 to treatment if it is heads, or to control if it is tails.
3. We repeat the process for the following 199 rows.
4. We then check the number of heads. If the number of heads is not 100, we repeat steps 1-4 until we generate a dataset with exactly 100 heads.

5. The first school recruited to the trial is assigned to treatment if row 1 is a treatment row, or to control if row 1 is a control row.
6. We repeat the process in step 5 for the next 199 schools recruited to the trial.

Sequential allocation is sometimes critiqued because it can introduce biases when the recruiter knows what the next allocation will be. In this study, this approach is acceptable because the evaluation team will conduct the randomisation and the recruiter will not know what the next allocation will be. The randomisation process will be recorded in the syntax and log files used to carry out the randomisation and included as an appendix in the evaluation report.

Participants

Non-selective state schools within England matched with a physics NQTs are eligible for the trial. For the purposes of this project, a physics NQT is defined as someone with a physics or 'maths with physics' degree or mechanical, civil or electrical engineering degree; and/or someone with a physics or physics and maths PGCE (or QTS-equivalent teaching certificate labelled as specialising in physics or physics and maths).

Eligible schools may be engaged prior to them recruiting a physics NQT. Likewise, physics NQT may be engaged in the trial prior to gaining employment at an eligible school. However, the unit of randomisation in this trial is a physics NQT and eligible school pairing. As a result, no eligible school or physics NQT will be deemed to be a participant in the trial until they have formed such a pairing and both the school SLT and the physics NQT have jointly signed a MoU. Schools and NQTs will sign a MoU which outlines the processes and expectations within the project, with the school MoU also signed by the SLT.

Sample size calculations

The developer originally planned to work with 300 teacher-school pairings, and the primary outcome was to be teacher retention. Original sample size calculations – see earlier version of the protocol – highlight that this would have allowed us to detect a reduction in turnover at a reasonable level.

Due to Covid-19, recruitment was more difficult, and we were forced to re-evaluate our previous target to 200 teacher-school pairings. This was approximately 70 teacher-school pairings per year, across three years. The primary outcome was changed to be teacher job satisfaction and our power calculations are based on having schools evenly split between intervention and control group with standard assumptions of power (0.8) and significance (0.05). We additionally assume that 30 per cent of the variation in the outcomes are explained by covariates in the model, based on analysis of teacher job satisfaction in Sims and Jerrim (2020). Altogether, this yields a minimum detectable effect size of 0.26.

For the secondary retention outcomes, we will use a survival analysis approach, estimating the probability of the NQT staying in employment one, two and three years after their NQT year. The minimum detectable effect size depends in large part on the number of units expected to have experienced the event by the end of the study. In this study, our secondary outcome measure is the 'event' of the physics NQT leaving state funded teaching (or leaving the school they were at during their NQT year) within three years. Based on Physics NQT teacher retention rates using official Department for Education (DfE) data[3], we assume 35 per cent of new physics teachers leave the profession within three years.

[3] This data comes from the Teachers Analysis Compendium 4 Entrants, Leavers and retention Statistics (Pilot) https://department-for-education.shinyapps.io/turnover-and-retention-grids/

The minimum detectable effect size also depends on a number of other things. First, the proportion of the variation in the outcome that can be explained by covariates. There is a lack of guidance for this in the empirical literature. We pragmatically assume it to be 30% but also note that our final result is insensitive to changing this to 20% or 40%. Second, the standard deviation of the outcome variable. We derived this from a simulation based on official DfE data on physics NQT retention rates. We assume this is 0.79. We make the standard assumptions of a power of 0.8 and a two-tailed significance test at the 0.05 level. With a sample of 200 teachers we estimate the hazard ratio (which is the appropriate measure of effect size in survival analysis models) of 0.60. This means that we would be able to identify a 40 per cent reduction in the odds of leaving the profession in the treatment group compared with the control group.

It should be noted that a 40 per cent reduction in odds is not the same as a 10 per cent reduction in probability. To see the difference between probability and odds consider that the probability of flipping heads on a fair coin is 0.5, whereas the odds of flipping heads is the probability of flipping heads over the probability of flipping tails, or 0.5/0.5 = 1. This is not particularly intuitive, so we provide a worked example of what a hazard ratio of 0.60 means in practice.

At the outset of the trial, data from the DfE shows that physics NQTs have a 0.35 probability of leaving the profession after three years. Their odds of leaving the profession are therefore 0.35/0.65 = 0.53. Our trial is powered to detect a 40% reduction in these odds. Reducing odds of 0.53 by 40% gives an odds of 0.32. An odds of leaving of 0.32 is equivalent to a probability of leaving of 0.24 (0.24/0.76 = 0.32). Hence, if the baseline probability of leaving the profession after three years was 0.35, our trial would be able to detect the effect of the intervention if it reduced the probability of leaving to 0.24, or below (i.e., a reduction of the probability of leaving the profession after 3 years of 11 percentage points).

We calculated the minimum detectable effect size using the Power Cox command in the Stata software.

		OVERALL
Effect size		0.26
Mean of outcome measure		3.67
Standard Deviation of outcome measure		0.72
Proportion of variance explained by covariates		0.3
Alpha		0.05
Power		0.8
One-sided or two-sided?		2
Number of schools	Intervention	100
	Control	100
	Total	200

Outcome measures

Primary outcome

The primary outcome will be NQT job satisfaction. This measure will come from an online survey set to be implemented towards the end of participant's NQT year. The measure is based on Thompson and Phua (2012). Respondents are asked how far do you agree with these statements about your job?

- I find real enjoyment in my job
- I like my job better than the average person
- I am seldom bored with my job
- I would not consider taking another kind of job
- Most days I am enthusiastic about my job
- I feel fairly well satisfied with my job
- In the last term, I have seriously considered moving school
- In the last term, I have seriously considered leaving teaching altogether

Responses are coded - 1 strongly disagree, 2 disagree, 3 neither disagree/agree, 4 agree, 5 strongly agree. The last two items are new and are reverse coded. Mean scores of all items will be considered.

Secondary outcomes

The retention analysis will consider the time to leaving teaching in the state sector in England, and the time to leaving the school they started their NQT year. This will use data from the SWC collected in November each year.

Analysis plan

All quantitative outcomes would be modelled on the basis of intention to treat (ITT). The primary outcome will be NQT job satisfaction, discussed above. We will estimate a linear regression model including a dummy variable indicating whether the NQT was in the intervention or control group. The estimated impact will be converted into a Hedges' g effect size (Hedges, 1981).[4]

For both of the retention outcomes we will use a survival analysis approach using Cox Proportional Regression incorporating the treatment condition. We will also report results from a Weibull proportional hazard model as a robustness check. We will analyse data from two years after completion of the intervention to allow time for a reasonable number of job exits to be observed. Descriptive analysis will also be conducted, using survey data, on the retention and job move intentions of NQTs reported at the end of their NQT year. This will allow us to consider intentions to move into teaching jobs outside of the state sector. Such jobs will not be identified in the School Workforce Census data that will be used for the main retention analysis.

Survival analysis takes into account both: the time to an event and the event status, which records if the event of interest has occurred (or not yet). This allows us to correctly incorporate information from censored (completed) and uncensored (ongoing) job spells when estimating important model parameters. Thus, the total teaching years contributed in both the treatment and control groups can be compared. We will also use Cox Proportional Regression, which does not make assumptions about the baseline hazard function but does assume that treatment and control hazards are proportional to each other over time. This will allow us to

[4] EEF Statistical Analysis Guidance: https://d2tic4wvo1iusb.cloudfront.net/documents/evaluation/evaluation-design/EEF_statistical_analysis_guidance_2018.pdf

control for all observable variables, increasing power and soaking up any residual bias from the randomisation. Following Clotfelter et al. (2008), we will also report results from a Weibull proportional hazard model as a robustness check. In this case, the time-to-event would be the period between the first day of the first term of employment as a physics NQT and leaving employment in state funded education in England. To allow time for a reasonable number of job exits to be observed we propose conducting our analysis using data from two years after completion of the intervention.

Given the modest sample, no subgroup analysis will be conducted, but we will analyse missing data. The analysis will be detailed in the Statistical Analysis Plan (SAP), which, in line with the EEF procedures, will be produced three months after randomisation.

Additional regression analysis will examine the relationship between the ‘matchedness’ of timetables (dosage) and teacher job satisfaction and teacher efficacy. We know from earlier work that there is considerable variation in the ‘matchedness’ of timetables in both the intervention and control groups. We will exploit this variation for a sample that includes pooled data from both groups. Here we address RQ4 from page 10 by assessing the following questions:

- Is a change in the degree of ‘matchedness’ associated with a change in teacher job satisfaction?
- Is a change in the degree of ‘matchedness’ associated with a change in teacher efficacy?

This analysis only requires NQTs to provide timetables and complete the survey towards the end of the academic year. Importantly, no intervention is required. This means that for the most recent cohort, recruitment can continue for longer, potentially giving us a larger sample for this analysis.

Implementation and Process Evaluation

A robust and in-depth implementation and process evaluation (IPE) is vital to ensure we understand how the KEEP Teaching intervention is implemented, and the extent to which the logic model (see Figures 1 and 2) adequately describes the factors and mechanisms underlying the intervention as well as the key conditions for success and any barriers to implementation. Our IPE will take a mixed methods approach. We outline the RQs, data collection and analysis below.

Additional research questions were added in Autumn 2020 in response to COVID-19. These are presented below and make explicit reference to the pandemic - sub-questions (v) and (vi) under RQ1, and sub-question (iv) under RQ3.

We recognise that the characteristics of job satisfaction may have changed due to COVID-19. Nevertheless, RQ4 allows us to explore the factors influencing job satisfaction within both the intervention and control groups.

Research questions

In the process evaluation, we will address the following research questions:

RQ1. What are the barriers (and opportunities) to modifying timetables for physics NQTs?

i. How far can the intervention be delivered with fidelity?
ii. How responsive are schools: how many cycles of guidance are required?
iii. How far is guidance tailored for each department?
iv. What are the common barriers (and opportunities) in the processes of timetabling within schools?
v. What influence did COVID-19 have on timetabling for physics NQTs during national lockdown?
vi. What influence did COVID-19 have on timetabling for physics NQTs outside of period of national lockdown?

RQ2. How far can physics NQT timetable 'matchedness' be achieved by science departments and schools?

i. What is the variation in 'matchedness' across intervention schools?
ii. What is the variation of 'matchedness' between intervention and control schools?
iii. What factors influence the level of 'matchedness' achieved?

RQ3. What factors influence the impact of the intervention?

i. What is the perceived quality of the guidance? How far do those involved in timetabling report that the guidance leads to change?
ii. Is the intervention timely to change timetabling practices?
iii. How far does the intervention differ from normal practice in the control schools?
iv. How far did COVID-19 influence schools' capacity to implement the intervention?

RQ4. What causal processes reflected in the logic model can be supported?

i. What role does a 'sense of purpose' through teaching in subject, play in job satisfaction?
ii. What role does reduced workload, from teaching in subject, play in job satisfaction?

iii. How does time for reflection and adaptations effect pedagogical content knowledge and self-efficacy?
iv. What negative effects do increased number of groups play in relation to 'pastoral efficacy')

Implementation and process evaluation data collection

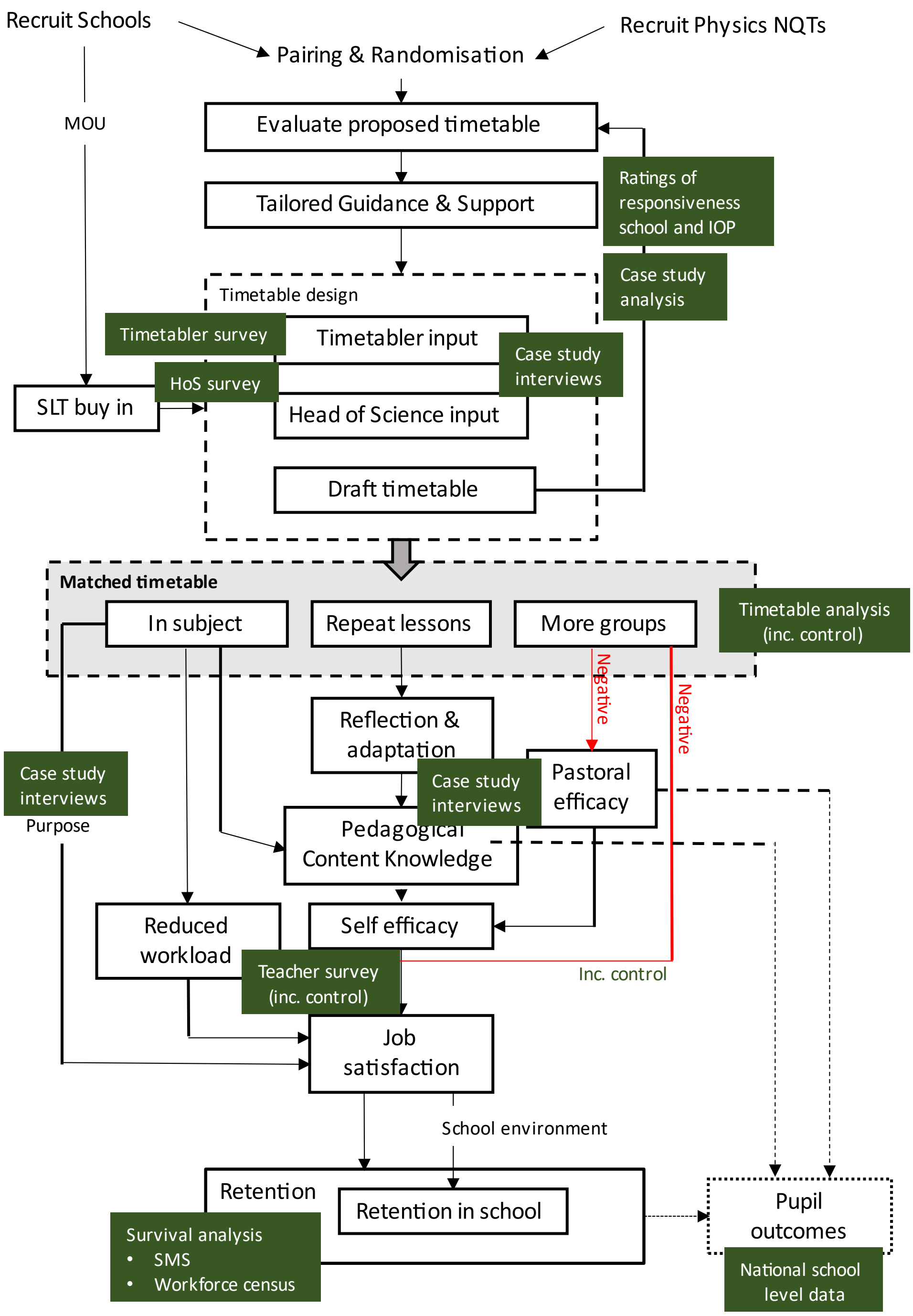
Data collection will involve timetable analysis, analysis of guidance and support materials, surveys, case studies and interviews as set out below. These will be conducted by UCL researchers.

The IPE covers the EEF dimensions for efficacy trials programmes, as specified in Humphrey et al. (2016) "*Implementation and process evaluation (IPE) for interventions in education settings: An introductory handbook*" as well as in the EEF's IPE guidance.[5]

[5] EEF IPE Guidance: https://d2tic4wvo1iusb.cloudfront.net/documents/evaluation/evaluation-design/IPE_guidance.pdf?v=1630567483

IPE Dimension / Factor	IPE RQs	Meaning	Data
Fidelity	RQ1 (i)	The intervention has happened: IOP have given tailored guidance after evaluating draft timetable. Nothing beyond this is provided (e.g. mentoring)	Timetabler and Head of Department (HoD). Case study follow up interviews.
Dosage	RQ2	Level of 'matchedness' in NQT timetable.	Timetable 'matchedness' analysis
Quality	RQ3 (i) RQ3 (ii)	Guidance: clarity of guidance, process, timeliness of the intervention.	Guidance analysis in case studies.
Reach	RQ2	A measure of how many make a change.	HoD surveys and 'matchedness' analysis.
Responsiveness	RQ1 (ii)	Responsiveness of school – how much contact and cycles of guidance are required.	HoD/Timetabler surveys, ratings of school responsiveness from IOP.
Programme differentiation (and the assessment of 'usual practice' at baseline and endpoint)	RQ2 (ii) RQ3 (iii)	What is the difference in 'matchedness' between intervention and control? What is the difference between intervention and normal practice in determining timetables?	'Matchedness' analysis, HoD/Timetabler surveys (intervention and control)
Monitoring of control group	RQ2 (ii) RQ3 (iii)	Did the control modify their timetable? As a result of being part of the trial? Was there any contamination from intervention schools?	Control survey – timetabler / head of science
Adaptation	RQ1 (iii)	The differences in guidance and support offered to different schools.	Case studies – monitoring changes to guidance over years of trial.

Figure 2 – KEEP Teaching logic model with data sources



Timetable analysis

Through an initial scoping study, we have developed an approach to timetable analysis which allows the ‘matchedness’ of timetables to be evaluated.

Timetables will be collected from each school/NQT pairing within their first term of teaching in the school. Timetables of NQTs in control schools will also be collected in the Autumn term of each year that the trial runs. This analysis defines ‘matchedness’ in relation to:

1) Teaching as many lessons as possible within specialist subject. This is operationalised as the proportion of classes which are within the teacher’s specialism. Specialism may be physics, or physics with maths, depending upon training route and is self-reported. Where a teacher teaches combined science, the rota of topics is examined to estimate the proportion of lessons which are physics across the year.

2) Repeating lessons over the year (e.g., by having classes in the same year group). This is operationalised as the number of groups who are taught in the same year group as one or more other groups, as a proportion of the total number of unique groups taught.

3) Having as few groups as possible. This is operationalised as the number of unique groups as a proportion of lessons taught per rotation (one or two weeks in most schools). This accounts for part-time working although NQTs are usually full time.

The concepts used to define ‘matchedness’ are included in the table below:

Concept	Operationalised	Comments
Lessons within Specialism	Number lessons in self-stated specialism□Number	Self-report specialism in order to simplify
Repeated lessons	Number of groups in same year as another group□	Groups in same year is best proxy for repeat lessons
Not too many pupils	Number unique groups □Number lessons taught □	Groups = pupils in same class. NQTs usually FT.

Surveys

Timetabler and Head of Department surveys investigate the processes involved in determining NQT timetables, in both intervention and control schools. This was originally intended to involve a single survey in September/October for each cohort. However, due to COVID-19, timetabler and Head of Department surveys were delayed for cohort A in order to avoid additional stress to these teachers and were delivered towards the end of the year along with the survey of NQTs. We found that this supported direct comparison of the accounts of NQTs with their school colleagues at a single time-point, allowing us to triangulate ratings and identify any differentiation in consideration at school versus department level. It also allowed us to capture the reflections of timetablers and Heads of Department on the impact of timetables across the school year. For these reasons, and to enhance comparability across cohorts, we delivered all surveys at the end of the school year for later cohorts also.. Quantitative responses provide feedback on the ease with which the timetables are adapted, and on the ratings of the guidance as well as on IOP's support in this process. Qualitative responses detail any barriers and considerations during this process.

Teacher surveys in both intervention and control schools will investigate the influences upon job satisfaction specified within the logic model (see Figures 1 and 2).

Surveys are short and use online technology. They use previously validated items where possible, for example in using existing tools for rating quality of guidance and support. The job satisfaction scale is adapted from Thomson and Phua (2018) with two new items added; the teacher efficacy scale is the short form from Tschannen-Moran and Woolfolk-Hoy (2014); and the workload stress scale is adapted from TALIS (2018) also with two new items added.

Case studies

Case studies provide data to not only triangulate the findings of surveys, but also provide greater depth of understanding around the mechanisms, barriers and affordances of providing matched timetabling. They also explore the influences upon job satisfaction and retention postulated within the logic model. Data collected was intended to include:

- Online interviews with the timetabler and head of department ~ 40 per year: intervention and control schools.
- Online interviews with NQTS ~ 20 per year: intervention and control schools.
- Visits to schools for face-to-face meetings with timetabler, head of department, senior leaders, NQTs and other science teachers ~ 5 per year: intervention schools.
- Gathering of materials and guidance exchanged between the IOP and school ~ 5 schools per year: intervention schools.

Case studies were to be selected at random within the first year of evaluation. In subsequent years we intended to also follow schools who have previously been involved, to gain a longitudinal picture. However, we anticipated these being rare given the need for a school and physics NQT pair each year.

Due to COVID-19 we were not able to conduct case studies with cohort A. It was not tenable to visit schools and we also wanted to minimise impact of our study during the pandemic. Case studies for cohort B were conducted online towards the end of the year (May-June 2021) to also minimise impact in autumn 2020. Although we originally planned to interview timetablers in Autumn, we found that the capacity to talk to the timetabler, head of department and NQT within case studies in the summer presented advantages in evaluating on the impact of the intervention and timetabling more broadly. We therefore continued to conduct case studies in summer for subsequent cohorts.

Responsiveness ratings

Head of Department and Timetable surveys include items evaluating how responsive the IOP were in providing guidance. It was originally intended that the IOP would also provide a simple rating of how responsive the school was in enacting guidance, including how much chasing was required and iterations of guidance and timetable drafting. In practice, schools did not enter into iterative processes of timetable construction with the IOP, despite offers of continued guidance from the IOP.

Implementation and process evaluation data analysis

Surveys will be analysed descriptively and, where appropriate comparisons can be made, using inferential statistics. The case study data and interviews will be analysed thematically (e.g. Braun & Clarke, 2006) and informed by the survey results.

Non-compliance analysis

Within this trial, compliance is closely related to dosage, which denotes the level of 'matchedness' of the timetable. This is therefore to be analysed within the IPE through timetable analysis.

We also anticipate that some schools may provide a matched timetable at the start of the NQT year, but may then change the timetable over the year (for various reasons). We therefore use survey items within the teacher survey to capture this in June/July of the NQT year, and analyse this in relation to non-compliance.

Cost Evaluation

We will follow the December 2019 EEF Guidance on Cost Evaluation[6] in estimating the costs of the delivery of the intervention. Costs will be reported as an average cost over three years per school who is in receipt of the KEEP Teaching intervention. With regards to the direct marginal costs, we will estimate the costs of providing the advice and guidance on timetabling by the developer, and any costs incurred by schools as well as any additional resources. We will collect cost data from the developer via a short interview and either a pro-forma or developer records. In addition, we will collect data on costs incurred by schools through the process evaluation (through case studies and Head of Science surveys).

[6] EEF Cost Evaluation Guidance: https://d2tic4wvo1iusb.cloudfront.net/documents/evaluation/evaluation-design/Cost_Evaluation_Guidance_2019.12.11.pdf?v=1630567719

Ethics and Registration

The trial will seek approval from the UCL Institute of Education Research Ethics Committee.

We intend to process personal data for public interest purposes. (See data protection below). Nevertheless, teachers will sign a MoU where they accept the eligibility terms and conditions set out in the MoU and accept that their contact details and teacher number can be released to the research team. We will provide an opportunity for teachers to withdraw their own data from any data processing as part of the research to ensure that they have no objection to their data being processed in this way. This will demonstrate that the processing does not impinge on anyone's rights and meet our responsibilities under the BERA Ethical Guidelines for Educational Research (particularly regarding informed consent, openness and disclosure).

Outcomes of the project will be publicly reported through an EEF evaluation report and subsequent academic publications. No outcomes will include reporting that could allow for the identification of schools or teachers that participated in the research. The impact estimates will be reported as aggregated statistics while the implementation and process evaluation reporting will ensure that any references to individual schools, teachers are anonymised or removed, where residual risk of identification remains. Impact evaluation data will be securely shared with the EEF's Data Archive as part of their strategy for long term follow-up.

The trial will be registered with the ISRCTN (www.controlled-trials.com) following publication of this evaluation protocol.

Data Protection

Data will be processed in line with data protection legislation (including the General Data Protection Regulation, GDPR), and in line with the interests of the participants. The project will be registered with the UCL Data Protection Officer.

UCL and IOP both perform different roles in this project and use personal data for different purposes, relying on different legal bases. UCL are using the lawful basis known as the 'public task' basis in its capacity as a public authority in connection with its core purposes of research and innovation. Please see UCL's Statement of Tasks in the Public Interest for further information. IOP are using the lawful basis known as the 'legitimate interests' basis. IOP has a legitimate interest in improving the retention of Physics Newly Qualified Teachers, which is consistent with its Charter and status as a charitable organisation.

Head teachers and NQTs are asked to sign an agreement that accepts eligibility terms and conditions set out in a memorandum of understanding (MOU) for each party. The MOUs also provide contact details and for the NQT: agreement that their unique teacher number can be released to the research team; and for the head teacher, details about the school: Local Authority area, county and establishment numbers, Ofsted rating, the percentage of pupils ever eligible for Free School Meals and the percentage of pupils who have English as an Additional Language.

A school and NQT will only be deemed to be part of the project if they have both agreements have been signed.

Head teachers, NQTs and other teachers involved in the research are provided with a Data Privacy Notice, which describe how and why we use their personal data. Their rights and how to exercise their rights in relation to their personal data, are also explained.

The data we hold will be kept securely at all times, transferred using secure (encrypted) methods, and kept on secure computer systems at UCL and IOP's offices under password protection. We will never disclose the name of the school or any personal data collected from the study in any report arising from the research, and we will not include any information that could otherwise identify you or your school.

Personal data will be processed by IOP only for the purposes of this research project. Personal data will not be kept for longer than is necessary for this purpose. After such time, the data will be securely destroyed.

UCL will de-identify information wherever possible (anonymisation or pseudonymisation). Information where individuals can be identified will, as such, be kept for a minimum amount of time and in accordance with the research objectives. For some aspects of the research project, UCL cannot de-identify information as it is necessary for achieving the outcome of the research. For such aspects, UCL will need to store personal information as part of the research for the duration of the project and for a defined period after the project has ended. This is usually defined by external regulations but may be defined by our own policies and procedures. Personal data will be processed by UCL only for the purposes of this research project. Personal data will not be kept for more than 10 years, in line with UCL's policy on storing research data, after such time, personal data will be anonymised. Further details about how long personal information obtained for research is kept can be found in UCL's Data Retention Schedule.

Personnel

IOP *Delivery Team:*

- The Project Manager will authorize work packages for team(s), including deployment of field workers. They will be accountable for: recruitment of 100 participants per year; providing guidance on of matched timetables; and attrition prevention
- The Project Coordinator will assess, prepare and deliver guidance on timetables – responsible for quality of 'matchedness' and provide day to day support for schools.
- The Marketing Officer will be responsible for: marketing and recruitment strategy content and delivery; and recruitment of 100 eligible participants per year and management of their data. They will intervene initially if any attrition to dissuade if possible

UCL Institute of Education Evaluation Team:

- David Wilkinson will be the PI and will be responsible for the overall direction of the evaluation and will lead the impact evaluation.
- Dr Mark Hardman will lead the IPE and will contribute to all other aspects of the evaluation.
- Dr Sam Sims will work on the impact evaluation and will contribute to all other aspects of the evaluation.
- Professor Jeremy Hodgen will advise on design and analysis and undertake internal quality assurance.
- Dr Marian Mulcahy will work on the IPE and will contribute to all other aspects of the evaluation.
- Research Assistants Haira Gandolfi and Claire Pillinger will undertake the fieldwork and analysis relating to the IPE and assist with report-writing.
- An Administrator will provide day-to-day support to the project, including supporting data collection.

Risks

Risk	Likelihood	Impact	Action
Failure to recruit	Low / Moderate	High	• Establish timeline for recruitment involving a variety of methods • Regular developer and evaluator team contact
Failure to gain data from schools on retention	Moderate	High	• Use School Workforce Census (SWF) as back-up. This may lead to a delay in reporting as SWF data is released with delay.
Attrition of schools / teachers	Moderate	Moderate / High	• Over-recruit schools/teachers for efficacy trial (subject to developer's agreement) • Appropriate financial incentives • Regular contact with intervention and control schools • Allocate staff time to school liaison at key data collection points • Regular developer and evaluator team contact
Loss of staff in the UCL team	Low / Moderate	Low	• UCL IOE has a large staff team and would reallocate staff
Fidelity	Moderate	Low / Moderate	• Monitor through process evaluation

Timeline

Recruitment will take place in 3 cohorts to get school and NQTs combinations in place for the NQT year:

- Cohort 1: September 2019 - July 2020
- Cohort 2: September 2020 - July 2021
- Cohort 3: September 2021 - July 2022

Below the timetable for Cohort 1 is outlined, with additional detail for the recruitment of the later cohorts. Activities for Cohort 2 take place one year later and activities for Cohort 3 take place two years later.

An extension for Cohort 3 will also be collected whereby recruitment is allowed beyond the data when timetables are collected. This will allow us to have a bigger sample to examine associations between timetables and outcomes.

Two reports will be submitted. The first report considering job satisfaction, results from the IPE and the regression analysis assessing the association between timetable 'matchedness' and job satisfaction will be submitted in March 2023. An addendum report focused on teacher retention using data from the SWC will be submitted in March 2026.

Dates	Activity	Staff responsible/ leading
Jan – May 2019 Sep 2019 – May 2020 Sep 2020 – May 2021 Nov 2021 – Apr 2022	Recruitment and Initial data collection – Cohorts 1, 2 , 3 and 3+	IOP
Jan – May 2019 Sep 2019 – May 2020 Sep 2020 – May 2021	Randomisation – Cohorts 1, 2 and 3	UCL
May – Jul 2019 May – Jul 2020 May – Jul 2021	Support with timetabling – Cohorts 1, 2 and 3	IOP
Sep 2019 - May 2020 Sep 2020 – Jun 2021 Sep 2021 – Jun 2022	Collect timetables – Cohorts 1, 2, 3 and 3 + Survey, and fieldwork with timetablers, NQTs and Heads of Science - Cohorts 1,2 and 3	UCL
Sep/Oct 2020 and Sep/Oct 2021	Retention data collection with NQTs – Cohorts 1 and 2	UCL
Nov 2022 – Mar 2023	Job satisfaction and IPE analysis	UCL
Mar 2023	Report 1 – Job satisfaction report submitted	UCL

Nov 2025 – Mar 2026	Job retention analysis	UCL
Mar 2026	Report 2 – Job retention report/addendum submitted	UCL

References

Bueno, C. & Sass, T. (2016). The effects of differential pay on teacher recruitment, retention and quality. In 2016 Fall Conference: The Role of Research in Making Government More Effective. Appam.

Clotfelter, C., Glennie, E., Ladd, H., & Vigdor, J. (2008). Would higher salaries keep teachers in high-poverty schools? Evidence from a policy intervention in North Carolina. *Journal of Public Economics, 92* (5), 1352–1370.

Dolton, P. J. (2006). Teacher supply. *Handbook of the Economics of Education*, *2*, 1079-1161.

Donaldson, M. L., and Johnson, S. M. (2010) The price of misassignment: The role of teaching assignments in Teach for America teachers' exit from low-income schools and the teaching profession. *Educational Evaluation and Policy Analysis, 32*(2), 299-323.

Feng, L., & Sass, T. R. (2016). The Impact of Incentives to Recruit and Retain Teachers in "Hard-to-Staff" Subjects. *Journal of Policy Analysis and Management,* 37(1), 112-135.

Fredriksson, P., Öckert, B., & Oosterbeek, H. (2012). Long-term effects of class size. *The Quarterly Journal of Economics*, 128(1), 249-285.

Goldhaber, D., Krieg, J., Theobald, R., & Brown, N. (2014). *The STEM and special education teacher pipelines: Why don't we see better alignment between supply and demand*? CEDR Working Paper 2014-3.

Hill, A. and Jones, D. (2018) A teacher who knows me: The academic benefits of repeat student-teacher matches, Economics of Education Review, Volume 64, 1-12, June 2018.

Hill, H., Rowan, B., and Ball, D. (2005) Effects of Teachers' Mathematical Knowledge for Teaching on Student Achievement. American *Educational Research Journal, 42*(2), 371-406.

Kelly, S. (2004). An event history analysis of teacher attrition: Salary, teacher tracking, and socially disadvantaged schools. *The Journal of Experimental Education*, *72*(3), 195-220.

MAC [Migration Advisory Committee] (2016). *Partial review of the Shortage Occupation List: Review of Teachers*. London: Migration Advisory Committee.

MetLife (2006) The MetLife Survey of the American Teacher: Teachers, Parents and the Economy. Metropolitan Life Insurance Company 2006.

Mocetti, S. (2012). Educational choices and the selection process: before and after compulsory schooling. *Education Economics*, 20(2), 189-209.

Moor, H., Jones, M., Johnson, F., Martin, K., Cowell, E., & Bojke Pharmerit, C. (2006). *Mathematics and science in secondary schools: the deployment of teachers and support staff to deliver the curriculum*. Department for Education and Skills Research Report RR708. London: Department for Education and Skills.

Olmos, F. (2010). *Square peg in a round hole: Out-of-field teaching and its impact on teacher attrition*. University of California, Irvine and California State University, Los Angeles. Available from:

https://search.proquest.com/openview/ee24378b85bb6f8b3c49bef107d49e9b/1?pq-origsite=gscholar&cbl=18750&diss=y

Schanzenbach, D. W. (2006). What have researchers learned from Project STAR?. *Brookings Papers on Education Policy*, (9), 205-228.

Sims, S. (2018). Modelling the relationships between working conditions, teacher job satisfaction and retention. Available from: https://samsimseducation.files.wordpress.com/2018/11/workingconditions_081118.pdf

Sims, S. and Jerrim, J. (2020) *TALIS 2018: teacher working conditions, turnover, and attrition*. Department for Education Statistical working paper, March 2020

Smithers, A. & Robinson, P. (2000). *Coping with teacher shortages*. Centre for Education and Employment Research, University of Buckingham.

Smithers, A., & Robinson, P. (2008). *Physics in schools IV: Supply and retention of teachers*. Centre for Education and Employment Research, University of Buckingham.

Worth, J. & De Lazzari, G. (2017). *Teacher* Retention and Turnover Research*. Update 1: teacher retention by subject*. Slough: National Foundation for Educational Research.